

COMMENTS OF SUSAN VON STRUENSEE, JD, MPH  
to the  
Request for Information and Comment on the National Institute of Standards and Technology's  
Artificial Intelligence Risk Management Framework  
86 FR 40810 (July 29, 2021)  
Agency/Docket Numbers:  
Docket No. Docket Number: 210726-0151

---

Thank you for the opportunity to comment on the National Institute of Standards and Technology's development of a framework that can be used to improve the management of risks to individuals, organizations, and society associated with artificial intelligence (AI).

The World Economic Forum's Centre for the Fourth Industrial Revolution launched a project, *Unlocking Public Sector in AI*, which offered AI procurement guidelines for government and public-sector organizations in November 2018. The Forum's goal was to help officials better understand AI and mitigate potential risks. Acquisition processes are a critical way to mitigate the risks of AI.

Federal procurement law needs to be harnessed in service of ethical algorithmic governance. Professor David S. Rubenstein writes how federal procurement law can and must be retrofitted to meet the unique challenges of algorithmic governance. His work provides a principled and pragmatic approach for doing so. Professor Rubenstein also writes (footnotes omitted) "A recent article by Mulligan and Bamberger warrants special mention here: not only for its important contributions, but also for its analytical route. Descriptively, they argue that "[g]overnment responsibility for policymaking is abdicated" when the adoption of AI systems is governed by procurement, because the embedded policies escape administrative law requirements of public participation and reasoned deliberation. Moreover, Mulligan and Bamberger urge federal and state agencies to shift from a "procurement mindset" to a "policymaking mindset" when acquiring AI systems." He focuses critical attention on the need for AI ethics in federal procurement law. His work steers the conversation in a new direction—namely, toward procurement law's positive potential. More than a marketplace, the AI acquisition gateway must be reimagined as a policymaking space. His articles begin the difficult work of integrating ethical AI into federal procurement law. Rubenstein, David S., Acquiring Ethical AI (Oct. 1, 2020). Florida Law Review, Vol. 73, 2021, Available at SSRN: <https://ssrn.com/abstract=3731106>

The references cited below and attached to this RFI, show the major thought given to government procurement and AI. But those of us charged with implementation need more than top level frameworks that highlight principles and provide general guidance. Have you seen any real-life solicitations, RFPs, or contracts for responsible AI that can serve as models for integrating AI ethics into the acquisition process, any model RFPs or solicitations for governments to use, even as an annex section? Is anyone developing templates, model contract clauses, model RFPs? Can this work be developed via governments? There is some excellent work out there advising what should be in the AI RFPs to integrate AI Ethics, but no models, examples, or templates for government procurement officials to use.

Can the Artificial Intelligence Risk Management Framework provide model contract clauses, model content, model RFPS and solicitations with AI ethics language and templates for practical

implementation and customization? The Framework would be most helpful to focus on Ethical AI in Procurement and provide model RFPs, requirements, contractual language, to ensure we are integrating AI ethics into the acquisition process. Providing model language and templates speeds up implementation. Please include templates and model language as part of the framework.

The World Economic Forum, *AI Government Procurement Guidelines*, and its *AI Procurement in a Box* alludes to developing model solicitations, and RFPs and model contract clauses but none are accessible on-line and the few I found are very rudimentary and I have attached them. Do you agree they need to be developed to get procurement officials to use them-to hasten the Responsible AI adoption process?

Enstrom and Ho describe that the US Federal government is using AI in its Social Security Administration decisions and Security and Exchange Commission (SEC) prosecution decisions, but I could not find the RFPS or the contract agreements to look at the AI Ethics language and how AI ethics is or should be put into the contracts via the RFPS.

Intellectual Property model clauses also needed to be developed. Governments need to be much more explicit about the risks inherent in AI methodologies without AI ethics, and then foster market competition around this challenge. They can do this via RFPS, solicitations, and contracts.

The way to improvement begins with what government asks of its AI vendors. To set the right conditions, the government's market solicitations should contain prompts that explicitly tie ethical AI to methodologies, and committing to those things in the contract. We need to see templates and model solicitations and contract clauses showing what that would look like.

I found very few government solicitations for AI that have an AI ethics component. I attach one I did find at the Crown Commercial Services of the UK.

<https://www.crowncommercial.gov.uk/agreements/search?q=artificial+intelligence>

As you likely know, Canada requires algorithmic impact assessments for government uses of AI, and is integrated into the government's AI acquisition practice. Algorithmic Impact Assessment Tool - Canada.ca

Canada has a pre-certified list of AI vendors, which are on what is the equivalent of a GSA GWAC (or Multiple Award Schedule)? Artificial intelligence source list - Software Acquisition Reference Centre – Buying and Selling – PSPC ([tpsgc-pwgsc.gc.ca](http://tpsgc-pwgsc.gc.ca)). Attached is a Canadian RFP for the AI pre-certified vendor list. They mention AI ethics (at 2.4.1 see below). I do not see questions on how they have applied AI ethics in past work, lessons learned etc.

A view of the Canadian system, the paper, Artificial Intelligence Policy and Funding in Canada: Public Investments, Private Interests Main Findings:

1. Public investments in AI technologies primarily benefit the private sector
2. Even though Canada has federal AI policy, there is no national government AI strategy
- Companies linked to human rights abuses can pre-qualify as government AI suppliers
3. Concentrations of power provide advantages to a handful of entities

The below is the only AI Ethics language I saw in a RFP for AI by the Canadian government, it is not sufficient and better templates with model language needs to be developed:

*Mandatory Criteria 2.4 in Canadian RFP*

*Canada intends to pre-qualify suppliers based on the following mandatory criteria:*

*2.4.1 AI Ethics: Supplier must describe how they address ethical considerations when delivering AI. This could include experience in applying frameworks, methods, guidelines or assessment tools to test datasets and outcomes.*

The implementation of process-based governance frameworks is suggested in the UK Guidance for Understanding AI ethics and safety. This provides a basis to integrate norms, values, and principles informing procedures and protocols that define the project workflow. The Alan Turing Institute calls it a ‘PBG (Process-Based Governance) Framework’. It is helpful, but contains no templates or model language for AI ethics in RFPs or contracts. What is needed are concrete examples showing how the guidance looks in action, when applied.

As Professor Rubenstein states, “the governments in-house capacity challenges are a major concern. It is one thing if agencies must rely on vendors to satisfy the governments demand for AI tools. It is quite another if the government does not have the resources and capacity to responsibly manage AI acquisitions.” Hopefully, the framework will include the types of templates, model clauses, and model solicitations to assist procurement officials integrate AI Ethics into the Responsible AI acquisition process.

Respectfully Submitted,

Susan von Struensee, JD, MPH

**Sources:**

DataEthics.eu, *White Paper on Data Ethics in Public Procurement of AI-based Services and Solution*, (April 2020) <https://www.dataethics.eu/wp-content/uploads/dataethics-whitepaper-april-2020.pdf>

Rubenstein, David S., *Acquiring Ethical AI* (Oct. 1, 2020). Florida Law Review, Vol. 73, 2021, Available at SSRN: <https://ssrn.com/abstract=3731106> (PRE\_PRINT COPY-IS BEING UPDATED)  
Abstract-Artificial intelligence (AI) is transforming how government operates. Federal agencies use the technology for law enforcement, adjudication, rulemaking, inhouse management, and the delivery of public services. Algorithmic governance brims with promise and peril. Under the right conditions, AI systems can solve complex problems, reduce administrative burdens, and optimize resource allocations. Under the wrong conditions, AI systems can lead to widespread discrimination, invasions of privacy, and the erosion of democratic norms. The United States has pledged its commitment to principles of “ethical AI,” including transparency, accountability, fairness, and human rights. But proselytizing is not actualizing. A burgeoning literature has emerged to square algorithmic governance with the precepts of constitutional and administrative law. Federal procurement law, however, remains a dangerous blind spot in the reformist agenda. The government’s pent up demand for AI systems far exceeds its inhouse capacity to develop, field, and monitor this powerful technology. Accordingly, many if not most of the tools of algorithmic governance will be procured by contract from the technology industry. This Article intervenes with a principled and pragmatic agenda for acquiring ethical AI. First, it provides an original

account that aligns the ambition of algorithmic governance, the imperative of ethical AI, and the complexities of procurement law. Second, the Article argues that procurement law is not only uniquely situated, but also well suited, to serve as a checkpoint and catalyst for ethical algorithmic governance. Third, the Article prescribes a set of concrete regulatory reforms to center ethical AI throughout the procurement process: from acquisition planning through market solicitation, negotiation, and contractual award. The outsourcing of algorithmic governance raises a host of challenges that constitutional law and administrative law are ill equipped to handle. Procurement law will not solve all the challenges of algorithmic governance. Just as surely, the challenges of algorithmic governance cannot be solved without procurement law.

David S. Rubenstein, FEDERAL PROCUREMENT OF ARTIFICIAL INTELLIGENCE: PERILS AND POSSIBILITIES, The Great Democracy Initiative, December 2020

<https://greatdemocracyinitiative.org/document/federal-procurement-of-artificial-intelligence-perils-and-possibilities/> Abstract: The proliferation of Artificial Intelligence (AI) use by federal agencies raises urgent questions about how these new technologies should be regulated. Today, AI procurement helps streamline processes in agencies like the Social Security Administration, the Food and Drug Administration, Homeland Security, and more. Under the right conditions, algorithmic governance can be an incredibly useful tool, however, under the wrong conditions, it can lead to widespread discrimination, invasion of privacy, and the degradation of democratic principles. Yet, much, if not most, of the AI used by federal agencies will be procured from a virtually unregulated private market. In this report, David. S Rubenstein shows that when the government acquires AI, it is often procuring the policy choices of the nongovernmental actors who designed the technology.

Rubenstein outlines a plan for ethical AI procurement, going forward: mandating the creation of a government-wide inventory that includes clear information on AI systems used by federal agencies, requiring agencies to prepare “AI risk assessment” reports prior to acquiring AI services, and integrating ethical AI consideration into existing regulations for source selection. Federal procurement of AI services must be reimagined as more than just a marketplace but rather a policymaking space that promotes trustworthy and ethical AI.

Engstrom and Ho, ALGORITHMIC ACCOUNTABILITY IN THE ADMINISTRATIVE STATE, 37 YALE J. ON REG. (2020)

<https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1563&context=yjreg>

Storia Law, *8 Points about AI Development Agreements that can be learned from the “Contract Guidance on Utilization of AI and Data”* at <https://storialaw.jp/en/service/bigdata/ai-13>

Rebecca S. Eisner and Brad L. Peterson, *Smart Licensing of Artificial Intelligence*, (May 16, 2019) <https://www.mayerbrown.com/en/perspectives-events/publications/2019/05/smart-licensing-of-artificial-intelligence>

Rebecca S. Eisner, (Sept. 2020) <https://www.mayerbrown.com/-/media/files/perspectives-events/publications/2020/09/tbfall20ofnoteipt.pdf>

Practical Law, *Expert Q&A with Rebecca Eisner of Mayer Brown LLP on artificial intelligence (AI) licensing*, <https://www.mayerbrown.com/-/media/files/news/2019/01/expert-qanda-on-artificial-intelligence-ai-licensing-w0219801.pdf>

Leslie, David, *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector*, (June 10, 2019). Available at SSRN: <https://ssrn.com/abstract=3403301> or <http://dx.doi.org/10.2139/ssrn.3403301>

This guide, written for department and delivery leads in the UK public sector and adopted by the British Government in its publication, 'Using AI in the Public Sector,' identifies the potential harms caused by AI systems and proposes concrete, operationalisable measures to counteract them. It stresses that public sector organisations can anticipate and prevent these potential harms by stewarding a culture of responsible innovation and by putting in place governance processes that support the design and implementation of ethical, fair, and safe AI systems. It also highlights the need for algorithmically supported outcomes to be interpretable by their users and made understandable to decision subjects in clear, non-technical, and accessible ways. Finally, it builds out a vision of human-centered and context-sensitive implementation that gives a central role to communication, evidence-based reasoning, situational awareness, and moral justifiability.

Mulligan, Deirdre K. and Bamberger, Kenneth A., *Procurement As Policy: Administrative Process for Machine Learning*, (October 4, 2019). Berkeley Technology Law Journal, Vol. 34, 2019, Available at SSRN: <https://ssrn.com/abstract=3464203> or <http://dx.doi.org/10.2139/ssrn.3464203>

Abstract: At every level of government, officials contract for technical systems that employ machine learning—systems that perform tasks without using explicit instructions, relying on patterns and inference instead. These systems frequently displace discretion previously exercised by policymakers or individual front-end government employees with an opaque logic that bears no resemblance to the reasoning processes of agency personnel. However, because agencies acquire these systems through government procurement processes, they and the public have little input into—or even knowledge about—their design or how well that design aligns with public goals and values. This Article explains the ways that the decisions about goals, values, risk, and certainty, along with the elimination of case-by-case discretion, inherent in machine-learning system design create policies—not just once when they are designed, but over time as they adapt and change. When the adoption of these systems is governed by procurement, the policies they embed receive little or no agency or outside expertise beyond that provided by the vendor. Design decisions are left to private third-party developers. There is no public participation, no reasoned deliberation, and no factual record, which abdicates Government responsibility for policymaking. This Article then argues for a move from a procurement mindset to policymaking mindset. When policy decisions are made through system design, processes suitable for substantive administrative determinations should be used: processes that foster deliberation reflecting both technocratic demands for reason and rationality informed by expertise, and democratic demands for public participation and political accountability. Specifically, the Article proposes administrative law as the framework to guide the adoption of machine learning governance, describing specific ways that the policy choices embedded in machine-learning system design fail the prohibition against arbitrary and capricious agency actions absent a reasoned decision-making process that both enlists the expertise necessary for reasoned deliberation about, and justification for, such choices, and makes visible the political choices being made. Finally, this Article sketches models for machine-learning adoption processes that satisfy the prohibition against arbitrary and capricious agency actions. It explores processes by which agencies might garner technical expertise and overcome problems of system opacity, satisfying administrative law's technocratic demand for reasoned expert deliberation. It further proposes both institutional and engineering design solutions to the challenge of policymaking opacity, offering process paradigms to ensure the "political visibility" required for public input and political oversight. In doing so, it also proposes the importance of using "contestable design"—design that exposes value-laden features and parameters and provides for iterative human involvement in

system evolution and deployment. Together, these institutional and design approaches further both administrative law's technocratic and democratic mandates.

Naudé, Wim and Dimitri, Nicola, *Public Procurement and Innovation for Human-Centered Artificial Intelligence*. IZA Discussion Paper No. 14021, Available at SSRN: <https://ssrn.com/abstract=3762891>  
Abstract-The possible negative consequences of Artificial Intelligence (AI) have given rise to calls for public policy to ensure that it is safe, and to prevent improper use and misuse. Human-centered AI (HCAI) draws on ethical principles and puts forth actionable guidelines in this regard. So far however, these have lacked strong incentives for adherence. In this paper we contribute to the debate on HCAI by arguing that public procurement and innovation (PPAI) can be used to incentivize HCAI. We dissect the literature on PPAI and HCAI and provide a simple theoretical model to show that procurement of innovative AI solutions underpinned by ethical considerations can provide the incentives that scholars have called for. Our argument in favor of PPAI for HCAI is also an argument for the more innovative use of public procurement, and is consistent with calls for mission-oriented and challenge-led innovation policies. Our paper also contributes to the emerging literature on public entrepreneurship, given that PPAI for HCAI can advance the transformation of society, but only under uncertainty.

The World Economic Forum, AI Procurement in a Box: Challenges and opportunities during implementation, World Economic Forum, June 2020

Workshop participants explored various themes related to the governments use of AI and how procurement plays a role in government adoption of the technology.

[http://www3.weforum.org/docs/WEF\\_AI\\_Procurement\\_in\\_a\\_Box\\_Challenges\\_and\\_Opportunities\\_during\\_implementation\\_2020.pdf](http://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_Challenges_and_Opportunities_during_implementation_2020.pdf)

The World Economic Forum, *AI Government Procurement Guidelines*, (Sept. 2019)

[http://www3.weforum.org/docs/WEF\\_Guidelines\\_for\\_AI\\_Procurement.pdf](http://www3.weforum.org/docs/WEF_Guidelines_for_AI_Procurement.pdf)

*Exploring Blockchain Technology for Government Transparency: Blockchain-Based Public Procurement to Reduce Corruption* | World Economic Forum (weforum.org), June 2020,

[http://www3.weforum.org/docs/WEF\\_Blockchain\\_Government\\_Transparency\\_Report.pdf](http://www3.weforum.org/docs/WEF_Blockchain_Government_Transparency_Report.pdf)

*WEF\_AI\_Procurement\_in\_a\_Box\_Workbook\_2020.pdf* (weforum.org)

[http://www3.weforum.org/docs/WEF\\_AI\\_Procurement\\_in\\_a\\_Box\\_Workbook\\_2020.pdf](http://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_Workbook_2020.pdf)

European Commission, *Emerging technologies in public procurement*, (June 2020)

[https://ec.europa.eu/growth/single-market/public-procurement/digital/emerging-technologies\\_en](https://ec.europa.eu/growth/single-market/public-procurement/digital/emerging-technologies_en)

Emerging technologies can transform public procurement. From automating repetitive administrative tasks to providing unprecedented information and analysis regarding spending patterns and project results, new technology can enable better decisions, lower costs, and increase transparency. In 2019, the Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs commissioned under the ISA2 a study on the uptake of emerging technologies in public procurement. This study examined how public authorities around the globe are using new technologies when procuring goods and services. Covering technologies including artificial intelligence and machine learning, big data and data analytics, blockchain, robotic process automation, augmented and virtual reality, internet of things, and drones, the study features:

- a longlist of 96 projects (as of January 2020) where these new technologies have been explored or used by public authorities for procurement

- 20 detailed case studies presenting the issue addressed, impact, cost, requirements and risks of particularly interesting emerging technology projects
- a final report presenting the overall approach and findings of the project, including 10 recommendations for the application of emerging technologies to public procurement
- Update (22 June 2020): 10 new projects have been added. 3 are from Portugal, 3 are from Estonia, and the remaining 4 are from Slovenia, the Netherlands, Italy and Spain. The most frequently used emerging technologies in these new projects are business intelligence, blockchain and artificial intelligence.

European Commission, *Study on up-take of emerging technologies in public procurement*, (Feb. 27, 2020) <https://ec.europa.eu/docsroom/documents/40102>

*Confronting Bias: BSA's Framework to Build Trust in AI* / BSA | The Software Alliance.  
<https://ai.bsa.org/confronting-bias-bsas-framework-to-build-trust-in-ai>

Raji, et al, *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, <https://dl.acm.org/doi/pdf/10.1145/3351095.3372873>

Mark Treveil, Nicolas Omont, Clément Stenac, Kenji Lefevre, Du Phan, Joachim Zentici, Adrien Lavoillotte, Makoto Miyazaki, Lynn Heidmann, *Introducing MLOps*, November 2020,  
<https://www.oreilly.com/library/view/introducing-mlops/9781492083283/>

More than half of the analytics and machine learning (ML) models created by organizations today never make it into production. Some of the challenges and barriers to operationalization are technical, but others are organizational. Either way, the bottom line is that models not in production can't provide business impact. This book introduces the key concepts of MLOps to help data scientists and application engineers not only operationalize ML models to drive real business change but also maintain and improve those models over time. Through lessons based on numerous MLOps applications around the world, nine experts in machine learning provide insights into the five steps of the model life cycle--Build, Preproduction, Deployment, Monitoring, and Governance--uncovering how robust MLOps processes can be infused throughout.

This book helps you:

- Fulfill data science value by reducing friction throughout ML pipelines and workflows
- Refine ML models through retraining, periodic tuning, and complete remodeling to ensure long-term accuracy
- Design the MLOps life cycle to minimize organizational risks with models that are unbiased, fair, and explainable
- Operationalize ML models for pipeline deployment and for external business systems that are more complex and less standardized

*Guidelines for AI procurement, A guide to using artificial intelligence in the public sector*, (June 8, 2020) <https://www.gov.uk/government/publications/guidelines-for-ai-procurement/guidelines-for-ai-procurement>

DHS AI Strategic Plan

[https://www.dhs.gov/sites/default/files/publications/21\\_0730\\_st\\_ai\\_ml\\_strategic\\_plan\\_2021.pdf](https://www.dhs.gov/sites/default/files/publications/21_0730_st_ai_ml_strategic_plan_2021.pdf)

U.S. Government Accountability Office, *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities, Highlights of GAO-21-519SP*, available at <https://www.gao.gov/assets/gao-21-519sp-highlights.pdf>.

## **RM6200 Artificial Intelligence Dynamic Purchasing System - Data Ethics Letter of Understanding**

Purpose: to ensure high standards of ethical conduct are upheld when adopting technologically assisted decision making in the public sector, in accordance with the principles and recommendations in the Committee on Standards and Public Life's report [Artificial Intelligence and Public Standards](#).

It is important that suppliers who bid for work under the RM6200 Artificial Intelligence (AI) Dynamic Purchasing System (DPS) are committed not only to delivering the technical elements of the procurement but also delivering ethically where a buyer has stated that there is an ethical dimension to their tender.

The Office for Artificial Intelligence (AI), Government Digital Service (GDS) and Alan Turing Institute published [ethical principles for data-driven technology](#) in the jointly issued [A Guide to Using Artificial Intelligence in the Public Sector](#), in June 2019.

The Department for Digital, Culture, Media and Sport published a collection page in July 2020, with main [data ethics and AI guidance](#). Public servants working with data and AI will use this collection of guidance when buying technology, products or services under the RM6200 Artificial Intelligence DPS. It is important that suppliers are aware of the standards and frameworks that will affect the buying decisions of Buyer organisations and will adhere to these as appropriate.

Suppliers may be asked to provide evidence of how the government's [Data Ethics Framework](#) principles have been followed during the development and implementation of the technology, product or service, at the award of an Order Contract.

The following list of requirements for Artificial Intelligence suppliers is an example and not exhaustive, and may be developed during the DPS Contract Period and by the Buyer organisation.

### Transparency and explainable AI

- The Supplier should describe the capabilities in the business to ensure the outputs of the AI technology are explainable, and that this explanation is widely available and understandable to a non-expert audience.

### Ethical considerations relation to data limitations, fairness and bias

- The Supplier should identify data limitations and implement strategies to address these data limitations.
- The Supplier should be able to describe the approach to eliminate (or minimise) bias, ethical issues, or other safety risks as a result of using the service.

- The Supplier should be able to describe how they have ensured that the data used to power the AI solution is sufficient in quantity, accuracy and relevance to the data available, and what measures have been taken to mitigate bias in the model.
- The Supplier should be able to demonstrate how they consider the skills, qualifications and diversity of the team developing and deploying AI systems.

#### Consent and Control

- The Supplier should adopt legally sound and ethical consent for processing and capturing data throughout the full lifecycle of the solution and be able to describe the level of human decision-making at critical points.

#### Privacy and cybersecurity

- The Supplier should be able to describe their privacy and cybersecurity approach for the proposed solution, in particular how the data will be protected.

The Supplier shall cooperate in good faith with CCS to develop efficiency tracking performance measures for Data Ethics Performance Indicators in accordance with RM6200 Artificial Intelligence DPS Schedule 4 (DPS Management) clause 4 (How the Supplier's Performance will be measured), if required to do so by CCS.

Suppliers appointed to the RM6200 Artificial Intelligence DPS will continue to meet government standards, guidelines and regulations as they develop in this industry.

In signing this letter of understanding you acknowledge that where a Buyer has stated that there is an ethical dimension to their procurement, you will only bid for work where you are willing to deliver both ethical as well as technical dimensions of a tender.

Signed by the Authorised Representative of .....  
*[insert company name]*

Signature .....

Name (please print) .....

Position .....

Date .....



# AI PLAYBOOK FOR THE U.S. FEDERAL GOVERNMENT

**EMERGING TECHNOLOGY COMMUNITY OF INTEREST**  
**Artificial Intelligence Working Group**

Date Released: January 22, 2020

## Synopsis

This Playbook immediately follows the ACT-IAC Artificial Intelligence/Machine Learning Primer and proposes a process and a series of phases to support the United States Federal Government in its understanding and application of artificial intelligence (AI) technologies to support its mission. Each phase contains a set of key activities organized in functional areas that go beyond just the technical aspects of AI but include management, people, process, and acquisition areas.

AI has the potential to help government mitigate fraud, reduce errors, and lower the cost of paper-intensive processes, while enabling collaboration across multiple divisions and agencies to provide more effective and efficient services to citizens. Moreover, the adoption of AI may also allow government agencies to provide new value-added services to citizens which can generate new sources of revenue and achieve agency objectives.

*This page is intentionally blank.*

### **American Council for Technology-Industry Advisory Council (ACT-IAC)**

The American Council for Technology (ACT) is a non-profit educational organization established to create a more effective and innovative government. ACT-IAC provides a unique, objective, and trusted forum where government and industry executives are working together to improve public services and agency operations through the use of technology. ACT-IAC contributes to better communication between government and industry, collaborative and innovative problem solving, and a more professional and qualified workforce.

The information, conclusions, and recommendations contained in this publication were produced by volunteers from government and industry who share the ACT-IAC vision of a more effective and innovative government. ACT-IAC volunteers represent a wide diversity of organizations (public and private) and functions. These volunteers use the ACT-IAC collaborative process, refined over thirty years of experience, to produce outcomes that are consensus-based. The findings and recommendations contained in this report are based on consensus and do not represent the views of any particular individual or organization.

To maintain the objectivity and integrity of its collaborative process, ACT-IAC does not accept government funding.

ACT-IAC welcomes the participation of all public and private organizations committed to improving the delivery of public services through the effective and efficient use of IT. For additional information, visit the ACT-IAC website at [www.actiac.org](http://www.actiac.org).

### **Emerging Technology of Community of Interest**

ACT-IAC, through the Emerging Technology Community of Interest, formed an Artificial Intelligence Working Group to give voice to and provide an authoritative resource for government agencies looking to understand and incorporate AI/ML technology and functionality into their organizations. This working group includes government and industry thought leaders incubating government use cases. The ACT-IAC Emerging Technology Community of Interest (ET COI) mission is to provide an energetic, collaborative consortium comprised of leading practitioners in data science, technology, and research, engaged with industry, academia, and public officials and executives focused on emerging and leading technologies which transform public sector capabilities.

### **Disclaimer**

This document has been prepared to contribute to a more effective, efficient, and innovative government. The information contained in this report is the result of a collaborative process in which a number of individuals participated. This document does not – nor is it intended to – endorse or recommend any specific technology, product, or vendor. Moreover, the views expressed in this document do not necessarily represent the official views of the individuals and

organizations that participated in its development. Every effort has been made to present accurate and reliable information in this report. However, ACT-IAC assumes no responsibility for consequences resulting from the use of the information herein.

**Copyright**

©American Council for Technology, 2020. This document may not be quoted, reproduced and/or distributed unless credit is given to the American Council for Technology-Industry Advisory Council.

**Further Information**

For further information, contact the American Council for Technology-Industry Advisory Council at (703) 208-4800 or [www.actiac.org](http://www.actiac.org).

## Table of Contents

Guide to Reading this Playbook .....	8
INTRODUCTION .....	10
Phase 1 – Problem Assessment .....	13
Phase Inputs	13
Key Goals	13
Key Participants	13
Key Considerations	16
Key Activities	19
Management.....	19
People .....	19
Process .....	20
Technology.....	21
Acquisition.....	21
Key Outcomes	21
Engaged.....	21
Defined.....	22
Phase Outputs	22
Decision Gate	23
Phase 2 – Organizational Readiness .....	24
Phase Inputs	24
Key Goals	24
Key Participants	25
Key Considerations	25
Approach Guidance.....	28
Readiness Checklist.....	28
Key Activities	30
Management.....	30
People .....	30
Process .....	30
Technology.....	30
Acquisition.....	31
Key Outcomes	31
Phase Outputs	31
Decision Gate	32
Phase 3 – Solution Selection .....	34
Phase Inputs	34
Key Goals	35
Key Activities	35

Management.....	35
People .....	36
Process .....	37
Technology .....	38
Acquisition.....	39
Key Outcomes	40
Phase Outputs	41
Decision Gate	41
Phase 4 – AI Implementation.....	43
Phase Inputs	43
Key Goals	45
Key Considerations	45
Key Activities	47
Management.....	47
People .....	48
Process .....	50
Technology .....	52
Acquisition.....	57
Key Outcomes	59
Phase Outputs	60
Decision Gate	61
Phase 5 – AI Integration.....	63
Phase Inputs	63
Key Goals	63
Key Participants	64
Key Considerations	64
Key Activities	65
Management.....	65
People .....	65
Process .....	66
Technology .....	68
Acquisition.....	70
Key Outcomes	71
Phase Outputs	71
CONCLUSION.....	72
GLOSSARY.....	73
ACKNOWLEDGEMENT.....	76
Authors and Affiliations	76
APPENDICES .....	78

Appendix A - AI Functionality Template .....	78
Appendix B – Playbook Navigation .....	80
Framework Flow .....	80
REFERENCES .....	81

## Guide to Reading this Playbook

### Where do I start?

The INTRODUCTION contains a graphic that represent the overall process to help organizations implement and integrate artificial intelligence (AI) solutions. Each phase of the process contains a more detailed graphic that summarizes the objective of the phase, its key activities and outcomes, inputs, and outputs. It provides a preview and a good starting point for each phase.

### I am a senior executive, what should I focus on?

Each playbook phase is composed of phase inputs, phase outputs, key activities and outcomes, and a decision gate. It is recommended that you read the introduction, the summation of each of the five phases, and the conclusion. The inputs/outputs of each phase as well as the phase decision gate are useful in understanding the functionality of AI capabilities.

### I am on the management team, what should I focus on?

As part of the management team, you need to understand the overall process. The diagrams are a good start. You should also focus on the “management” key activity category of each phase. You can also look at the “management” key activity category of each phase to get an understanding of the role of the management team.

### I am a data architect, scientist, and develop AI algorithms, how should I read this playbook?

This playbook is written from a top down perspective. The first part provides the strategic goals and drills down through the operational objectives and tactical requirements.

### I am on the acquisition team, what should I focus on?

Each phase has an “acquisition” key activities category. There is also an acquisition section starting on page 57.

### I am on the development team, what should I focus on?

You should have a good understanding of the inputs needed to implement the technical solution. Focus on the “technology” key activities category of each phase.

### What is AI?

The [ACT-IAC Artificial Intelligence/Machine Learning Primer](#)<sup>1</sup> provides the definition of AI, limitations, risk, ethical considerations, and the impact it can have on organizations.

### What about my workforce?

Each playbook phase has a set of key activities that are grouped in high level categories. The category labeled “People” highlights key activities regarding the workforce. Also, the Readiness Phase handles organizational readiness.

**What phase(s) should I follow for a proof of concept?**

All the phases should be used to produce a proof of concept with only the most valuable functions of the use case developed.

**Can I use the playbook in an agile manner?**

This playbook is intended to be used in an iterative way. One can go through the playbook phases multiple times to develop proofs of concept, pilots, and full implementation. Each iteration will dive deeper in the activities of each phase. Iterations can also occur to refine data in each phase and adapt the roadmap.

**How do I handle a use case with multiple organizations?**

Each phase deals more or less with understanding the organization(s) undergoing the implementation of the solution. If you know more than one organization will be involved, when you look at the organizational aspects of the solution, address from the start the need for a multi-organization solution

## INTRODUCTION

As noted in the ACT-IAC Artificial Intelligence/Machine Learning Primer<sup>1</sup>, artificial intelligence may be applied to help government reduce fraud, errors, and cost of paper-intensive processes, while enabling collaboration across multiple divisions and agencies to provide more efficient and effective services to citizens. The adoption of AI may enable government agencies to provide new value-added services and serve as a catalyst to modernize IT. How can agencies turn that potential into reality?

**Understand the technology using the ACT-IAC primer:** Over a dozen federal agencies and a variety of industry partners collaborated to develop the ACT-IAC Artificial Intelligence/Machine Learning Primer which provides government with an introduction to AI, outlines its related technologies, and presents several potential use cases.

**Incorporate AI functionality using the ACT-IAC playbook:** The ACT-IAC Artificial Intelligence Working Group developed this playbook to guide government in taking the appropriate steps and developing the necessary plans to appropriately implement this technology to achieve the goals of their specific missions.

**AI and data-centered organizations:** AI has the potential to significantly impact both business processes as well as provide the foundational capabilities to achieve the objectives that fulfill the goals of the organization.

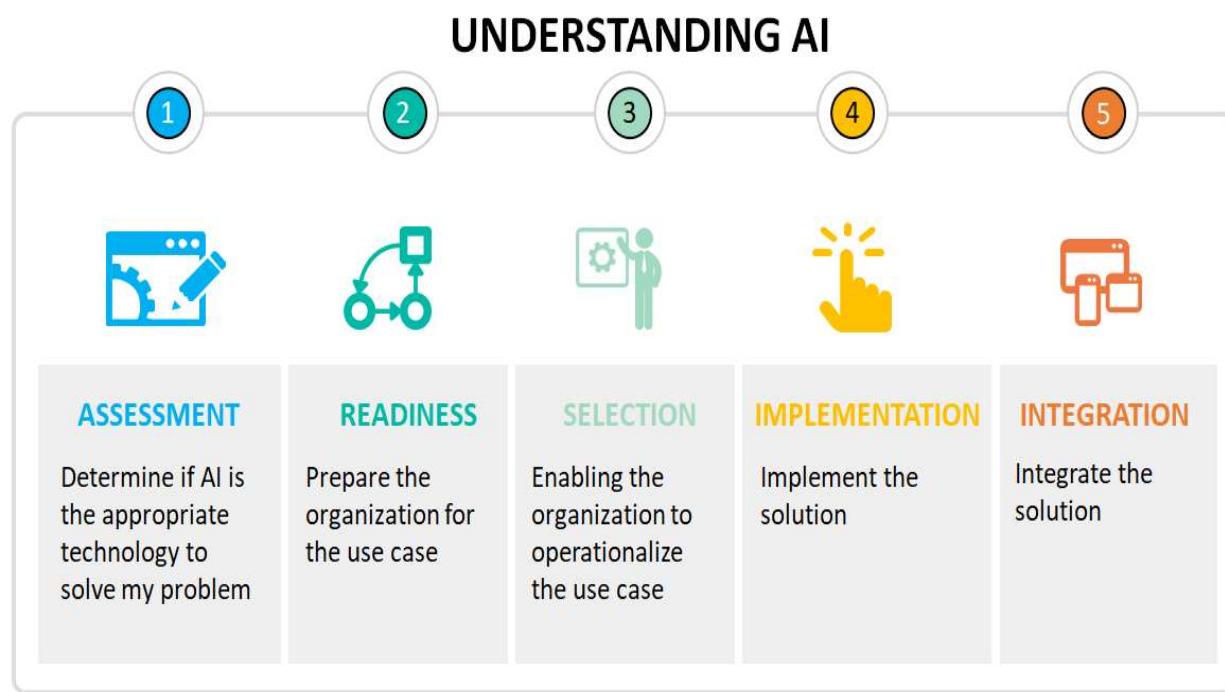
This playbook applies the concepts of the ACT-IAC playbook framework (Appendix B) as well as the General Services Administration's Modernization and Migration Management (M3)<sup>2</sup> unified shared services framework to help government achieve successful outcomes and reduce risk during an AI deployment. The progression of this framework ensures the application of the appropriate models to optimize available resources in order to deliver the most effective solution.

By leveraging this playbook during each iteration of AI implementation, organizations can understand the actions necessary to deliver minimally viable product (MVP), proof of concept (POC), pilot/limited fielding, initial operational capability (IOC), and a fully operational system to support their overarching objectives. Not all topics in each phase may apply to all iterations. However, this playbook should remain useful as a step by step process to deliver solutions and provide a scalable product that is sustainable through the lifecycle of development, implementation, and recapitalization.

It is also important to note that at scale and fully implemented, AI technology will probably require several solutions and provide an evolutionary catalyst to solve and support the evolutions and transformation of the entire organization. Therefore, it is important that at the

onset, any organization interested in leveraging AI will need to define the appropriate stakeholders and the group (network peers) that will participate in the steps outlined in the playbook. It serves as a catalyst to the necessary cultural transformation at the core of change management essential for the evolutions of today's organizations.

As government efforts move through implementation of this new and rapidly developing technology, contributions to this playbook (e.g., additional best practices, lessons learned, and feedback) are welcome to keep this resource current, comprehensive, and effective in meeting the needs of government.



*Figure 1: ACT-IAC AI playbook phases*

**Phase 1 - Problem Assessment:** Develop a vision and business objectives through various assessments to ensure the AI solution addresses a specific use case and delivers results that optimize services and operational delivery.

**Phase 2 - Organizational Readiness:** Engage AI subject matter experts and consider the nuances that accompany an AI solution to prepare the organization. This includes creating a project management office, as well as the establishment of AI-tailored business, functional and technical requirements, and implementation plans.

**Phase 3 – Solution Selection:** Conduct a thorough investigation of business consideration, types of AI requirements, deployment models, and procurement options to enable optimal provider selection to achieve the desired end state.

**Phase 4 - AI Implementation:** Customize and configure AI solution to meet the organization's operational objectives.

**Phase 5 - AI Integration:** Integrate AI solution into the organization's infrastructure.

## KEY ACTIVITIES

	Management	People	Process	Technology	Acquisition
 ASSESSMENT	<ul style="list-style-type: none"> <li>Establish AI Inventory &amp; Definition Set</li> <li>Capture the Need and Use Cases for problem</li> <li>Document the objective trying to be achieved</li> </ul>	<ul style="list-style-type: none"> <li>Define who will use the AI</li> <li>Workforce Readiness (Knowledge &amp; Capability &amp; Skill)</li> <li>Willingness (Perception of Value Benefit vs Consequence)</li> </ul>	<ul style="list-style-type: none"> <li>Map the use case to the AI</li> <li>Define the ethical boundaries for the AI</li> <li>What is the impact of the AI</li> </ul>	<ul style="list-style-type: none"> <li>Assess how sophistication and maturity of the AI</li> <li>Evaluate the AI's fitness for the intended use</li> <li>Identify capability differentiators</li> </ul>	<ul style="list-style-type: none"> <li>Capture the solution value and outcomes</li> <li>Define your constraints (Cost &amp; Schedule)</li> </ul>
 READINESS	<ul style="list-style-type: none"> <li>Change Management</li> <li>Bring together broad stakeholder group</li> <li>Establishing Working groups</li> </ul>	<ul style="list-style-type: none"> <li>Uncover Stakeholder &amp; User concerns</li> <li>Identify Skill Gaps</li> <li>Workforce transition</li> <li>Change Management</li> </ul>	<ul style="list-style-type: none"> <li>Analysis of business process to be automated</li> <li>Assess As-Is State</li> <li>Gap Analysis</li> <li>Identify key users</li> </ul>	<ul style="list-style-type: none"> <li>Assessing Current Infrastructure</li> <li>Data Sources</li> <li>Identify integration points</li> <li>Determine AI COTS vs Open Source Tools</li> </ul>	<ul style="list-style-type: none"> <li>Available Budget/Schedule</li> <li>Market Research—tools &amp; capabilities</li> <li>Review Vendor Capabilities</li> <li>RFI</li> </ul>
 SELECTION	<ul style="list-style-type: none"> <li>Identify right problem for AI</li> <li>Allocate Budget</li> <li>Bring Right Team together</li> <li>Follow DevSecOps approach</li> </ul>	<ul style="list-style-type: none"> <li>Identify skill gaps</li> <li>Inform staff on pros &amp; cons of AI</li> <li>Train SMEs on AI &amp; specific SME tasks around training</li> </ul>	<ul style="list-style-type: none"> <li>Review Data availability and completeness</li> <li>Develop change management plan</li> </ul>	<ul style="list-style-type: none"> <li>Assess technology platforms that fit the need</li> <li>Consider Explainability</li> <li>Define Architecture</li> </ul>	<ul style="list-style-type: none"> <li>Define Outcomes</li> <li>Plan Purchase of solutions selected</li> </ul>
 IMPLEMENTATION	<ul style="list-style-type: none"> <li>Change &amp; Communication Management</li> <li>Establish an Iterative approach</li> <li>Workforce planning</li> <li>Governance &amp; Oversight</li> <li>Stakeholder Updates</li> </ul>	<ul style="list-style-type: none"> <li>Resource Allocations</li> <li>Continuous Skill Audit and Training</li> <li>Fill skill gaps</li> <li>Deliver Required Training</li> <li>Assign Project Manager</li> </ul>	<ul style="list-style-type: none"> <li>Define Requirements</li> <li>Establish operations model</li> <li>Identify Integration Points</li> <li>Obtain ATO or IATT</li> <li>Conduct Proof-of-Concept / Pilot if applicable</li> </ul>	<ul style="list-style-type: none"> <li>Setup, Configure &amp; Deploy Platform &amp; Tools</li> <li>Develop solution in an iterative manner</li> <li>Data Preparation</li> <li>Model verification for Bias</li> <li>Integrate with existing applications</li> </ul>	<ul style="list-style-type: none"> <li>Administrator contract</li> <li>Modify Contract</li> <li>Prepare follow-on acquisition</li> <li>Award follow-on acquisition</li> </ul>
 INTEGRATION	<ul style="list-style-type: none"> <li>Monitor, schedule, budget and velocity</li> <li>Auditability dynamic algorithm updates</li> <li>Establish Risk Security, Communication Plans</li> </ul>	<ul style="list-style-type: none"> <li>Education</li> <li>Outreach</li> <li>Change Management</li> <li>Monitor skill gaps</li> <li>Incentivize adoption</li> </ul>	<ul style="list-style-type: none"> <li>Execute Change Management</li> <li>Execute Scaling Process</li> <li>Auditability dynamic algorithm updates</li> <li>Execute Risk Management</li> </ul>	<ul style="list-style-type: none"> <li>Provision technology resources</li> <li>Configuration management</li> <li>Link AI solution to inputs and outputs</li> </ul>	<ul style="list-style-type: none"> <li>Administer and monitor contract performance</li> <li>Prepare contract for scaling and reuse</li> <li>Contract Modification</li> </ul>

Figure 2: AI Playbook phases and key activities matrix

## Phase 1 – Problem Assessment

The first phase is designed to help decision makers create the most value through their AI initiative. It includes tools to ensure that the initiative is designed to address a specific use case and advancing mission goals, even if that is not an AI solution. Inputs and outputs artifacts are organized in 3 categories - AI, Business Need, and Governance Risk & Compliance (GRC).

### Phase Inputs

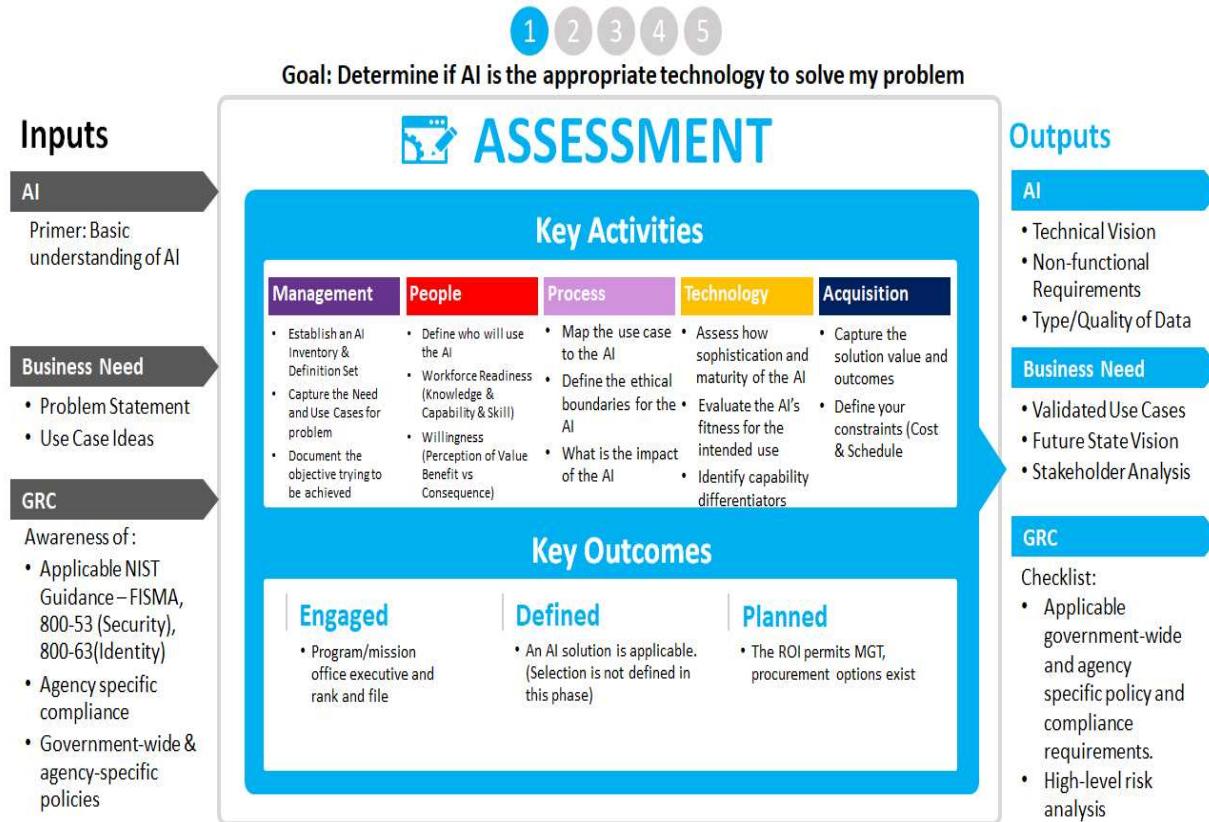


Figure 3: assessment phase (1) summary

### Key Goals

- Determine if AI is the appropriate technology solution.
- Reassess outcomes of later playbook stages to validate and ensure that the AI solution is still the best option to fulfill the goals and objectives of the organization.
- Develop future reassessment questions.

### Key Participants

- Business Sponsor/Advocates/Executives/Strategists/Program Manager/Stakeholders

- Technologist/Enterprise Architects/Computer Engineers/Security and Risk Managers
- Functional Data Stewards/Architects/Scientist/Visualizers/Subject Matter Experts (*SME*)
- AI Technology SME/Product Manager/Programmer/Integrator

## **"DO I NEED AI?" Assessment Questionnaire**

The following are key assessment questions to consider as a preliminary guide for those considering an AI approach. Afterwards, determine your total score to gain insight into the possibility for a substantial Return on Operations (*ROO-increased effectiveness*) and Return on Investment (*ROI-increased efficiency*) from the development (*integration*) and application (*implementation*) of the proposed AI approach.

**\*\*NOTE:** This is a notional table and the level of importance associated with each question may differ given the specific emphasis applied to each and the use case being assessed. Assign Points based on the Attribute Importance Rank (with suggested weights). You may adjust the weight of questions as they apply to your use case.

(5 – critical, 4 – very high, 3 – high, 2 – moderate, 1 – slightly, 0 – not at all)

### 1. Does the use case clearly, and accurately describe the problem to be solved?

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

### 2. Does the use case accurately outline current processes in place?

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

### 3. Does the use case align the goals and objectives with desired outcomes?

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

### 4 Does the use case need greater insight from the data?

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

### 5. Has sufficient data been identified for the use case?

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

### 6. Does the use case identify what data is required and available, accessible, and accurate?

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

### 7. Is the data from the use case annotated and curated? (Does the data contain metainformation?)

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

### 8. Does your use case largely need manual process automation? (That is to determine if only RPA is needed)

0 (Not at all)	-1 (Slightly)	-2 (Moderate)	-3 (High)	-4 (Very High)	-5 (Critical)
-------------------	------------------	------------------	--------------	-------------------	------------------

**9. Is there a predictive element to the use case? (Assumptions and testing made based on prior data)**

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

**10. Have other technologies successfully been applied to address elements of the use case? (Could you somewhat solve your use case with an existing solution?)**

0 (Not at all)	-1 (Slightly)	-2 (Moderate)	-3 (High)	-4 (Very High)	-5 (Critical)
-------------------	------------------	------------------	--------------	-------------------	------------------

**11. Does the data fit for purpose (*descriptive modeling*) and is it operationally relevant (*predictive modeling*)?**

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

**12. Are the authoritative data sources of the use case, organized, structured, deconflicted, and matriculated?**

0 (Not at all)	1 (Slightly)	2 (Moderate)	3 (High)	4 (Very High)	5 (Critical)
-------------------	-----------------	-----------------	-------------	------------------	-----------------

**13. Could the result of the use case change how conformance requirements need to be applied? (e.g., personally identifiable information [PII], classified etc.)?**

0 (Not at all)	-1 (Slightly)	-2 (Moderate)	-3 (High)	-4 (Very High)	-5 (Critical)
-------------------	------------------	------------------	--------------	-------------------	------------------

**14. Does the use case contain ethical considerations and is there a potential for bias? (In the data, algorithms, or aggregation process)**

0 (Not at all)	-1 (Slightly)	-2 (Moderate)	-3 (High)	-4 (Very High)	-5 (Critical)
-------------------	------------------	------------------	--------------	-------------------	------------------

## Questionnaire Results:

The total score will serve as a preliminary assessment for those considering an AI approach. While useful, this is still only a guide for consideration and further investigation. Thorough engineering analysis and practices should still prevail.

### Assessing Your Score:

In order to assess the applicability of an AI approach, the total score will guide the reader whether an AI approach would be beneficial (high score) and where it is less likely (may still be applicable but needs additional scrutiny).

#### If your score is 18 or below:

A score of 18 or below typically represents a small ROO/ROI and limited applicability from an AI approach. Consider that while the score may be low, your situation may still warrant deeper analysis as there can be a compelling reason to continue with an AI approach that did not fall into the standard categorization.

#### If your score is between 19 and 40:

A score of between 19 and 40 could typically be supported with an AI approach but is not an overwhelming natural candidate. These situations can have powerful reasons that can still drive an AI approach, yet they might also have mitigating factors that make a traditional approach a better alternative. In these situations, a more thorough analysis is typically needed.

#### If your score is 41 or higher:

A score above 41 typically represents a compelling ROO/ROI and strong applicability that would benefit significantly from an AI approach. It is strongly recommended to consider the costs and benefits of an AI approach in these instances while still considering other additive and

mitigating factors in the organization, strategic direction, interdependencies, and related items.

## Key Considerations

With the word AI being used everywhere, it is important to separate reality from hype when it comes to which use cases can actually benefit from an AI solution. Consider the following advice and best practices when evaluating AI for any use case

An essential resource to any organization considering AI is the GSA *Emerging Citizen Technology Atlas*<sup>2</sup> which provides a clear snapshot into potential use cases and programs.

**Demonstration of Capabilities/Minimal Viable Product (MVP):** Set goals and objectives for each AI use case by defining a schedule for the MVP/POC demo.

- Establish a high-level framework to prioritize the use cases based on the assessment questions:
  - Data availability/completeness
  - Business value/outcome
  - AI technology maturity
  - Compliance assessment (legal, regulatory, etc.)

- Confirm exploration of the use case scope against the knowledge repository of the organization for re-use or lessons learned to benefit from any previous application of the AI.

**Set Your Foundation:**

- Introduce incentives to encourage workforce innovations to spark use cases partnerships across functions.
- Create governance objectives that empower vs restrict use case exploration.

**Business Capabilities and AI Capabilities:**

- Consider mapping your business capabilities to your AI capabilities.
- Publish a list of existing capabilities across the organization.
- Create blueprints that maximize usage of their existing capabilities.

**Build AI Architectural Blueprint for Future Phases:**

Develop a vision and a plan for the additional requirements and challenges that will need to be addressed if your solution moves into a prototype phase and subsequent operational pilot phases. This should encompass modernization and integration with legacy systems in consideration of infrastructure requirements to host AI applications. Currently FedRAMP Authority to Operate (ATO) accreditation of the AI application is a viable option<sup>4</sup>. Additional options should include the necessary activities essential for major change management components. The viability of these opportunities should take into account policy, process, operational, and cultural requirements.

**Building or Taking an Inventory:**

Current algorithms, dashboards, questionnaire/checklist objectives statements, and computer macros across an organization can be used as the basis to understand how AI might map to business processes.

**Build Once, Use Many:**

- Leverage innovative partnerships with focused or niche domain players that can contextualize exploration of use cases and accelerate MVP/prototypes.
- Use mature tools that can integrate with existing technology stacks to minimize your technology debt.
- Focus on the use case ability to exploit or maximize the value proposition/ROI.

Ultimately, the organization should examine the desired technologies and subsequent capabilities that can be enabled by the future state AI solution. Building a working blueprint of the technical architecture presents a powerful tool for defining the scope and phases of a comprehensive AI implementation. Strategic scaling will enable organizations to optimally address pain points and align stakeholders while tackling one priority area at a time. This will ultimately accomplish the transformational objectives that advance mission goals.

### Emphasize ROI and Benefits:

Emphasize ROO/ROI while making an assessment. Examine the solution's common costs/benefits to provide increased effectiveness and deliver more efficiencies from their AI solution. Include design thinking based on personas and a prioritization matrix around value versus complexity. A MVP should prove viability of an AI solution with ROO/ROI measures to ascertain potential operational gains and resource savings in effectiveness and efficiencies respectively. Also important to consider is the reduction of risk in their ability to meet their mission goals. Ultimately, the ROO/ROI considerations should include:

- **Gains in effectiveness of productivity** – Effort and cost currently utilized on reconciliation to determine the impact on ROO provided by an AI solution, when exchanging data or assets.
- **Gains in efficiency and cost savings** – Effort and cost it currently takes on reconciliation to determine the ROI provided by an AI solution, when exchanging data or assets.
- **Incremental gains** – Implementation in small increments, keeping to a true agile methodology. This is not a lift and replace but a gradual shift to a strategically-assured, positive ROO/ROI.
- **Cloud first and shared services** – Provides for an agile service delivery model which is more adaptive in nature with low entry cost and more consistent delivery of productivity over time.
- **User experience** – Seamless interface and ease of use by users to derive the benefits of the AI experience.
- **Reducing risk** – Understand the ROO/ROI that AI can provide as a result of the reduction in risk.

### Incorporate Regulations/Mandates:

AI has the potential to traverse large swaths of data and generate new forms of data aggregation requiring impact assessments against standards such as National Institute of Standards and Technology (NIST)<sup>5</sup> along with other legal and regulatory considerations (*GDPR*<sup>6</sup>, *HIPAA*<sup>7</sup>, *Personally Identifiable Information, Data Sensitivity*). Organizations should review the use case to understand the application of standards around use of AI and develop risk management plans around underlying technologies that support the use case as it relates to the ethics, mission goals, and business objectives. As organizations seek to establish levels of governance and enforce assessment standards that drive new outcomes, the goals and objectives achieved by the AI use cases should be reviewed at each phase and iteration to assure existing regulatory, legislative, and policy guidance surrounding the use cases are being fulfilled.

## Key Activities

### Management

- Establish an AI inventory and definition set for your organization:
  - Engage executive sponsor and key stakeholders from different functional domains (*missions/business, finance, HR, IT, etc.*) and explore use cases within each domain.
  - Organizations may be at different levels of maturity with regards to their use of AI. It is important to know what capabilities may already exist within the organization and ensure an established inventory of AI technologies and common use cases has been captured to provide a baseline and perspective for the assessment of AI and its applicability for a specified use case.
  - For early adopters of AI who may be uncertain of its applicability for a specific use case, the Primer can help organizations ensure there is a common understanding of the appropriate AI terminology, frameworks, models, and lexicon so that attributes of AI's components can be deconstructed and assessed as part of the use case alignment.
  - By establishing a common AI inventory and standardized definition set for the organization, the reusability of assessments for use cases can offer an enhanced benefit in helping agencies revisit and accelerate their assessment and progress through later stages of the playbook.
- Capture the needs and the use cases of the problem statement:
  - Establish preliminary priorities of use cases based on benefit, data/technology readiness, etc.
  - AI offers a wide variety of opportunities to solve organizational problems and answer important questions with predictive and qualitative elements using cross functional structured and unstructured data elements.

### People

- Define who will use AI:
  - The adoption of emerging technologies generate a ripple effect in the organization. Identifying individuals and user groups impacted in the creation, operation, maintenance, and benefit from an AI solution shape the stakeholder landscape.
  - Identification of users, stakeholders, and populations affected should follow an established practice and protocol, so that categorization and identification can be reliably replicated for multiple problems as candidates for an AI solution.
  - For the specific candidates for an AI solution, defining the necessary functions with the help of subject matter experts, process owners, and support organization will identify the stakeholders essential to coordinate and collaborate the implementation of the AI solution.

- Workforce readiness (knowledge/capabilities/skills):
  - Successful adoption of an AI solution ultimately depends on the workforce integrating the advanced technology into established business processes. A comprehensive strategic communication plan is essential to creating a collaborative partnership essential to taking advantages of AI capabilities.
  - Ensuring a catalogue of available training for the workforce to include options for the skills and capabilities inherent in the AI-driven solution such as establishing a Learning Management System (LMS) that can track and report participation in the coursework to optimize the propensity for knowledge transfer.
  
- Willingness (perception of value benefit versus consequence):
  - The executive level awareness and endorsement of technology is an efficient and effective way to enhance the agency mission and increase the impact of resource investments.
  - Assessment of the critical success factors and lessons learned from prior efforts in adopting technology, as well as the tolerance for changes in established business processes, are predictive tools for the willingness to adopt AI as a viable solution.

## Process

- Map the use case milestones to AI implementation:
  - Align the use case objectives by taking into consideration the underlying business processes tied to specific outcomes (the Primer outlines viable AI solutions that can be incorporated).
  - Remain focused on the problem to be solved, mindful of the available options and resulting opportunities of AI capabilities that can deliver solutions (leverage the frameworks and models of this playbook to monitor and manage progress).
  
- Define the ethical boundaries for AI: Assessing the viability of AI as a potential solution set for the defined problem includes identifying the risk of cognitive, cultural, and computational bias of AI as it pertains to data, algorithms, and aggregation to ensure outcomes are in compliance with ethical considerations:
  - Safeguarding the personal nature of data – identify risk points where the solution intersects with Personally Identifiable Information (PII), either in the defined outcomes or data sets.
  - Autonomous systems parametric framework – identify risk points for the spillover of AI into the larger enterprise architecture through association of data sets and user populations from the use case.
  - Cultural and cognitive bias – identify risk points for unconscious bias embedded in the data sets, the problem statement, or the business process to drive and deliver pre-determined outcomes.

- Impact of AI on the organization:
  - The ability of the enterprise to accommodate a change in respect to advanced technologies is a key factor in assessing the impact of AI on the technology architecture of the organization.
  - The integration impact of AI on the workforce, not only in skills and competencies but also on morale and continuity of mission, has to be included in the assessment of the viability for the adoption of AI.

## Technology

- Assess the sophistication and maturity of AI. Technology should comply with enterprise architecture principles.
- Evaluate the AI solution for its intended use:
  - Using the Primer as a guide for defined AI and their expected results
  - Assess the alignment of the desired outcomes for the resolution of the defined problem statement to ensure the fitness of AI capabilities for the prescribed solution
- Identify the capability distinctions, differences, and differentiators of the available AI solutions.

## Acquisition

- Capture solution metrics (value and outcomes):
  - Produce desired value propositions to help the organization respond with viable, sustainable, and effective options that capitalize on opportunities to solve current problems.
  - Compare to current and past use cases to ascertain the available options to procure capabilities.
- Define your constraints (cost and schedule):
  - Acquisition can look to determine what is available within the agency's own contracted services.
  - Explore existing contract vehicles can reveal others within the organization that are utilizing AI capabilities.
  - Identify current and past AI services through Shared Management Offices<sup>8</sup> can determine “Best in Class” contract vehicles (*OMB M-19-13<sup>9</sup>*).

## Key Outcomes

### Engaged

**At the executive level**, integration of AI use case is endorsed to facilitate the overarching potential to fulfill the strategic intent.

**At the program/mission office**, confidence that AI integration is possible and that an AI solution will enhance the outcomes of business processes and meet all ethical requirements.  
**In the general workforce**, acceptance of AI capabilities, preparation for future skills, and competency development to achieve the desired outcomes.

### Defined

Careful consideration of the problem statement, in the context of the underlying business processes, will afford the means to appropriately apply existing technology within the environment that assures AI solution are appropriately applied.

### Phase Outputs

The following artifacts generated during the Assessment Phase support the Organizational Readiness and the subsequent phases.

#### Business Need

- Capture the stakeholder vision of the desired operational end-state.
- Identify the deliverables necessary to achieve the ascribed goals.
- Document the use case 4P's:
  - **Problem:** What is the negative impact of the current system?
  - **Process:** What are the steps that created the problem?
  - **Potential:** What are the preferred outcomes or alternative end-state?
  - **Proposal:** What solutions are available to resolve the problem?
- Outline the necessary support that must be leveraged during the Readiness Phase.

#### AI

- **Technical vision:** Examination of the agency infrastructure, as well as options for external and shared services used in assessing the viability of the AI solution which informs the platforms and infrastructure readiness activities
- **Non-functional requirements:** Identification of associated data sets, underlying business processes, and workforce capability that are essential in the adoption of AI capabilities and their associated boundaries and readiness assessment activities

#### Business Need

- **Valid Use Case:** Identification and validation of the use case against the defined problem set narrows the readiness assessment activities vital to supporting a successful resolution of the problem.
- **Future state vision:** Defining the value proposition and anticipated return on operations/investment that provides the framework upon which serves as a baseline for readiness and future adoption phases.

- **Stakeholder analysis:** Through identifying the impact of AI in the organization, as aligned to the internal and external stakeholders, and capturing expectations and reservations, the assessment phase provides the framework for stakeholder engagement and communication throughout the AI adoption lifecycle.

## Decision Gate

Consider the score on the “Do I need an AI” assessment questionnaire as a guide to proceeding with an AI approach.

- If the score is 41 and above, highly recommended to commence the Organizational Readiness review.
- If the score is between 19 and 40, recommended to commence the Organizational Readiness review.
- If the score is between 5 and 18, recommended further review of the scope, the inputs and the assigned weights before determining if the proof of concept is applicable for AI prior to continuing readiness review.
- If the score is 5 or below, recommended the proof of concept is not appropriate for AI.

## Phase 2 – Organizational Readiness

The purpose of this phase is to prepare enterprises and organizations for AI efforts and define key supporting activities to ensure organizational readiness. The structure and activities of the AI Readiness Phase are similar to other emerging technology readiness guidelines or strategy frameworks, such as GSA's Modernization and Migration Management (M3)<sup>10</sup>. However, there are nuances specific to AI that should be understood and considered before an organization undertakes an AI initiative.

### Phase Inputs

This phase leverages artifacts generated from the assessment phase as illustrated in Figure 5, which outlines the inputs of this phase.

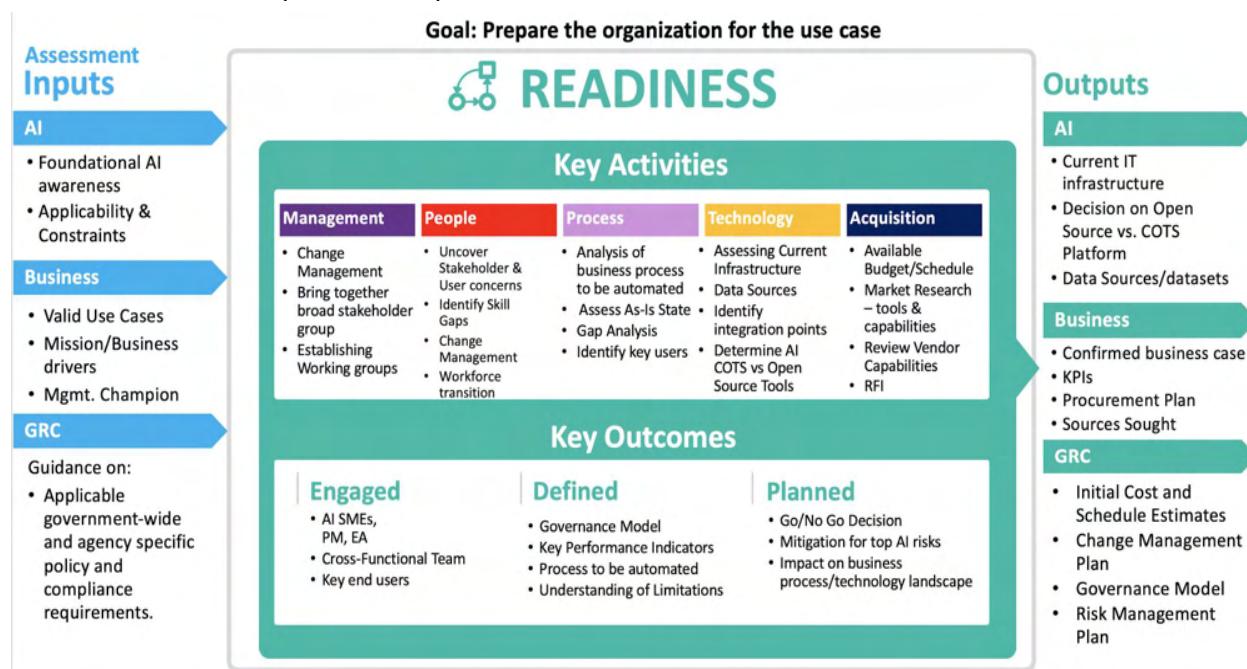


Figure 4: Readiness Phase (2) summary

### Key Goals

This phase prepares enterprises and organizations for AI efforts by defining required organizational capabilities for success. This phase aims to increase the likelihood of success by providing guidance based on best practices and lessons learned in the supporting activities:

- Integrate** AI teams to implement the technological capabilities.
- Define** the scope of the AI solution and processes around it.
- Assess** risks and establishing risk mitigation strategies.
- Ascertain** existing systems' integration readiness.
- Analyze** selected key performance indicators' (KPI) evaluation readiness.

## Key Participants

To help ensure success, the key participants listed below must be identified and engaged throughout this phase:

- Product owners/managers who undertake overall programmatic of the Readiness Phase
- AI subject matter experts (SME) who may or may not be from the agency initiating the program
- Subject matter experts from lines of business and systems with potential AI integrations
- An enterprise architect who is well versed in the current platform topology
- A data architecture knowledgeable on the applicable AI frameworks and models
- An information systems security officer/engineer to ensure access compliance
- An Information manager who can ensure privacy policies and compliance requirements

The participants may include other stakeholders in part or whole of this phase. For example, end-users could serve as the voice of the customer during requirements discovery and definition. Legal and HR teams can be included optionally to ensure policy compliance.

## Key Considerations

AI can start with a minimum viable product (*MVP*), prototype, or pilot program which can be scaled across the enterprise. However, the following are some key considerations that AI evangelists, Chief Information Officers, and Enterprise and Data Architects should consider as part of the Readiness Phase activities.

No.	Consideration	Description	Analysis	Takeaway
1	Readiness assessment	For the selected MVP, assess the people, process and technology readiness for the use case(s) that disrupt least number of business touch points but yet have highest scope of improvement	Starting small allows for demonstrating an emerging technology like AI to be refined rapidly. It enables the stakeholders to get the first-hand experience and allows the high-level concepts to become tangible. Choosing and assessing the readiness of a process that primary stakeholders have complete control of can allow for better governance, faster change management and ROO/ROI assessment. In the Readiness Phase, assess the readiness for the smallest scope by building the constructs to get prepared for technology expansion	Assess the readiness for a process that is controlled end to end for demonstrating the highest value
2	Change management approach	Have a user-centric change management approach, starting right from project initiation	Change management for emerging technology areas should not be considered as only a post-rollout activity. It should be a key factor considered for all the phases from assessment to production. The way impacted users understand, learn and adopt the solution becomes the most important factor for	Change management should be the top priority to maximize learning and adoption of the proposed solution

No.	Consideration	Description	Analysis	Takeaway
			demonstrating the key benefits of the platform over time.	
3	Project management approach	Decide on a project management approach that allows management of all the network participants and their activities	While setting up the AI PMO and governance processes, it is critical to decide the project mgt approach for the initiative. Given that AI technology is still emerging, agile development would be favored over traditional waterfall, spiral development, or iterative approaches. Agile allows cross-functional, cross-partner teams to remain continuously involved in product development. This aspect is also critical for success of any AI initiative, given the number of participants and responsibilities	Agile product management is best suited to ensure continuous stakeholder involvement and response to continuously changing landscape
4	Consortiums	Join or, in rare cases, create consortiums of members that have common goals	AI ecosystems typically involve multiple parties in an industry working together in a consortium to support and leverage an AI platform. It is often better to choose the consortium and become a participant once an organization has assessed its use cases and scope	Best to be part of an industry consortium to get maximum benefits from a given AI ecosystem
5	Enterprise integration	Determine the context of the AI system	For most enterprise use cases, AI technology will be part of the core infrastructure and should be able to integrate seamlessly with other legacy systems	Create a concept of ops to propose solution of context/vision
6	Value transfer risks	Identify and manage value transfer risks for the value transfer use cases	An AI solution needs to manage the risks that were being handled by the central intermediaries whom they aim to eliminate. These include fraud detection, key management, asset security and other risks associated with the value transfer network	Risk management of people, process, and technology to create partial risk guarantee for security, fraud, and costs
7	Consensus mechanism	Define the consensus mechanism	Readiness Phase activities should include rethinking conceptual models for Interagency Agreements and/or Memorandums of Understanding/Agreement to shift away from a centralized security approach which may need education for information security & procurement teams to understand complexities & evolving needs of AI security	Create common understanding on consensus and security mechanism and corresponding participant liabilities and responsibilities
8	Performance expectations	Establish pragmatic performance in terms of metrics	AI are not a replacement of traditional high-performing systems, but are complementary technologies meant to solve different problem domains/use cases	Create realistic non-functional requirements for AI capability
9	Framework-based design	Est guidelines for an AI technology framework that is modular, reusable and extendible	The technological landscape is fluid. Projects based on today's solutions will have to be reworked or re-implemented onto the eventual leading platforms in the future. Consider government wide initiatives using a shared	AI is still an emerging technology. Aim to create modularity and re-use of capabilities

No.	Consideration	Description	Analysis	Takeaway
			services/platform approach and open source software.	
10	Cross-functional team	Establish a cross-functional government team	In addition to enterprise IT and business and functional teams, AI initiatives must engage with customers in this phase. The governance team must ensure to engage risk management, regulatory compliance, IT operations, HR, legal teams, etc. to ensure that the requirements of these stakeholders are recorded appropriately	Create commitment, partnerships, & draft charters to identify tools that support inter/intra organizational development
11	Talent management	Define the skillset and training needed to implement and maintain AI initiatives	Organizations will need experienced IT talent who can implement and maintain AI solutions, as well as support network participants. Government agencies may have to rely on technology partners and third-party vendors who have a working knowledge of different AI ecosystems	Consider training and developing internal talent for continuity while leveraging external talent
12	User experience	Establish user-centric design guidelines	AI is generally considered a backend technology which end-user facing systems rarely see directly. That may or may not be true for all the use cases. Other than the underlying code and algorithm, every user touch point must be designed with user-centricity focus. All users – such as backend, administrators and enterprise users – should get the same quality of experience as the end users. Laying the ground rules for design right from Readiness Phase helps in enterprise-wide adoption in the long run and in covering all the non-functional requirements, such as privacy, confidentiality, security and personalization	User experience is critical for enterprise-wide adoption across every user touchpoint. Iterative agile approach with product ownership and Lean UX techniques can be utilized to ensure the best user experience
13	Emerging tech specific risk management	Understand the agency's risk appetite and plan, communicate, mitigate and discover risks continuously	Agencies that do not accept risk may not be willing to be involved in AI, as this is an evolving technology. Risks related to emerging technology must be managed as the top-most governance activity for such agencies	Manage risks with the focus on change management, technology immaturity, availability & sustain skills
14	Expansion strategy	Create an implementation strategy that allows the program to expand in a risk-controlled manner	Starting small in a controlled business process with high impact on the day to day transactions of end users is the best strategy. But that should not mean postponing the strategy planning and design for expansion, its associated risks, and mitigation strategies	Start by ID options, acknowledge opportunities, & communicate implementation gaps/integration requirements

## Approach Guidance

In most cases, the Assessment Phase will precede the Readiness Phase to ensure use case selection and business relevance for the effort has been determined. Although rare, some government agencies may have assessment and readiness phases running in parallel. This may occur when an agency has already completed a proof of concept and is planning for a larger project based on the proof of concept or integration with an external agency that has already implemented AI.

## Readiness Checklist

When assessing organizational readiness to participate in an AI project, it is important to review your entire organization and ask critical questions designed to assess all possible risks. This will ensure there is sufficient visibility on how to best proceed.

- Organizational capability to execute the project:
  - Do we have a clearly defined use case and problem statement?
  - Is AI confirmed as the right approach for the problem? How else could it be solved?
  - Do we clearly understand our approach? Do we want to build a prototype or pilot?
  - Is there a clear definition for success and measurement methodology and targets?
  - What is the adaptability/scalability of solution for future changes in mission objectives?
- Acquisition:
  - Portability of model and AI solution components
  - Ownership of code and/or IP rights
  - Visibility into model, methodology, and results
  - Most appropriate acquisition path for solution
- Governance:
  - Who will lead?
  - What authority do they have to execute changes?
    - From a policy level
    - From a technology level
    - From a business process level
- Identified capability requirements:
  - SME (internal or need to contact)
  - Technical (internal or need to contract)
  - Talent (internal or need to contract)
- Availability of data (content, quality, rights)

- Identified resource requirements and commitment:
  - What is the budget estimate?
  - What is the strategy for getting budget approval (e.g., connect to the mission, an administration priority, compelling budget demand)?
    - Sustained budget availability
  - Was there a review of the initial vs. total cost?
    - Initial development, operations, and maintenance, as well as fully scaled capability, cost?
  - What is the expected schedule?
  - What key personnel are needed? (especially those who are matrixed)
- Stakeholder identification - traditional and subject-specific:
  - Is there an understanding of stakeholders and their interests (positive/negative)?
  - Is the following known:
    - The expected level of stakeholder involvement, commitment and influence (especially how this influence can work for or against the project)
    - Do we understand stakeholder (and user) attitudes as it relates to experimentation, risk, failure (learning vs. compliance), and automation vs. augmentation?
    - Is there a strong executive sponsor?
- Identification and communication of risk factors:
  - Traditional
  - Ethical, bias, privacy
    - Management of risks - Mitigate, transfer, or accept
    - Compliance or other policies conflicts - Difficulty in resolving
    - Unintended consequences - Business process/ tech impacts
- Has this or a similar approach been done internally or at another organization to improve the ability to execute, gain buy-in, have relevant lessons learned?
- Change management
  - Two-way communication between stakeholders
    - **Send** - Status, buy-in, and adoption
    - **Receive** - How the change is being felt
    - **Feedback** - Should the initiative adjust
  - Training on AI and broader mission (near term and long term)
  - Insights and learning (more broadly)
  - Technology (feedback loop)
  - AI model (feedback loop)
  - Intended impact on business processes

- Workforce adaptation (skills, quantity, transition)
- Technology
  - How much customization is needed to address the use case?
  - Are we taking a data or use-case centric approach for AI?

## Key Activities

Activities in this phase vary depending on the type and scope of selected use cases. Below is a notional activity guideline to prepare an organization for AI implementation.

### Management

Stand up an AI Program Management Office (PMO) and governance office:

- Establish authority for the oversight and management of the proposed AI solution
- Establish change management processes
- Identify stakeholders to form a working group
- Establish AI solution oversight and management practices including meeting cadence, reporting content and audience, and escalation procedures for addressing issues beyond the authority of the AI solution PMO
- Identify applicable Enterprise Architecture (EA) guidelines
- Develop procurement planning
- Identify key skills and resources required for the AI solution
- Confirm the mission/business drivers and value assumptions created during the Assessment Phase

### People

- Identify genuine stakeholder and user concerns
- Identify workforce skills gaps
- Analyze workforce transition process

### Process

- Analyze business processes to be automated/augmented
- Identify business rules/heuristics/mental models of the problem
- Identify key users of the process
- Assess As-Is process and gaps

### Technology

- Assess current infrastructure
- Map user needs to data sources
- Identify interface and integration needs

- Consider the relative strengths and weaknesses of proprietary vs. open source alternatives for tools or platform selection relative to the characteristics of the AI solution

## Acquisition

- Compare initial cost and schedule estimates to most likely resource availability
- Perform market research to identify potential candidates for required tools and capabilities
- Identify potential vendors for services that may need to be acquired
- Issue one or more RFIs to support information gathering if needed

## Key Outcomes

To ensure an organization's readiness for an emerging technology like AI, several internal and external factors have to be assessed and new areas need to be defined and established. The list below highlights definitions and high-level plans, which are further refined in subsequent phases and throughout the lifecycle of the initiative, resulting from the Readiness Phase:

- Identification of the key SMEs and network of AI solution participants.
- A defined governance model for managing the development and implementation of the proposed AI solution as well as communicating progress and elevating issues outside the authority of the AI solution management governance.
- Clarity of the use case to be addressed.
- Identification of risks, constraints and limitations.
- Risk mitigation approaches for the most significant AI risks for the solution.
- Identification of intersection or integration points with existing people, processes and technology.
- Formal review of the environment and factors within the assessment and readiness phases to determine whether the proposed AI solution is reasonably viable.
  - This should be in the form of a formal decision gate (GO or NO GO).
  - AI is a new field with new technologies and challenges some of which may not be fully understood.

## Phase Outputs

The following artifacts generated during the Readiness Phase support the Selection Phase and the subsequent phases. They should be leveraged to ensure the alignment to initial vision, continuous discovery, monitoring and mitigation of risks and continuous feedback to the stakeholders for forthcoming implementations:

- Current IT infrastructure: cloud access and infrastructure, Secure Data pipelines
- Current authorizations: authorities to operate (ATO), policies, and DevOps practices
- Target state concept of operations (CONOPS)

- Confirmed business case which can be solved with an AI solution
- Determination if use case is unique and requires custom solution or can be evaluated against commercial off-the-shelf (COTS) solution
- Identified data sources and datasets necessary for solution with awareness of their current state, location, and access
- KPI and measurement baselines and defined outcomes for success
- Initial Cost, Schedule Estimates, and Procurement Plan
- Established Change Management Plan with communication strategies to address top risks
- Risk Management Plan for top risks identified during the evaluation
- Request for Information (RFI) to leverage industry insights, proposals, approaches and solutions

## Decision Gate

After a successful Readiness Phase, which includes a review of model outcomes, impact assessment of any changes, and proper evaluation of risks, it is time to make a go/no-go deployment decision. If progress has stalled prior to reaching required accuracy, the task is to assess the problem, decide whether to revisit the model design or some aspect of the solution, and even reconsider the original business objective. All stakeholders should have a joint understanding of the responses to the questions highlighted in the Decision Gate section.

### Proposed system:

- What are the key business capabilities of the proposed system?
  - How does this solution align to their mission
  - Is there already an available solution that does the same thing?
- Is this project in their area of expertise?
- Who are the key participants in the proposed AI initiative?
- Who and what will be impacted? What are their roles? What will be the impact?
- How will the onboarding/separation happen?

### Strategy and governance:

- What is the proposed governance and management structure?
- How will this project impact existing policies?
- Will it be necessary to rewrite policy for solution to work?
- What are the key technological, business context, security, performance, user experience, program management and governance related risks specific to the proposed AI solution?
- Are all key stakeholders aware of these risks?
- How will key risks be managed?
- Are KPIs defined and baselined?
- Is there an estimated timeframe and resources required?
- Does the initial schedule and estimated cost allow for agile product development?
- Can the timeline and cost can be recalibrated based on ongoing learning?

- What is the procurement strategy for the proposed program?
- How will security be managed?
- How will change be managed for the impacted people, processes, and systems?
  - Who needs to be informed and when?
  - What systems are in place for enabling two-way communication?
  - What people and groups in our organization are most likely to be resistant and why?
  - What specific processes or workflows are impacted by this change?
  - How do we plan to communicate this change effectively?
- Do we have any missing information to advance to selection? If so, what is the plan to identify the missing information?

At the end of a successful Readiness Phase, the stakeholders should have a joint understanding of the responses to the questions highlighted in this section.

## Phase 3 – Solution Selection

Selection comes after the Readiness Phase and the organization has determined that it has the necessary requirements to implement an AI solution. This phase focuses on selecting the right tools, policies, people, etc., for a successful implementation of the AI project. This section is critical in setting the agency for a successful implementation of AI technology. Selection of the right AI solution for an agency will depend on several factors that will be discussed here. This section will attempt to shed some light on key considerations that should be evaluated before implementing an AI solution.

Main outputs from the Selection Phase are:

- Selection of the right AI technology for the use case.
- Understanding your organization's capabilities to support AI implementation.
- Aligning your organization's acquisition strategy with high-value AI solutions.

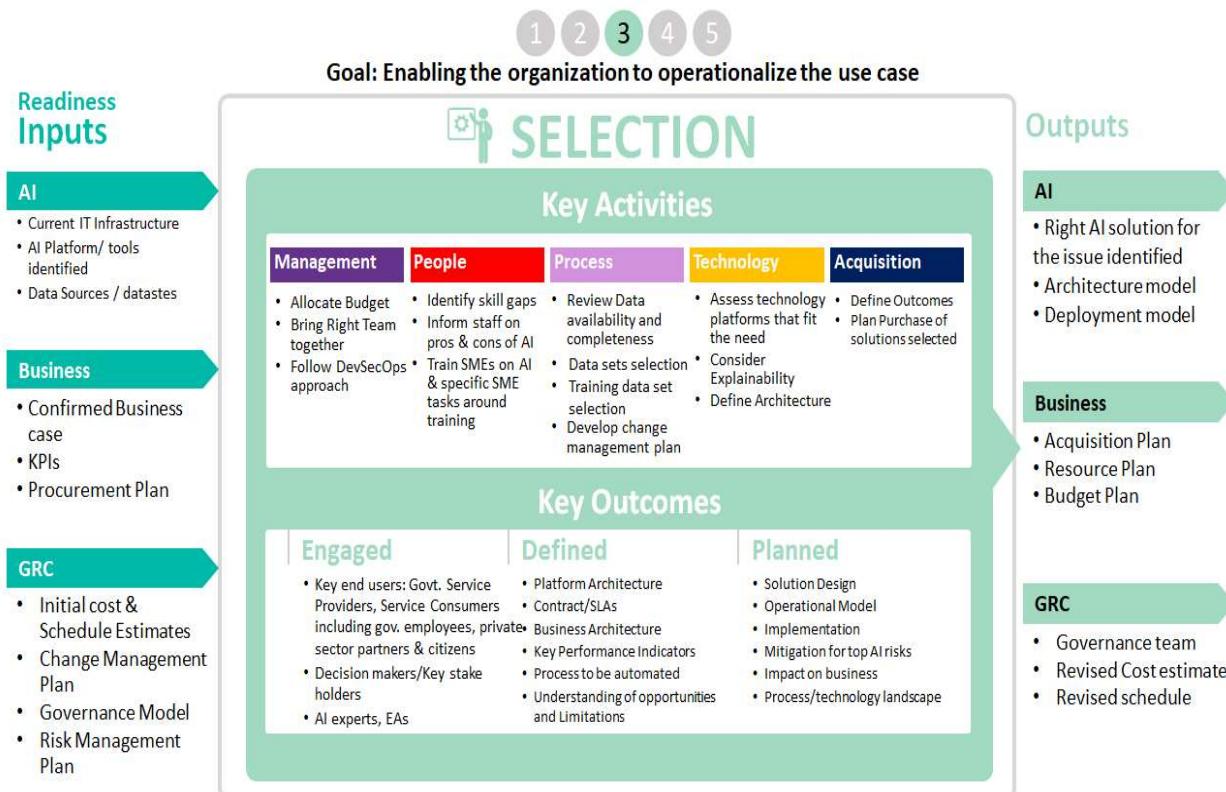


Figure 5: Selection Phase (3) summary

## Phase Inputs

The inputs for this phase will take the outputs from the Readiness Phase to expand and explore various concepts and categories that should be considered for a successful selection of the right technology solution for the selected use cases. Selection of an AI solution will have to address

business and technology considerations. **Business consideration** involves clearly defined business problem statements and requirements, selection of the right team and skills to ensure successful implementation, availability of acquisition vehicles, and budget to implement solutions. **Technology consideration** involves selection of the right technology tools for the selected use case. AI is a vast field of technology so the selection of a solution should follow a methodical due diligence process. It is also imperative to make sure that complete and quality data sets required for building the AI models are available. The previous phase should provide the verification needed to move into the selection section. Business needs and outcomes for the AI solution must be clearly identified before embarking on selection.

## Key Goals

The goal of this phase is to identify the solution that best fits with the organization's use case and outputs of the Assessment and Readiness phases. This can extend into selecting training, and worker types and skills. It should also address which policies need to be considered in choosing the use or procurement of AI.

Key goals also include the selection of the appropriate procurement contract vehicles, tools, and support specialists to set the organization up for a successful implementation by providing the right recommendations and tools. As a result, the organization can proceed into implementation based on the outputs from this phase.

## Key Activities

### Management

Management of the selection of the organization's use case and AI solution will accomplish these activities:

- Allocation of appropriate budget for the selected use case aligned with the organization's strategic priorities.
- Recruitment of an effective project team (e.g., experience and skills).
- Creation of a governance process that will evaluate and oversee AI project selection.
- Establishment of the AI Governance council to support the following:
  - Legal, regulatory, and compliance review to set up accountability with AI outputs.
  - Scientific verification and validation to confirm the AI algorithm has been tested on a valid data set.
  - Ethical evaluation and usage guidelines to determine whether or not and to what extent stakeholders are informed about the role AI is playing in the business scenario.
  - Organizational deployment and change management for training staff on what is expected and the correct actions to be taken when using AI.

- Development of timeframes according to an AI project management roadmap associated to the organization’s performance plan.
- Selection of the software development methodology and process for AI development.
- Ensuring the right team members are involved based on the methodology and the process selected.

## People

Preparing people is as essential as preparing data. It requires organizations planting the seed of an AI mindset. The following “seeds” or anchors for such a mindset reflect three universal truths about AI. These should serve as starting points for effectively building people’s understanding, engagement, and role in an organization’s AI journey.

### Diversified AI

AI must be built by people who understand the business and domain problem (not solely the technology). Data scientists and engineers are key for tooling but front-line people often dictate AI’s success. When thinking about “diversifying” AI development, consider those responsible for the day-to-day administration of the workflow in question (e.g., customer support agents, security admins, doctors, field technicians, etc.). No matter the application, it is the domain experts and end user employees who best understand where the breakdowns occur, where products fall short, where they spend most of their time, and where customer sensitivities lie. Moreover, the gravity of some applications (e.g., credit scoring or medical treatment recommendations) demand diverse perspectives, multi-disciplinary expertise, and workflows to monitor domain dynamics.

### Directional AI

The use of AI in the organization takes on a new meaning because data is analyzed through deep uses of algorithmic programming. AI is a complex predictive analytic that combines the relationships between databases and desired situational learning. The AI programming analysis enhances decision support, informed continuously by new data sources, APIs, integrations, regulations, techniques, security reconfigurations, and cultural changes. In fact, the role of “AI managers” may well be an emerging job function and one worth investing in upskilling. The key accomplishment of using AI is to engage human end-users with the benefits of greater, previously unknown insights, enabling the use of human judgments, and more comprehensive knowledge.

### Democratized AI

AI technology has matured over the years and with the advent of cloud-based solutions, it has become easier to implement basic AI capabilities without the need for highly skilled technical personnel. With a need for a fully customized AI solution, skilled technical personnel will be necessary to implement a full solution. It is imperative that the right training be provided to all stakeholders involved. Key functions supporting AI’s technical integration include tools with

intuitive and customizable user interfaces; easy-to-use dashboards; interfaces, “self-service” portals; training modules; and even using machine learning or chat bots to help facilitate user experience (UX) and model selection for far less technical user types.

### **Identify key people and ready each group accordingly**

For the majority of employees, readying for AI is not about training on new technology but embracing a new mentality. It is about building trust through education and quelling fears, as well as offering pragmatic ways to play an active role. It is also about conveying the accessibility of AI and articulating people’s roles in the broader picture. This should be framed and delivered as a growth opportunity for your people, giving them the opportunity to contribute to designing systems of the future.

Subject matter experts of all disciplines are also important for AI optimization. In some cases, SMEs could be those best suited to translate highly specialized domain expertise into AI-enhanced decision making (e.g., security, lawyers, doctors, accountants, scientists, etc.). Depending on the sector or use case, other specialized roles such as computational linguists, computational biologists, computational vision specialists, etc., may be necessary skill sets to bring in-house. Ethicists and behavioral scientists also provide critical perspectives when analyzing the designs, workflows, implications, and myriad risks of AI-enhanced processes or AI-defined products.

AI’s success is often driven by people’s willingness to adopt it. Thus, enterprises deploying AI are well advised to assess how people’s sentiments, fears, questions, and insecurities impact their proclivity to adopt. Organizations have an opportunity to use employees’ fears, uncertainties, and doubts (FUD) to pinpoint where and how to support people.

It is important to address concerns of job displacement by educating employees on the limitations of AI; articulating where AI will augment or accelerate human workflows; providing clarity on governance models; and by supporting employee upskilling and continued education programs. Investment in AI equals an investment in people. Early successes in the space show that the sum of human plus AI is greater than either alone. Therefore, business preparedness and investment in AI requires proportionate preparedness and investment in people.

### **Process**

Key process-related activities to operationalize an AI use case are as follows:

- Pick the right algorithm and data sets for the use case. The algorithms and data sets appropriate for one health care related use case might not be the right one for another-related use case.
  
- Employ a variety of AI toolsets and training methods for a problem.

- An important consideration is bias in the data sets and training sets. There are lots of examples where the data and humans building the algorithms had an inherent bias that was built into the AI algorithm.
- How clean is the data – avoid garbage in and garbage out (GIGO):
  - In some instances, you do not want sensitive personal information and need to make sure the data has been scrubbed of PII.
  - A significant step toward obtaining high-quality data is to understand the goals of the stakeholders, collect data from the various stakeholders, integrate different data sets together, and sort out inconsistencies so that the data is accurate and rich.
  - Structured versus unstructured data processes to manage and handle both of these types of data.
- Importance of setting up pilot projects. For AI implementations, it is important to start small and scale-up. This also helps with bias detection and elimination.
- Set up goals that the pilot should accomplish, be open to learning from the pilot, and utilize various methods to achieve this end goal. The pilot will help determine what the long-term project will be and if the value proposition makes sense for the organization.
- Use iterative, agile methods to facilitate the requisite change.
  - Develop a change management plan outlining how the organization manage transitions and transformation.
  - Create a communications plan outlining the goals, objectives and milestones.
  - Establish a mechanism to disseminate guidance to adopt, implement, and train the organization.

## Technology

AI encompasses a vast set of technologies and capabilities and AI solutions ranges from language translation to decision support to self-driving cars and more. It is imperative to understand the need and the technology that fits the need. As mentioned previously, understanding the problem statement and the outcomes expected are critical in the selection of the right technology. Technology solutions can be built using API based technology solutions, Open Source technology solutions, and internally developed solutions. It is worth mentioning that the last option provides the most flexible and nimble approach to the implementation, adaptation, and sustainment of the technology.

The selection of the type of technology to be utilized will need to be determined based on the use case at hand. For example, if there is a need for translation from one language to another, an API based solution might be a good fit. If there is a need for an AI to review the internal contract, policy documents using natural language processing (NLP) to summarize or extract

entities, etc., an open source based pre-trained model might be a good fit. If there is a need to review large amounts of numerical data from a control system and assess the data, make predictions etc., an internally built model might be a good fit. Therefore, it is important to understand the need and the desired outcome before selecting a technology.

In addition to understanding the need and outcome, another factor to consider in selection of a technology is the explainability of the solution. Justification and feasibility in AI involves the ability to understand why a certain selection decision was made. It is important to consider how the selection for AI solutions will accomplish the mission critical priorities. Thus, consideration of which AI technology to select for deployment given the operational need, infrastructure platform, and portal will determine the best eventual solution. These considerations will define the architecture and selection of the solution (e.g., Where will the solution be deployed? What are the security implications of deployment?).

## Acquisition

Acquisition tends to be the ‘long pole in the tent’ because it takes up a lot of administrative burdens to compile procurement packages and maintain thousands of contracts terms and conditions. Having disparate solutions by various agencies can cause data interoperability issues in the long term. The guidance below addresses many of these issues allowing federal agencies to maximize the government’s investment in AI acquisition plus cost avoidance.

Consider the efficiencies that can be gained through the sharing of acquisition strategies and mechanisms across departments and agencies. Examining the potential for shared services will benefit citizen services and save tax dollars in the long run. The Best-in-Class (BIC) acquisition designation identifies government-wide contracts that satisfy key criteria defined by the Office of Management and Budget (OMB)<sup>11</sup>. These solutions are vetted, well managed, and recommended, and in some cases required for use.

Interagency government-wide category teams have worked to designate over 30 BIC contracts to reduce the amount of effort individual buyers spend finding and researching acquisition solutions. Widespread adoption of BIC solutions will:

- Maximize the government’s shared purchasing power so agencies can leverage volume discounts.
- Help agencies operate more efficiently by reducing administrative costs and contract duplication.
- Expand the collection and sharing of government-wide buying data, leading to better-informed business decisions.

## Key Outcomes

To complete the ‘Solution Selection’ phase for an emerging technology like AI, several internal and external factors must be assessed, and in some cases, new areas need to be defined and established. The list below highlights the high-level plans, which are further refined in subsequent phases and throughout the lifecycle of the initiative, resulting from this phase:

- Business stakeholder SMEs and enterprise architects are engaged in developing a scalable solution.
- Business architecture and changes to the business processes are defined and agreed upon.
- Clear identification of the boundary for process automation and analysis of the impact on the overall (lean) business processes
- New business opportunities, limitations, and constraints (if any) are analyzed and defined.
- Mitigation plans are in place for the following risk categories:
  - Technology
  - Business process
  - People
  - Security
  - Performance
  - User experience
  - Adoption
  - Regulatory compliance
  - Enterprise Integration
- KPIs are defined and baselined for the selected business case.
- High level solution architecture design is planned and programmed to produce desired results.
- Enterprise-level platform architecture is defined to scale up the solution in future, if needed
- Implementation controls like estimated budget, schedule, master plans, and required type of resources are planned and programmed
- A procurement strategy is defined and coordinated with the acquisition community
- The revised operational model is defined, documented, and shared among key stakeholders

Many organizations have successfully launched AI pilots, but have not had much success rolling them out organization-wide. To achieve their goals, agencies need detailed plans for scaling up the solution that requires collaboration between technology experts and owners of the business process being automated. Because AI technologies typically support individual tasks rather than entire processes, scale up almost always requires integration with existing systems and processes. Ensure business process owners discuss scaling considerations with the IT organization during the solution selection phase.

## Phase Outputs

The Selection Phase is a critical step in ensuring the right direction for implementation. Analysis of the need, outcomes, technology, process, and people are done in this phase and produces a set of deliverables that the implementation team will use to implement the AI solution. This phase will ensure the right technology, create the base architecture for the solution, and provide recommendations on high level deployment model. From a business perspective, this phase creates a path for implementation by allocating the right budget, bringing the right leaders and stakeholders, and creating acquisition tools critical for success.

## Decision Gate

Following are the Decision Gate questions on solution selection:

### Data related questions:

- Is the data discoverable and available?
- Is the data organized so that it is fit for use?
- Is the data in a structured or unstructured format so it is searchable?
- Does the data need to be cleaned before training?
- Are there vast amounts of training data available to train the model?
- Does the data contain any sensitive information?
- Is there training data available to train models?

### Technology related questions:

- Is the explainability of AI important?
- Where will the solution be hosted (cloud versus on-premise?)
- Does the AI solution need to be an open-source solution?
- How do AI models perform over time? (Do the accuracy and quality of the results increase or decrease?)
- Is it scalable to meet current and future requirements?
- What best practices to leverage (CI/CD, DevSecOps, Agile, containerization, etc.)?

### People related questions:

- Is there internal talent available to accomplish the selected technology?
- Should external talent be utilized to accomplish the solution using selected technology?
- Do SME's and other stakeholders understand the scope and implications of AI?
- What are the short term and long-term skill needs and gaps to develop and implement the AI solution?
  - What strategies to address such skills demand?
  - What change management is in place to drive adoption and who lead the change?

**Management and acquisition related questions:**

- Are there contract vehicles available to accomplish the selected technology?
- Is there budget allocated to accomplish the selected use case?
- Are resources programmed to sustain the AI project?
- Is management onboard with the solution that is being presented?

## Phase 4 – AI Implementation

In the Implementation Phase, the inner workings of the selected solution are completed and tested. It closely examines the technical implementation of the components of AI, as well as the operational aspects, such as governance and security posture, to ensure the optimal operations of the AI solution. Key activities and outcomes for management, technology, people, process, and acquisition are also examined. Figure 6 below provides a phase summary.



Figure 6: Implementation Phase (4) summary

## Phase Inputs

To proceed with implementation, leverage the outputs from the previous Selection Phase. These typically include an established Conceptual AI Architecture, Operational Model, and Development Lifecycle. The business must have a Business Architecture, Resource Plan, and defined Acquisition strategy milestones and success criteria. The Governance, Risk, and Compliance (GRC) area must have a revised Cost and Schedule Estimate, Acquisition Plan, and Operational Model. These are described below.

### AI Inputs

**Selected Platform/Tools:** The necessary tools and technologies required to implement the use cases should have been identified. This includes the deployment model (Cloud versus on-premise), language/tool used to develop the algorithms (e.g., Keras, TensorFlow, PyTorch), big data processing platforms needed (e.g., Spark, Elasticsearch), ETL or data cleansing tools (e.g.,

EADEV2), reporting and visualization needs (e.g., Tableau). In addition, any real-time processing needs should have been identified along with the necessary tools (e.g., Kinesis or Kafka).

**Conceptual Architecture:** The overall blueprint of the solution should be finalized, preferably with the results of any proof-of-concepts conducted. This includes the data sources, primary transformations required, AI model used, and target audience.

**Data Sources:** The sources of data that will be used for implementation should be identified. Besides the frequency of data refresh (real-time versus batch), history, data format, storage format, presence of PII, and any accessibility constraints must be identified and agreed upon.

### **Business Inputs**

**Business Architecture:** A high level diagram of the business process, business rules, technology, roles, and responsibilities that, collectively, achieve the business benefits being sought. The Business Architecture should be defined in sufficient detail to develop functional requirements for the solution and to provide clear, verifiable metrics demonstrating success.

**Resource Plan:** A description of AI funding streams, workforce requirements, business initiatives, and technical system dependencies that are likely to impact or be required by a successful solution implementation.

**Acquisition Milestones:** The key products and services required for a successful implementation over the anticipated duration of the solution implementation.

**Success Criteria:** quantified, business defined goals, endorsed by the appropriate level of management, that objectively demonstrate the benefits of the business case is achieved.

### **GRC Inputs**

**Revised Cost & schedule estimate:** The multi-year Lifecycle Cost Estimates (LCCE), Independent Government Cost Estimates (IGCE), Program/Project budgets, and/or Operational Budgets. These estimates may include, but are not limited to, direct, indirect, and other direct costs (ODC) associated with initial implementation; Operations and Maintenance (O&M); & Development, Modernization and Enhancement (DME) costs related to applicable commodities, technical services, and non-technical services needed to implement and maintain the AI solution.

**Information Security, Architecture & Policy Compliance:** The required FIPS rating and associated security frameworks and controls; business, data exchange, infrastructure, software, and system architectures; and list of federal, state, and/or local applicable statutes, regulations, policies, security, and governance compliance requirements.

**Risk Mitigation Strategy:** The analysis, general plan, mitigation strategies, and risk monitoring approach needed to manage and reduce organizational-, portfolio- and/or program-wide business, legal, operational, technical, and security risks. These include, but are not limited to, an inventory of all risk sources, identified risk events, the severity of each event's impact, and probability of each occurrence.

## Key Goals

The key goal of the implementation phase is to take the models, processes, and technologies determined in the previous phases and implement them so they can be integrated within the organization in the next phase. For instance, in manufacturing a cell phone, would be expected that at the end of the Implementation Phase, the subcomponents of the phone will be complete and tested before moving to the Integration Phase of incorporating a particular carrier's service, other applications, accessories, etc.

## Key Considerations

**Acquisition:** All acquisition activities should be performed in accordance with the Federal Acquisition Regulation (FAR)<sup>12</sup>. However, it is worth paying attention to methods of procurement that allow agility and flexibility to truly reap the benefits of AI. For example, usage of cloud-based platforms and technologies necessitates appropriate funding mechanisms to obtain the best benefit offered by these platforms. While a typical pay-as-you-go or subscription-based model using Time and Materials (T&M) type can be considered, this also exposes the government to considerable risk in terms of runaway costs that could endanger exceeding contract funding. On the other hand, purchasing computing resources in advance using options such as “reserved units” creates the risk of unused capacity that offsets the advantage offered by the cloud. As such, it may be prudent to consider a pay-as-you-go consumption-based model that places an upper limit on the usage of resources. This enables limits to be placed on the amount of memory used, the number of downloads, number of service instances, storage capacity, or bandwidth utilization. Doing this provides the organization the benefit of using only the resources needed while still providing the ability to control costs in order not exceed contract funding limits. This kind of model is being used at GSA’s cloud.gov. The pricing model should also be flexible to accommodate different types of environments. For example, the pricing of a sandbox environment can be substantially different since there could be several constraints placed on the sizes of instances allowed.

**Data as a Strategic Asset:** Data needed for AI could be sourced within the organization as well as outside of the organization and can be structured or unstructured. Traditionally, most Chief Information Officers (CIOs) were concerned with organizing and managing structured data collected from internal applications used for daily business processing, as well as external facing websites. This data is typically stored in databases that offer ACID (Atomicity, Consistency, Isolation, Durability) properties to guarantee validity against errors. CIOs should continue to

improve on managing the quality of structured data by investing in the right infrastructure, technologies, and data tools and techniques. The quality of data often defines the organization, and the ability to manage daily operations effectively through evidence-based decision has become crucial. Thus, data has become a strategic asset that empower and inform the evolution of the organization.

**Data Collection:** The era of social media and internet-of-things (IoT) has dramatically increased unstructured data such as text, videos, audio files, and machine logs. The ability to mine this unstructured data is becoming critically important for developing comprehensive analytics and AI applications that can provide competitive advantages to an organization. With the advent of cloud computing and cheaply available storage, CIOs could consider moving to a logical data store that link to data located and a variety of sources and contained in multiple channels. In this approach, data is directly stored in its original format without any transformations and is indexed and organized through tagging and metadata to provide easy access and search capabilities to users. This offers the opportunity to quickly distill traditional extract, transform, load (ETL) tools or custom queries for further downstream processing and storage as a result of structured data. This provides the means to instantaneously access it through well-defined data architectures regardless of the growing mass and complexity of the volume, variety, and velocity of data that doubles every 18-months.

**AI Skills:** To fully reap the kind of benefits offered by AI, organizations need to start investing in understanding core technologies and techniques, especially those around deep learning, their implications, challenges and constraints. This includes understanding the hardware requirements such as graphic processing unit (GPU) based Virtual Machines, and software requirements (such as Python, R) or out-of-the-box technologies (such as RapidMiner.) Organizations need to invest in developing internal human expertise and AI skills to lower costs from hiring expensive data scientists and machine learning engineers. To accelerate this process, one method that can be explored is to hire more recent college graduates who are trained in programming and algorithms and train them to develop machine learning skills. However, these graduates will need to be led by senior staff members who have the requisite experience to provide proper guidance.

**Understanding the Limits:** Organizations need to scope the constraints of AI based solutions. While the benefits are numerous, AI still has several drawbacks, limitations, and disadvantages. One of the biggest issues lies in the area of bias in algorithms. AI based algorithms are only as good as the data that is exposed to them. Any data that is exposed by human beings inherently contains some bias that skews the results. As such, data scientists and AI programmers need to be very careful to capture the correct sources of data which contains representative examples of a wide variety of use cases, so as to reduce the bias in the algorithm. Predictive algorithms in techniques such as Deep Learning and Neural Networks offer very limited visibility into the method or logic for their predictions. This makes it very challenging to use these techniques for

critical applications. The risk can be mitigated, to a certain degree, through comprehensive testing to ensure that predictions are being made accurately. Several organizations invest in months of parallel testing, comparing results obtained by an AI model with actual results, and tuning the model to improve results. Typically, the decision of what accuracy is “good enough” depends on the comfort level of the business stakeholders and the problem in question. The more critical a problem, the higher the accuracy needed. For example, it is ok for Alexa to misunderstand a command given, but it is simply not acceptable for a self-driving car to make a mistake in assessing road conditions. Machine learning engineers need to strive for continuous improvement to improve accuracy and ability to handle different situations and different use cases.

## Key Activities

### Management

**Establish an Iterative Approach:** Implementing an AI solution is an iterative process that requires numerous cycles or attempts to define, refine, and optimize the solution. In predictive analytics use case for example, data needs to be extracted and transformed as needed. Multiple algorithms need to be examined and a prototype model needs to first be developed. Engaging solution architects, change management catalyst, and program managers can and will facilitate efforts to achieve the ascribed objectives, coupled with modeling, managing, and measuring progress. Additionally, ongoing program management reviews (PMRs) will assure those features that need to be identified or existing features integrated or eliminated within the algorithms are addressed in the ongoing efforts to further refine the prototype model. This process is repeated multiple times until the desired accuracy is obtained or until it is determined that the model or data sets (or combination) is not good enough to be implemented in production. The fact that this model may not be used at all is an important distinguishing factor in AI solutions. Thus, establishment of an iterative mechanism to manage this process effectively is crucial. One example of an established iterative mechanism that could be used is the agile philosophy. However, other mechanisms that work for the organization can and should also be considered.

**Change and Communication Management:** This is critical during AI implementation given the uncertainty involved in the outcome of the process. Setting expectations and constantly communicating the consequences of the iterative process requires regular feedback sessions. Change management also becomes a key component to address common workforce concerns such as “Will my job be eliminated” as a result of the initiative. As it stands right now, AI technology is not mature enough to replace humans in the workforce. AI augments rather than replaces personnel. This message should be constantly and consistently reinforced throughout the organization.

**Workforce Planning:** As mentioned in Key Considerations, the key resources need to be adequately planned for ahead of time. There is a considerable shortage of skilled Data Scientists who not only understand the technology but also understand the business. As such, it is essential to plan early, secure, and identify these resources. Other important resources, such as SMEs who understand the problem and help in determining the success of the solution, should also be identified and engaged early in the Implementation Phase.

**Governance and Oversight:** Proper governance mechanisms need to be established to monitor the progress of the project. Due to the iterative nature of AI, it is easy to get consumed in repetitive cycles with no clear progress made, while using up a lot of capital. A programmatic approach, managed by a well-educated and well-informed governance committee, is necessary to ensure they stay on time, on target, and on task. Their experience is critical to understand the AI capabilities and its potential implementation impact. It is through ongoing progress checks of program management reviews (*PMR*) that the governance committee will be able to ascertain if the team is on the right track, needs additional time, have adequate resources, and are empowered to make decisions regarding further investments, as well as have the authority and advocacy to review all aspects of the project.

## People

The success of the AI solution's implementation relies, in large part, to the proper assignment of human capital to fit the requirements of the project. To this end, evaluate available staff to assess their skill levels and match suitable staff to the project. As the project is implemented, continuous skill auditing will be conducted as well as tailored training based on this skill assessment to fill any skill gaps that are identified in the course of implementation. From the outset of this phase, a project manager (PM) should be assigned to ensure that resources are being allocated effectively, skill assessments are being conducted, and most importantly, that adequate talent is properly aligned to assure success of the project. The PM will ascertain if the right mix of technical and business staff are assigned to the project. Continuous skill auditing and training are key activities to ensure that staff is performing at the highest level.

## Staffing Requirements

- **Business Analyst(s):** Working with the data scientists and their team, the business analyst is an expert on the goals of the project and provides ongoing feedback and guidance as the analyst reviews results from the project. This feedback can lead to minor adjustments and tweaks by the data scientists and their team. The analyst is also focused on documenting the process and works in conjunction with the project manager to ensure that the project is moving along at a reasonable pace and conducts ongoing process evaluation to determine if any adjustments are necessary.

- **Project Manager(s):** Collaborating with the data scientists and business analysts, the project management team sets realistic timelines and processes to evaluate progress on an ongoing basis. This involves an agile approach where members of the team are in regular contact with one another and share progress and challenges through practices such as daily stand-up meetings. The project manager provides a consistent follow-up to ensure that the teams are communicating in an effective manner.
- **Data Architect(s) and/or Enterprise Architect(s):** Specializing in the use of and application of data as a strategic asset. These often overlooked members are the key to ensuring the approach of how things are done fulfill the overarching objectives of why the project is being undertaken. They outline and provide the fundamental frameworks which are foundational to ensure that the resulting solution is first and foremost effective and produces efficiencies that afford the organization to benefit from working smarter not merely harder. It is their efforts that are vital to ensure the solution optimizes mission effectiveness.
- **Data Scientist(s):** Specializing in data analysis with a broader skill set than goes beyond deep technical skills, the data scientist understands the business problem and provides ongoing support and direction and how to go about using AI and data analysis to deliver meaningful results. The data scientist also has strong skills in software development and, when development work is required, has the ability to oversee, evaluate, and provide feedback to software developers.
- **Software Developer(s):** Some AI projects will require the expertise of software developers especially if there are significant programming requirements.
- **Machine Learning (ML) Engineers:** Technical staff focused on AI modeling and testing. They work under their direct supervision of the data scientists.
- **Testers:** Software testers which work in close collaboration with ML engineers in testing for defects and bugs. Regularly document testing in reports.

#### **Collaborative Partnerships including:**

The **Chief Data Officer** (CDO) organizes and structures data as a strategic asset to provide the essential facts upon which to make evidence-based decisions.

- Big Data (BD): Operationalizes data by produces an inventory, catalog, and dictionary to make it visible and accessible to feed a Data-Driven-Organization (2DO).
  - COLLECT: Organize WHAT data is be available to discover it and make it visible.
  - CORRELATE: Structure WHERE information is so it can be searched and made accessible.

The **Chief Analytics Officer** creates algorithms to apply the facts in ascribed relational information that can be aggregated to inform answers.

- Analytical Information (AI): Apply technology to automate business processes, gain insights, and customer engagements.
  - CHARACTERIZE: Relate HOW knowledge increases awareness so it can be understood.
  - CONTEXTUALIZE: Answer WHEN understanding is applied to the associated linkages.

The **Chief Knowledge Officer** enlightens contextual dependencies of those influence that affect outcomes of actions taken measured against the results to be achieved.

- Best Initiatives (BI): Explore strategic options that capitalize on opportunities to effectively plan and program to achieve organizational objectives.
  - COGNATE: Depending WHO applies it given their perspective/paradigm/trustworthiness.
  - COST/BENEFIT: Context WHY the alternative impact given the statistical probabilities.

The **Chief Strategy Officer** empowers understanding to identify the causality of options as related to action taken measured against opportunities relative to the desired results to be achieved.

- Creative Impact (CI): Ascertain how to transform organizations to a desired end state. Actualize their potential in order to gain competitive advantage.
  - CONCEPTIONALIZE: Enlighten as to available options given current potential.
  - CAUSALITY: Empower to ascertain the influences that impact available options.

The **Chief Transformation Officer** evolves the requisite wisdom to verify the knowns, validate the unknowns, and venture into the unknowable in a quest to explore questions.

- Matriculation Loop (ML): Think big, start small, and scale quickly to take an iterative approach to evolve as a benefit of a learning environment.
  - CONSEQUENCE: Evolve the resulting outcomes given current potential.
  - C.I.S.: Measure and monitor actions taken against results to be achieved.

## Process

Process related activities ensure that both the business processes being automated or augmented by the AI solution, as well as the supporting and related administrative processes of the organization, are made ready to fully take advantage of the benefits available in the AI solution. By taking a use case approach, organizations can develop a progressive process that provides the necessary results through an iterative process.

As organizations work to organize and structure their data to make evidence-based decisions, the result will illustrate the value of leveraging data as a strategic asset. By identifying the problem, process, and potential, the opportunity to create a programmatic proposal will identify what data is necessary, how it must be structured, and its availability upon which to propose a solution in this 4P process:

- **PROBLEM:** identify WHAT you are solving for and the negative impact it has on the organization.
- **PROCESS:** outline HOW the progressive steps influence it and perpetuate its impact.
- **POTENTIAL:** describe WHY the intended alternatives are important to the organization and the benefits to be derived from the multitude of options and related opportunities.
- **PROPOSAL:** ascribe WHERE/WHEN actions/requirements must be taken to fulfill milestones/objectives that fulfill outcomes/goals essential to obtaining the organizations vision.

It is through this use case approach that organizations can begin to understand the value proposition of data done right. Through the application of the 4P process, the most complex problems can be solved not merely fixed.

#### **Define Requirements:**

Any successful technology solution, AI included, is designed and implemented to fulfill the requirements of the solution's stakeholders. Defining these requirements, managing them over time, and ensuring clear communication between stakeholders and solution implementers are all critical tasks during the Implementation Phase. The result of effective requirements management during the Implementation Phase is a clear, well-defined set of requirements that are fulfilled by the production-ready system.

#### **Establish Operations Model:**

An operations model is a powerful tool in defining and effectively managing requirements for the solution. Consisting of both illustrations and text, the operations model provides a high-level, graphical representation of how the new solution will work, how it will interact with existing systems and processes, and how it will provide the forecasted business benefits. By developing the operations model in collaboration with key stakeholder groups and the design and implementation team, the project can ensure that stakeholder expectations and the system functionality are aligned and delivered in the production-ready system.

#### **Identify Integration Points:**

Integration points for the new system can be discovered, documented, and effectively addressed by drilling into the details of the operations model. Focusing on those points, where the new system interacts with existing systems and where any changes to business processes interact with existing and supporting business processes, provide clear and comprehensive

documentation of integration points for the new system. This information informs the upcoming production deployment effort, including technical deployment (e.g., changes to databases, data exchange processes such as ETL that will be required, and physical network modifications) and people-related deployment (e.g., changes to processes, business rules, or policies required for the new or augmented processes to be successful).

#### **Obtain ATO or IATT:**

With the production-ready system, requirements, and operating model completed, the project team has sufficient information to develop the system's security plan and obtain either an Authorization to Operate (ATO) or IATT (Intermediate Authorization to Test). This is a necessary step before full deployment can be initiated, and ensures the security of all information flows and assets related to the new solution.

#### **Conduct Proof-of-Concept / Pilot, if applicable:**

If the organization's assessment of the risk-reward tradeoff for the new system indicates that a pilot or proof-of-concept is warranted, that effort should be completed during the Implementation Phase. Completing the pilot or proof-of-concept enables the organization to learn about the risks, benefits, and operational considerations that work well for it, and to incorporate those learnings into its deployment planning. Proofs-of-concept and pilots can also generate valuable information providing initial "on the ground" insights to senior leadership that validate forecast, and demonstrate the potential benefits of the new capabilities.

Once these capabilities are demonstrated, a pilot can illustrate the potential benefits and cost savings to be actualized. As the benefits to be derived are realized, a pilot delivers initial operational capabilities (IOC). It is within this phase that they optimize effectiveness resulting in increased productivity and provide an initial and reoccurring return on investment. In time, a process refinement and efficiencies realization will result from reaching a fully operational capability. At this point, a decision about how to scale and create a sustainable system to meet current and future demands could be made.

## **Technology**

#### **Setup, Configure, and Deploy Platform and Tools:**

The tools and technologies selected in the previous phase need to be configured and deployed for implementation. The first step is to determine the number of environments and specifications for each environment. At least two distinct environments are recommended to be set up: one for development and testing, and another for production. Sometimes, depending on the nature of the problem, additional testing environments such as a separate user acceptance testing environment or a performance testing environment can and should be considered. There are two main aspects to setting up and configuring environments:

- **Infrastructure:** This involves selecting the hardware and operating system that suits the use case in question. Typically, it is recommended to consider GPU based cores that are optimized exclusively for data processing. Ideally, the infrastructure should also provide fast and easy access to storage with very low latency, considering the data-intensive nature of AI. The low latency becomes a critical requirement when dealing with cloud-based infrastructure, which does not house data in the cloud but rather uses data from on-premise applications. Furthermore, all infrastructure should comply with the agency's guidelines to ensure data security. This could include ensuring that data is encrypted based on FIPS 140-2 guidelines<sup>13</sup>.
- **Tools:** Within each environment, the required technological tools/applications need to be installed and configured. This could be a variety of capabilities to organize/structure the data, correlate and contextualize the information, and assess and analyze the knowledge. Through the application of technological tools, it is possible to more effectively collect and correlate data into information, and contextualizing knowledge to understand the causality/influence that creates the outcomes. The vast number of use cases may require a myriad of other specialized tools and access to open source Application Programming Interfaces (API) to consistently take a more scientific approach to understanding what is happening and how it is evolving over time.

#### **Data Preparation:**

This is one of the most critical and important stages for implementation. Data preparation consumes most of the time spent by a data scientist in developing AI solutions. Data preparation largely consists of organizing and structuring the data by applying an enterprise information model to make it possible to:

- **Discover data:** make it visible, produce an inventory, and identify its authoritative source.
- **Organize information:** make it accessible, create a catalog, and de-conflicted to make it fit for purpose.
- **Structure knowledge:** make it understandable, produce a dictionary, and validate its veracity so that it becomes operationally relevant.

This is accomplished through a process known as Extracting, Transforming and Loading (ETL) the data. It is within this process that data can be curated so that it is properly prepared for the assessment and analysis process. There are two options for ETL to make data fit for purpose:

- Extract data from the source and perform the necessary transformations for the application.
- Extract and transform data first and load/link the resulting data to plug into the algorithm.

The lower cost of data in the cloud has led to the former path as a preferred strategy. Doing so provides some advantages in terms of the ability to pick and choose the elements required for analysis, without having to go back repeatedly to the source of data. However, this approach also results in increased latency for extracting information due to the size of data being transferred, and increased complexity resulting from managing a larger amount of data. Ultimately, the approach selected depends on the use in question and the particular constraints of the situation.

To make data operationally relevant, the data sets must be refined based on feature selection. In this step, the data used to develop initial algorithms are narrowed down to the most critical data required for the analysis. The activity of identifying the most relevant attributes of data required for a use case is called feature selection. This is an iterative process in which new data is constantly added based on the needs of the use case and then refined based on the results obtained in the execution of the algorithms.

### **Developing Solutions in an Iterative Manner:**

This process consists of developing and executing the steps required to achieve the desired objective of the use case. In machine learning, this can consist of developing the required algorithms using various techniques such as Decision Trees, Regression, Random Forest models, Neural Networks, or Deep Neural Networks. Data prepared is pumped into the algorithms and outputs measured against expected results. This is done in an iterative manner, continuously adjusting and configuring the various parameters of the algorithms, and the amount and type of data captured for increasing accuracy and reducing bias. It is critical to understand the target accuracy expected to be achieved and the required threshold for success to prevent long trials and iterations which do not yield results.

### **Model Verification for Bias:**

As AI becomes a mainstay in today's information environment, the possible good or bad impact should always be considered. It is important to consider the influence that cognitive and cultural bias has upon evidence-based decisions. The consequences of this are vast and varied from its applications in criminal justice, credit/loans, recruitment, education, and clinical diagnosis, just to name a few.

Ethics has become a big part of the AI ecosystem. The implications that this has upon the digital environment allows the potential for bias to exist as a result of unfounded or substantiated frames of reference that create the perceptions and paradigms which produce skewed results. The impact of cognitive, cultural, and now computational bias is an area that must be characterized and quantified in order to effectively hold people, AI, and organizations accountable for their actions, outputs, and impact respectively.

It is imperative to identify the implications and outline these potential influences and how they impact biases in order to quantifiably assess and ascertain their impact. Therefore, it is essential to understand the influencers that create bias from data sets (*facts*), associate the information (*relationships*), ascertain the reliance between the interactions (*dependencies*), and assess the outcomes through confirmation (*reliability*). This process will help to determine the extent that the manifestation of the conditions and resulting causality will create negative consequences due to flawed facts, disconnected dependencies, and obfuscated or mischaracterized outcomes.

Through a comparative analysis of professional relationships, familiar frames of reference should be applied to the diverse and dynamic environment to derive a comparison of AI applications. Given a multitude of different perspectives (*backgrounds*), perceptions (*viewpoints*), and paradigms (*expectations*), it is critical to consider similar frameworks and models to apply, monitor, and manage AI capabilities. By preparing models to test and compare outputs of results temporally and contextually, the opportunity to identify and isolate bias can be revealed through a validation and verification process.

This has direct effect on how much humans trust the data/facts and the way they are assembled into algorithms/information, which is at the core of how useful the resulting data is to enlighten, empower, and evolve based on the circumstances. Thus, the evaluation must be based on three criteria in order to assess the level of trust and confidence people place in AI:

- Are the facts/data correct? Assess the veracity.
- Are the relationships/information dependable? Ascertain the reliability.
- Are the outputs/knowledge verifiable? Analyze the interpretation.

The subtlety of bias is often difficult to detect. Not all are bad and thus it is important to understand the types of bias and how it manifests:

- INTENTIONAL: Which is meant to harm
- UNINTENTIONAL: Which causes harm
- NECESSARY BIAS: To prevent harm

Intentional bias is the most destructive for it is intended to create harm through the exclusion of, and determent of others. Unintentional bias occurs more frequently because “we don’t know what we don’t know” and fail to test for or train. Necessary bias is introduced in engineering terms to provide parameters in order to assure the environment is effectively scoped and maintained with the boundaries of safety, operational limits, or the ascribed framework. It is often referred to as a tolerance and, if not properly identified, could be catastrophic.

Bias is not isolated to people. It overflows from their influence to affect all aspects of the AI adoption process. Thus, there is great probability that computers will promote prejudices, preclude appropriate perspectives, and even perpetrate paradigms that are contrary to the facts, formulas, and frameworks that create a false picture of reality. Whether flawed facts, faulty formulas, or distorted interpretations, bias can influence and affect outcomes at all three stages of AI's assessment, analysis, and outcomes:

- **COLLECTION:** Data can produce faulty facts given prejudice.
- **CORRELATION:** Information can be fused in a manner that creates incorrect perspectives.
- **CONTEXT:** Knowledge can be presented in a way to distort paradigms.

Since AI algorithms learn from large quantities of data, the machine learning models that the AI builds can amplify some of the biases inherently present in the data. Data-driven systems involve human judgment to sort and categorize the data; define the characteristics; and qualify attributes. Consideration must be given to who is evaluating, rating, and labeling the data. These questions must be considered to determine the existence of and impact of bias:

- How diverse is the team of raters?
- Has the tool used to label been tested for usability?
- Is the rating process consistent?
- Have you obtained user consent for data use?
- Have you considered multiple metrics for training? (e.g., short-term or long-term goals)
- Have you sampled the raw data?
- Have you obtained feedback from a diverse team?
- Does data have inclusive representation? (e.g., ages or geographies)
- Are assumptions documented and tested?
- Did you communicate data limitations with the users?

The following framework provide a means to model and measure bias in the data sets and the potential impact it can have:

- What is the goal of the AI solution and what outcomes are you achieving? Would these vary for different users and communities?
- Have you identified data sources?
- Are there any data outliers?
- Have you separated your training and test data sets?
- What is the distribution across a variety of parameters (e.g., age, geographical, ethnic, race, gender, etc.)?
- Have you used any open-source tools to review the data distribution? (e.g., Facets, What-If tool)

- Does data have blind spots? (language tools using “her” for certain professions like nursing and “him” for a CEO)
- Have you engaged a broader stakeholder group to review data sets? (e.g., legal, HR, policy, etc.)
- Have you tested data for unexpected/adverse impacts? (rate of false positives versus false negatives)
- Is there a plan for negative testing and stress-testing?
- Consideration of using data augmentation of synthetic data to ensure even data distribution?
- Explainability to ascertain “why is the algorithm making this recommendation?”
- How does the algorithm deal with unpredictable inputs?

While it is not necessary to show the entire math behind the machine learning model, determine the influencers and resulting impact relative to the narrative that connects the dots throughout the decision making the process. Ultimately, explainability is essential for regulated industries like healthcare and financial services to determine if quantifiable fact-based results are not influenced by subjective reasoning:

- What degree of explainability or interpretability do you need? (e.g., retail systems are not driven by as much regulation as healthcare systems)
- Have you identified features for interpretability? (some systems need more granularity than others)
- Is your model a white box or a black box? (this also depends on your platform of choice, especially if you do not use open-source tools)
- Are you able to provide a narrative around the AI solution? (e.g., why the recommendation was made)
- Have you considered exposing the decision tree of the machine-learning model?
- Have you optimized the output for understanding? (e.g., you can have a footnote for a recommendation that is not according to your preference)
- Did you include Model Confidence to determine degree of certainty? (example, can the system distinguish between dogs and cats)
- Have you considered using graphical representation to indicate certainty? (correlate the distinctions, difference, and differentials to illustrate contextual correlation)
- Can you design a test with minimal inputs that can provide the decision-making factors?
- Have you identified the causal relationships and not correlations whenever possible?

## Acquisition

An AI acquisition strategy is a critical aspect to enable and implement AI within an organization. This section describes the Acquisition phases, factors, and requirements that must be considered to procure, implement, and maintain AI.

## Acquisition Planning, Preparation, and Award

In accordance to [FAR- Part 7 Acquisition Planning<sup>14</sup>](#) and [Part 31 - Contract Cost Principles and Procedures<sup>15</sup>](#), to facilitate attainment of the AI acquisition objectives, an organization's contracting and AI program offices must develop an acquisition plan that identifies those milestones at which decisions should be made. The plan must address all the technical, business, management, and other significant considerations that will control the acquisition for implementing AI into an existing *Major system* or establishing an AI *Major system*. The specific content of plans will vary, depending on the nature, circumstances, and stage of the acquisition. In preparing the plan, the planner must follow the applicable FAR-Part 7.105 paragraph A/B instructions and include the agency's implementing procedures. Acquisition plans for service contracts or orders must describe the strategies for implementing performance-based acquisition methods or must provide the rationale for not using these methods outlined in the [FAR-Part 37.6 Service Contracting<sup>16</sup>](#).

According to FAR-Part 7.105, the plan should include, but not be limited to, a statement of need, acquisition background, objectives, applicable conditions, direct and indirect budget/funding/costs. It is these considerations that address accounting procedures, capability/performance requirements, delivery/performance-period requirements, acquisition streamlining protocols, contract options, source selection procedures, acquisition considerations, and trade-offs as they pertain to cost benefit analysis.

Once the Acquisition plan is completed, in accordance with [FAR-Part 6 Competition Requirements<sup>17</sup>](#) and [Part 14 - Sealed Bidding<sup>18</sup>](#), the organization's contracting and AI program offices would pursue a competitive or non-competitive contract solicitation for AI '*vendors for as to what is the availability to existing off-the-shelf (COTS)/ Government off-the-shelf (GOTS)* solutions and services. This includes, but not limited to, the preparation of acquisition artifacts to include, but not limited to, market research, Independent Government Cost Estimate (ICGE), Statement of Objectives (SOO), Performance work Schedule (PWS), and Statement of work (SOW). These options inform how contract solicitations and award process should be approached to include, but not limited to, the issuance of solicitation; executing and management of the bidding process; proposal selection criteria evaluation; proposal selection, offer, and the award based on the selection factors identified in the acquisition plan.

## Administer and Monitor the Contract

Contract and program offices must conduct auditing, administration, cost accounting, documentation, execution, reporting, and records retention for AI contracts in accordance with [FAR-Part 4 Administrative and Information Matters<sup>19</sup>](#), [Part 30 Cost Accounting Standards Administration<sup>20</sup>](#), [Part 42 - Contract Administration and Audit Services<sup>21</sup>](#).

## Contract Modification

During the administration of AI contract portfolio, if the contract and program offices identify changes, issues, or new AI-Related requirements, contract modifications may be necessary throughout all phases of the lifecycle of developing, implementing, integrating, maintaining, or decommissioning AI Technology in the systems that support the organization. The Contracting Office, along with the AI program office, executes the appropriate modifications in accordance with terms outlined in [FAR-Part 43 - Contract Modifications<sup>22</sup>](#).

## Prepare and Award Follow-on Contract:

As an AI contract is reaching its expiration, or if an existing contract is deemed to not support the organization's AI mission, goals, and objectives, the Contracting Office, along with the AI program office, may need to begin the follow-on acquisition planning stage to issue a new contract for ongoing development, implementation, integration, maintenance, recapitalization, or decommission of the AI Technology. If so, they would follow the guidelines outlined in the *Acquisition Planning and Award* subsection outlined above.

## Key Outcomes

### Engaged Outcomes

The key stakeholders that must be engaged are the AI Subject Matter Expert(s), the business owner and the contract team. These three groups of stakeholders are organized in a cross functional team.

- **AI SME:** Technical team that includes the architects, programs to assess AI solutions.
- **Product and Business Owner:** Organizational team which understands the operation problem.
- **Contract Team:** Acquisition specialist acting on behalf of the governments to purchase services and products that provide functionality and capabilities.
- **Cross Functional Team:** Representatives from each stakeholder and across government, industry, and academia.
- **Key Stakeholders:** Advocates, customers, suppliers, and employees.

### Defined Outcomes

The platform architecture, Initial Operating Capability (*IOC*) and AI models must be finalized with resulting AI/ML solution designed to meet the organizational business objectives. Along with these, critical AI/ML data and technological infrastructure must be defined and finalized.

- **Platform:** Architecture / IOC, AI models
- **Proposal:** AI/ML solution that meets business objectives
- **Portal:** AI/ML infrastructure setup

## Planned Outcomes

Once implemented, the AI model will have to be consistently refreshed and fine-tuned based on the feedback provided by the results and evaluation of the key stakeholders. To that end, a maintenance plan must be agreed to in order to ensure that any new data sets are added in a well-defined manner and any refinement is accurately noted and communicated.

- **Sustainment of AI model:** Monitor and measure outputs through a quantifiable assessment and analysis respectively
- **Maintenance Plan:** The care a feeding requirement to assure the viability of the tool
- **Data set addition and refinement:** Determination of how comprehensive, complete and correct the data sets are.

## Phase Outputs

### AI Outputs

Robust AI model: The goal at the end of implementation is a validated robust model that satisfies business user objectives and meets requirements. To achieve this requires several repetitive iterations of model development and training.

- Accessible model ready for integration: The results of the model should be readily accessible to users and or systems as required given the required inputs. This can be either via a user interface or an API, depending on the situation.
- Sustainment requirements: The methods, frequency, and any other requirements for training the model to ensure continued quality monitoring and measuring to ascertain the essential ecosystem is established.
- Code repository and versioning: Any code developed should be properly documented and managed to assure the validity and version is properly matured in the specified code repository.
- Production Environment and Deployment Pipeline: A separate production environment should be created to ensure segregation of duties and access control. All infrastructure, platform and application maintenance requirements should be identified. A deployment pipeline should be established to effectively manage the maturation of the code throughout all phases of the lifecycle; creation, testing, validating, sustainment, re-capitalization, and disposition.
- Testing: The AI model is repeatedly tested by different types of users for a wide variety of scenarios to ensure complete and comprehensive functionality and robustness. The results of testing should match necessary expectations set by business users. Included in this process incorporates testing non-functional but necessary requirements such as security, reliability, resilience, and performance. Included in this are the requirements to be approved and certified as outlined in the policies set for by the specific organization's authority to operate (ATO) requirements.

**Business Outputs:**

- Validated business case: An assessment of the total cost of ownership of the implemented solution, compared to the quantitative and qualitative benefits of the solution, with appropriate management endorsement. All costing and benefits calculations are verified for analytic accuracy, and all input data is verified by data stewards
- User Guide: A detailed “how to” manual for all selected user types, describing the implemented solution, typical use cases, step-by-step actions for each user type to successfully execute their respective use cases. The user guide also includes a standard glossary of terms, a description of all user administrative actions available, a Frequently Asked Questions section, and a description of “zero level” problem resolutions.
- Administration Guide: A detailed description of all administrative functions of the implemented solution, including user registration, system maintenance, escalation procedures, FISMA designation, and all relevant security controls.
- Resources Allocated: All resources required for successful, ongoing operation of the implemented solution are allocated, assigned, trained, integrated, and providing effective support.
- Success Criteria Metrics and Monitoring: All quantified metrics comprising the business case are specified, as are their data sources, frequency, access rights, and validation. A system is in place to regularly (if not automatically) pull metric data from associated sources, generate all required metrics, trending, sensitivity analysis, and visualization, available to appropriate audiences.

**GRC Outputs:**

- Governance Team/ Model Stood Up: The formalization of required governance team and steering committees needed for decision making, oversight, and compliance; the selection of frameworks, models, and standards that will govern AI implementation and operations.
- Cost Structure Finalized: Updated multi-year Lifecycle Cost Estimates (LCCE), Independent Government Cost Estimates (IGCE), Program/Project budgets, and/or Operational Budgets estimates and actuals. This may include, but not limited to, direct, indirect, and other direct costs (ODC) associated with initial implementation; Operations and Maintenance (O&M), Development Modernization and Enhancement (DME) costs related to applicable commodities, technical services, and non-technical services needed to implement and maintain AI capabilities.

## Decision Gate

At the end of a successful Implementation Phase, the organization is fully prepared for deploying the operational production system. All stakeholders should have a shared understanding of the responses to the questions in this Decision Gate section.

## Production System

- What are the key components of the production system and how do they align with fulfilling documented business requirements?
- What new information security controls are required?
- How does the new system affect current Business Continuity and Disaster Recovery documentation, planning, and operations?
- What current systems and processes will be impacted by the production system?
- How is each of these integration points being managed?
  - What mechanisms (technical, process, or people-related) are available to accommodate the potential improvement in the performance of the AI system as it learns over time?

## Strategy and Governance

- What potential bias exists in training data and how will it be managed to overcome/avoided?
- Do key stakeholders agree on the definition, measurement, and reporting of the new system's fulfillment of its business case?
- What is the procurement schedule and approach for the proposed program?
- What is the proposed schedule for governance and management structures?
- What are the key technological, business context, security, performance, user experience, program management, and governance related risks specific to the designed AI solution?
- Are all key stakeholders aware of the risks associated with the project (technical, talent, bias)?
- How will key risks be monitored, managed, and mitigated?
- Are KPIs defined and baselined?
- Do all key stakeholders endorse the deployment plan?

## Change Management

- **Who** are the key stakeholders in determining the successful deployment of the system? Who will be potentially affected by this solution?
- **What** are the key stakeholders' expectations for the new system? What are the key user concerns regarding the new system, and how are these going to be addressed?
- **When** are users of the new system going to be prepared? When is the overall business case for the new solution going to be assessed?
  - What will be measured?
  - Who will make the assessment?
  - When is the proposed project completion date?
  - How will it be reported?
- **Where** will the data come from? Where will the work be performed?
- **How** will the key stakeholders be engaged throughout the deployment? How will the effects of the new system impact the organization's strategy, mission focus, culture, and morale?

## Phase 5 – AI Integration

Once the organization has gone through Readiness, Assessment, Solution Selection, and AI Implementation phases successfully, this next phase integrates the AI solution into the organization's infrastructure.

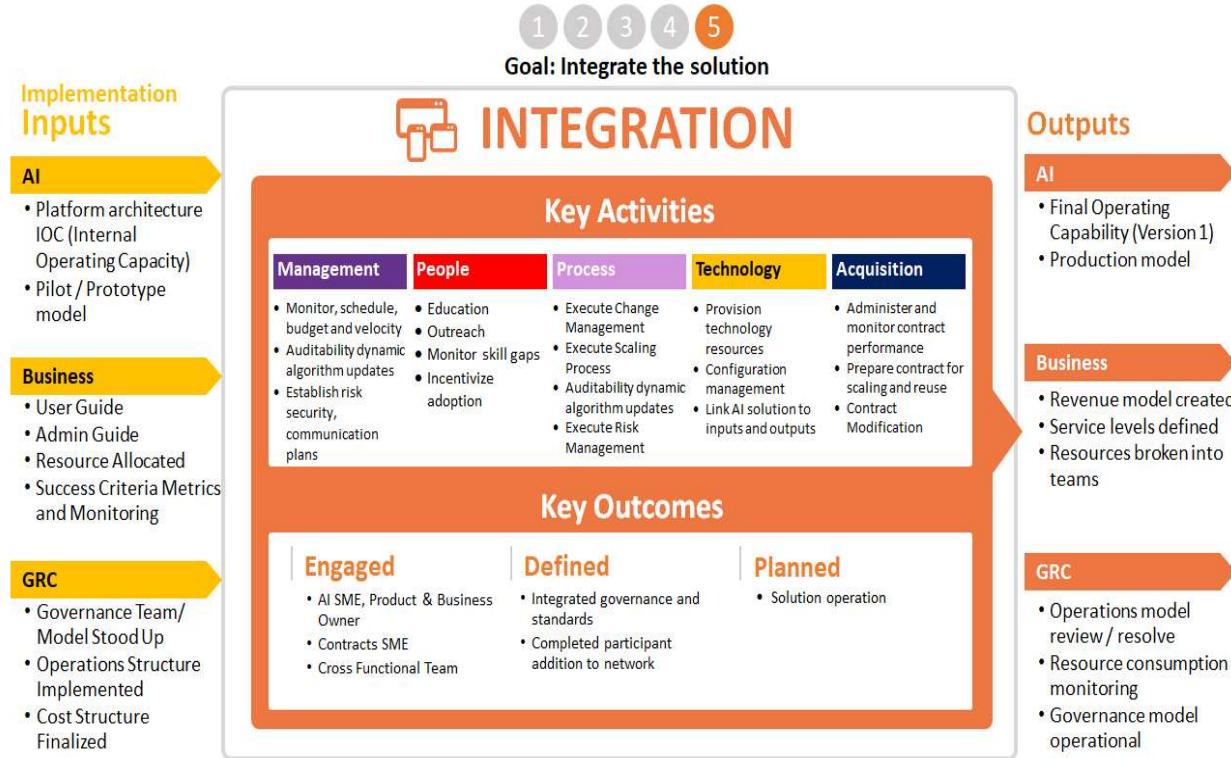


Figure 7: AI Integration Phase (5) summary

### Phase Inputs

Inputs to the AI integration phase are to be harmonized with outputs from the implementation phase, including definitions of data sources, platform architecture/Internal Operating Capacity (IOC) and a prototyped model accessible for integration. Inputs will define the functional state at end-production of the proposed AI solution in the context of pre-established specifications and technical requirements. The prototyped model should be implemented with the relevant data sources using an AI/ML architecture that is aligned with the organization's business objectives.

### Key Goals

The key goal of the AI integration phase is to get the AI solution, processes, strategies developed and implemented in the previous phases, and integrate them into the organization that will operate and maintain the AI solution.

## Key Participants

Key participants for successful AI integration could include:

- Product owners/managers to undertake overall management and governance of proposed AI solution.
- AI subject matter experts who may or may not be from the agency integrating the program.
- Stakeholders from the lines of business and systems with technical, process, or data intersections or integration with the proposed AI solution.
- An enterprise architect who is well versed with AI and data architectures to lead the execution of the integration framework into a production model capability.
- Change process leader driving change management:
  - Configuration planning
  - Item identification
  - Change control
  - Configuration status accounting
  - AI/ML configuration audit
- Executive sponsor (e.g., CIO at enterprise level project, director of business line)

## Key Considerations

With an operational solution, it is time to ensure the AI solution does not break the rest of the environments within each participating organization. Consider a situation where there is a security attack and operations within a participating organization is unaware on what the escalation protocol is or how to prevent the security attack.

Here are some questions that need to be answered in this phase:

- Have all the operational procedures been integrated into the IT operations of the organization(s)?
- Did the operational and user workforce receive proper training?
- Has the system been tested to ensure it does not cause other peripheral systems to break?
- Did the interfaces between the AI solution and the systems or users that use the solution work to the satisfaction of the stakeholders?
- Have the governance procedures been documented and tested?
- Has a strategy to identify and mitigate potential sources of bias been established?
- Is AI/ML solution explainable?
- Are all required documentation for an ATO completed?

Consider adopting standard project/program management practices across all Phases and Key Activities of the integrated AI solution. Critical activities should consider adequate identification and management of risk, communication, security, education, ethics, and regulatory requirements.

## Key Activities

### Management

#### Monitor, Schedule, Budget and Velocity

Track costs of AI associated iterative activities, monitor resource utilization, and weave the project into existing business cadences to ensure it stays in line with the provided cost structure, change management, and the cadence of time to meet desired outcomes. Integration pace and activities should be adjusted as necessary based on defined success metrics, reflecting both business value and technical functionality.

#### Auditability of Dynamic/Automated Processes

Baseline and crosscheck configurations against policy and regulatory compliance, including verifying and documenting algorithms and configurations utilized. Schedule verifications of AI outputs to measure for deviations between intended and actual outcomes and to test for ethics and bias, utilize robotic process automation (RPA), or other automation techniques to perform auditing of AI activities. Ensure consensus for change management processes if the solution is implemented as an enterprise shared service and capture change success rates.

#### Establish Risk, Security and Communication Management Plans

Establish training outlines to ensure topic coverage for people, processes, and security concerns. Develop contingency plans for fail-safe posture for anomalous behavior. Develop thorough risk management planning early to minimize re-work once implemented. Utilize governance models to build tracking mechanisms for meetings, actions, and resolutions. The factor for AI enacted in all Risk and Security activities. Involve the security organization early and often to ensure proper consideration and expedite the ATO certification process.

### People

During this phase, it is critical to ensure the workforce stays current with the evolution of the technologies implemented along with similar competing and/or complementing technologies that could influence the future operation of the solution.

Outreach is important to keep stakeholders aware of the implemented solution and enables the operational team to stay apprised of what is happening outside of the integrated solution.

It is essential to leverage regulations, policies, executive letters, and industry best practices to continuously monitor and analyze what skills are considered necessary and otherwise preferred. The outcomes of the analysis are used to establish and maintain the baseline against which skills gaps the difference between skills that the government needs and wants and what the workforce offers are measured. Measuring the skills gap should be a continuously monitored activity to permit adjustments through training offered via multiple media sources.

The outcomes of the analysis are to assist the organization to identify skills needed to meet mission goals; additionally, it will be used to plan and execute employee development, retention, and recruitment activities. The government has started identifying competency, skill gaps and future requirements in the Federal workforce (e.g., Federal Acquisition Institute<sup>23</sup>, Office of Personnel Management<sup>24</sup>.)

Government must continuously monitor and manage its assets which include people. If an electronic employee file is established, it will assist in tracking the health of the workforce, measuring for knowledge, skills, and abilities and permitting for the reuse and sharing of those workforce resources.

It is also important to adopt the right kind of incentives to enable the workforce and support it at the time of integration of the solution and transition to operational status.

## Process

### Execute Change Management

- Change management is the guiding tool for personnel to swiftly adjust to and adopt new methods for performing and completing tasks within unfamiliar territory.
- Identify and leverage AI evangelists and early adopters across the organization early. These individuals who are excited by new technology and welcome change can help spread a positive message and help create quick win use cases to improve adoption across the organization. In some organizations, evangelists are identified during implementation and used to test the POC and provide valuable feedback to improve on the final product.
- Leveraging findings from the evaluation stage of potential roadblocks and key users impacted by the change; this is the point when the project teams need to begin executing the change management strategy, which includes activities such as:
  - Addressing potential resistance.
  - Communicating why change is happening and listening to feedback & concerns.
  - Preparing and equipping managers & supervisors.
  - Launch sponsorship activities.
  - Launch incentives/rewards systems to reinforce good change adoption behaviors.
  - Launch coaching sessions that align AI to cultural values and mission.
  - Develop training.
- These are key steps requiring alignment of project teams, leadership, and key stakeholders.
  - Leadership commitment to change management is critical for effective integration.
  - Limited awareness of need for resource allocation which could impede integration.

- Inform leaders of the critical connection between managing the people side of change, technology application, and resource allocation to assure the success of the project.
- Project teams who may traditionally be focused on just “switching on” need to value and understand the importance of change management and the positive effect it could have on their ability to deliver on time, on target, and on task.
- A programmatic approach that takes into consideration the necessary empathy on behalf of team leaders to deliver change management as a credible, structured and intentional approach, is essential if the team is to be driven by concrete milestones that produce desirable deliverables
- Create collaborative teams to include change management and project practitioners with transparent communication and clarity on roles for both teams.
- For AI impacting decision making, it is necessary for people at all levels to trust the algorithms’ suggestions and the facts that informed decisions.
- Leadership must empower employees to take the necessary actions as advised from the outputs of AI technological tools.

### **Execute Scaling Process**

Once Initial Operational Capabilities (IOC) is achieved, scaling is essential to meet Full Operational Capabilities (FOC) requirements. Scaling happens **vertically** (adding more people to an existing product or initiative) or **horizontally** (adding more teams and products).

- Establish an environment which is elastic and scales based on demand.
- Automate scaling practices where possible to remain flexible.
- Train internal teams on the environment, automated scaling infrastructure, and resiliency standards.
- Automated Testing and Delivery need to be in place to support system and process scaling.

### **Auditability of Dynamic/Automated Processes (e.g., Algorithm Updates)**

- Prioritize explainability and keep algorithms as simple as possible.
- Set benchmarks and increase complexity only if it adds value.
- Address data quality and quantity to improve suitability of algorithms
- Deliver open source, reusable algorithms to increase collaboration and awareness across agencies and across teams.

### **Execute Risk, Security and Communication Management Plans**

- Leverage plans created in early stages.
- Provide training to teams on configuration, use, and how to execute.

- Communication Plans:
  - Communication plans should consist of the following 9 steps; steps 1-7 should be completed prior to Integration phase:
    - Identify your objectives
    - Choose your target audiences
    - Design your key messages
    - Select your communication methods
    - Plan for two-way communication
    - Establish your time frame
    - Draft a budget
    - Implement the plan
    - Monitor the results and look for ways to improve
  - When executing communication plans, always focus on “telling it like it is” and focusing on the “me” issues that will directly impact employees and management across the organization.
  
- Risk management
  - By this stage, the team should have a list of potential risks documented based on internal review during the Readiness phase, and concurrent findings through Selection and Implementation phases. Project managers should also be assigned to evaluate key risks.

## Technology

Because AI, like other technological advances, must fit into existing processes and it is critical to set up AI for success by properly maintaining the system and managing connections with other IT components. These steps begin with analyzing the enterprise architecture and ensuring provisioning or resources are aligned and integrated to meet the internal and external demand. Configuration management assures the appropriate allocation of resources essential to maintaining AI solution. Thus, close coordination with the enterprise architects is essential throughout all phases of the development, integration, implementation, and sustainment of AI systems. Therefore, it is crucial to model, monitor, and measure outputs against desired outcomes to ascertain the effectiveness of AI solutions.

Modeling and meeting user demand is the first step to integrating AI with existing technology. A solution that lacks sufficient speed or availability will simply not be adopted. On-premise solutions require a platform to proceed. Thus, the need to acquire the necessary hardware and software involves procurement, configuration, integration, and implementation. It is this traditional approach that creates considerable cost in both currency, coordination, and contraction to acquire. This lengthy, laborious process preclude the necessary agile approach to create an adaptive system that is able to adjust to an ever-changing environment.

If the AI solution incorporates deep learning, specialized hardware, such as Graphical Processing Units (GPUs), may be necessary to meet demands with enough speed and minimized price. If the technology is to run in a cloud environment, there will be an on-going cost, and the solution will also likely need cloud architects to handle provisioning and security. Regardless of whether the solution is on premise or in the cloud, agencies should expect to regularly update AI solutions for security updates, software upgrades, licenses if applicable, and other IT management needs and factor those costs into the business case.

Consideration should be toward collaborative partnerships focused upon a service approach to acquire capabilities. However, both solutions require regular configuration management to assure the sustainment of evolving capabilities that are sufficient to meet mission needs and updates to underlying systems, whether they are on-premises or in the cloud to support the AI solution. Therefore, ongoing interaction, coordination, and collaboration with IT managers throughout all phases of the project. This requirement extends to both users, suppliers, and analyzers of the AI capabilities. Therefore, ongoing communication is essential to create the necessary feedback essential to synchronizing and synergizing the application of AI technology.

The application of AI has many means to operationalize data as a strategic asset. Whether through API or RPA, care must be taken before updates to an application take action to ascertain if it is operating with prescribed parameters. Caution must be considered for changes to external APIs, for the results of which could easily change the formatting or features of the application. Likewise, an upgrade to an application used by an RPA or update to the website can confuse the AI and cause it to crash or incorrectly input or output data. The best solution is for the AI implementers to test its interactions with new software in a non-production environment and make updates before the linked software is deployed into production.

### **Provision Technology Resources**

- In order to integrate AI solutions with existing processes, agencies and companies must ensure that the solution has sufficient resources to meet user demand.
  - On-premise solution requirements covers hardware and software purchases, as well as ongoing configuration and sustainment requirements.
  - Cloud resource requirements have ongoing monetary components as well as the need for cloud architects to integrate new and evolving requirements.
- Plan to regularly update AI solutions (both on-premise and in the cloud) for security patches, software upgrades, configuration control, and other IT integration and license management requirements.

### **Configuration Management**

- Network and/or cloud architects must properly configure the AI solution so that users can access it, and it can access the resources essential for it to perform its mission.

- Coordinate with other administrators to ensure configuration changes are included in future updates so that routine software upgrades do not disrupt service.

### **Link AI Solution Inputs and Outputs**

- AI solution must be configurable to allow the linkage to ingest the necessary data that inform the algorithms and provide the answers through an aggregation process:
  - Ideally, APIs will link AI systems with input from RPA and NLP systems to collect and correlate the necessary data that informs the knowledge base essential to AI's evolving understanding.
  - If RPAs are used, care must be used to ensure that changes to the input system's user interface do not break the AI solution or cause it to look for data in the wrong location.

User feedback is required to ensure that users can understand the data generated by the AI and that the results are clearly presented to evolve the users' contextual understanding.

- Care must be taken to ensure that AI solutions properly outputs the data to the correct location.
  - APIs are the ideal method for outputting data from the AI solution to assure the means to collect and correlate data continuously.
  - If RPAs are used, the output method must be documented and provided to the appropriate systems developers to ensure that the RPA is able to function correctly in any system updates.

## **Acquisition**

### **Administer and Monitor Contract Performance**

The Contract and Program Office monitor the performance of the newly integrated technology from the perspective of contract compliance and business line satisfaction. Depending on the contract type and incentive structure, the contracting officer awards payments based on the contractor's ability to meet established service-level agreements (SLA) and performance metrics. Working with all stakeholders, the team measures the impact of the new AI enabled portfolio.

### **Prepare Contract for Scaling and Reuse**

Establish the scaling and reuse aspects of the contract during acquisition planning. Avoid “Scope Creep” by executing future service areas upon need. The statement of work (SOW) will also state the selected technology for the desired AI solution, express the component and configuration of the chosen platform, describe the required supplies or services needed by the government for implementation of the target technology, and define anticipated tasks necessary to implement the AI-enabled solution successfully. Refine the procurement

processes to ensure that this new AI acquisition model will provide a template for the next project. Continuous acquisition improvement, especially within the emerging technology landscape, positions the agency for clean acquisitions with minimal disruption to program execution.

### **Contract Modification**

Contract modification may be necessary to complete the project. The Contract and Program officials will collaborate and determine specific changes within the current scope for modification. Capture all contract modifications in the Acquisition's retrospective for continuous acquisition improvement.

### **Key Outcomes**

#### **Defined**

When applicable, the AI solution has now been integrated within the network architecture of participating legacy systems, stakeholders, and organizations. Inputs and outputs are tested for each of the hosted organizations to ensure expected outcomes according to pre-established acceptance performance criteria.

#### **Planned**

Solution Operation - The AI solution is now ready to be rolled out and operate.

### **Phase Outputs**

Once the integration is complete, the AI solution will be a Final Operating Capability (possibly Version 1 of many) and/or a Production Ready Model. The business will have a revenue model created, service levels defined, and resources broken into teams. Finally, the governance, risk and compliance (GRC) will have a reviewed or resolved Operations Model, costs monitored, and an operational Governance Model.

## CONCLUSION

The interest in artificial intelligence continues to gain momentum. Even though AI has been around for over a half a century, since its inception in 1956 at Dartmouth College by John McCarthy<sup>25</sup>, its potential is just now being actualized. It has the potential to impact every aspect of our government operations and, if properly applied, will undoubtedly be a big part of shaping our future.

As organizations begin to appreciate the distinctions between AI and tools, they will be able to better understand its unique ability to consistently compare and correlate facts into analytical information upon which to contextualize the distinctions, differences, and differentials of a multitude of available options and the opportunities they offer. It is the resulting contextual understanding between items and time which is at the core of AI capabilities.

This playbook outlines the process of considering the viability of this technology to solve problems and serve the organization. It provides the pathway to apply AI to collect, correlate, and contextualize information that will enlighten, empower and evolve contextual understanding. AI offers the ability to explore the art of the possible and apply the science of the probable which informs what influence are available (causality) and how they affect our environment (consequence).

The progression and process of building knowledge based on understanding provides the potential to make better evidence-based decisions. By using the frameworks, models, and steps presented in this Playbook, organizations will be well prepared to develop, implement, and integrate AI capabilities to optimize the effectiveness of their organization. Applying the definitions and types of AI found in the ACT-IAT Artificial Intelligence/Machine Learning Primer and the process described in this Playbook, the reader will be able to create AI solutions that serve their organization.

These resources prepare the organization for success in their efforts to leverage the capabilities of AI. They illustrate the benefits of applying cognition to leverage knowledge and becoming aware of the options available to drive strategic advantage. The result is the intuitive capacity of the machine to become aware, adapt, and learn to make wise choices. This agile approach assures that past actions compared to current circumstances are made discernible in order to achieve the ascribed future objectives. The resulting iterative approach to integrating and implementing lessons learned enables AI to serve as a transformational catalyst to make the vision of the future a reality today.

## GLOSSARY

**Acquisition:** Acquire with appropriated funds of supplies or services (including construction) by and for the use of the Federal Government through purchase or lease, whether the supplies or services are already in existence or must be created, developed, demonstrated, and evaluated. Acquisition begins at the point when agency needs are established and includes the description of requirements to satisfy agency needs, solicitation and selection of sources, award of contracts, contract financing, contract performance, contract administration, and those technical and management functions directly related to the process of fulfilling agency needs by contract.

**Acquisition planning:** The process by which the efforts of all personnel responsible for an acquisition are coordinated and integrated through a comprehensive plan for fulfilling the agency need in a timely manner and at a reasonable cost. It includes developing the overall strategy for managing the acquisition.

**Commercially available off-the-shelf (COTS):** Any item of supply that is, (i) A commercial item (as defined in paragraph (1) of the definition in this section); (ii) Sold in substantial quantities in the commercial marketplace; and (iii) Offered to the Government, under a contract or subcontract at any tier, without modification, in the same form in which it is sold in the commercial marketplace.

**AI Computer Software:** Means (i) Computer programs that comprise a series of instructions, rules, routines, or statements, regardless of the media in which recorded, that allow or cause a computer to perform a specific operation or series of operations; and (ii) Recorded information comprising source code listings, design details, algorithms, processes, flow charts, formulas, and related material that would enable the computer program to be produced, created, or compiled.

**Contract:** A mutually binding legal relationship obligating the seller to furnish the supplies or services (including construction) and the buyer to pay for them. It includes all types of commitments that obligate the Government to an expenditure of appropriated funds and that, except as otherwise authorized, are in writing. In addition to bilateral instruments, contracts include (but are not limited to) awards and notices of awards; job orders or task letters issued under basic ordering agreements; letter contracts; orders, such as purchase orders, under which the contract becomes effective by written acceptance or performance; and bilateral contract modifications. Contracts do not include grants and cooperative agreements covered by 31 U.S.C.6301, et seq. For discussion of various types of contracts.

**Contract modification:** any written change in the terms of a contract.

**Contracting:** means purchasing, renting, leasing, or otherwise obtaining supplies or services from nonfederal sources. Contracting includes description (but not determination) of supplies and services required, selection and solicitation of sources, preparation and award of contracts, and all phases of contract administration. It does not include making grants or cooperative agreements.

**Contracting office:** an office that awards or executes a contract for supplies or services and performs post award functions not assigned to a contract administration office

**Direct cost:** any cost that is identified specifically with a particular final cost objective. Direct costs are not limited to items that are incorporated in the end product as material or labor. Costs identified specifically with a contract are direct costs of that contract. All costs identified specifically with other final cost objectives of the contractor are direct costs of those cost objectives.

**Indirect cost:** any cost not directly identified with a single final cost objective, but identified with two or more final cost objectives or with at least one intermediate cost objective.

**Major system:** combination of elements that will function together to produce the capabilities required to fulfill a mission need. The elements may include hardware, equipment, software, or any combination thereof, but exclude construction or other improvements to real property.

**Market research:** collecting and analyzing information about capabilities within the market to satisfy agency needs.

**Statement of Objectives (SOO):** means a Government-prepared document incorporated into the solicitation that states the overall performance objectives. It is used in solicitations when the Government intends to provide the maximum flexibility to each offeror to propose an innovative approach.

**Performance Work Statement (PWS):** a statement of work for performance-based acquisitions that describes the required results in clear, specific and objective terms with measurable outcomes.

**Statement of Work (SOW):** is typically used when the task is well-known and can be described in specific terms, provides explicit statements of work direction for the contractor to follow, and can also be found to contain references to desired performance outcomes, performance standards, and metrics.

**Government off-the-shelf (GOTS):** A software and/or hardware product that is developed by the technical staff of a Government organization for use by the U.S. Government. GOTS

software and hardware may be developed by an external entity, with specification from the Government organization to meet a specific Government purpose, and can normally be shared among Federal agencies without additional cost. GOTS products and systems are not commercially available to the general public. Sales and distribution of GOTS products and systems are controlled by the Government.

**Offer:** a response to a solicitation that, if accepted, would bind the offeror to perform the resultant contract.

## ACKNOWLEDGEMENT

The AI Working Group thanks the authors and contributors who provided a tremendous amount of time, hard work, and good humor to bring AI Playbook for the U.S. Federal Government to completion. The AI Working Group would like to also thank all our government, industry, and academia collaborative partners who provided invaluable feedback as reviewers.

## Authors and Affiliations

This paper was written by a consortium of government and industry. The organizational affiliations of the authors and contributors are included for information purposes only. The views expressed in this document do not necessarily represent the official views of the individuals and organizations that participated in its development.

Fatima Akhtar	IBM
Gil Alterovitz	National Artificial Intelligence Institute, Veterans Affairs
Sandy Barsky	United States Government
G. Hussein Basaria	Maven Group LLC
Janelle Billingslea	Department of Health and Human Services
Denise Blady	Defense Information Systems Agency
John Gustavo Blair	Fairfax County Economic Development Authority
Anil Chaudhry	Department of Homeland Security
Chakib Chraibi	Department of Commerce
Johnny E. Davis, Jr.	National Credit Union Administration
Frederic de Vaulx	Prometheus Computing, LLC
Deborah Detwiler	Maven Group LLC
Latecia Engram	Department of Health and Human Services
Ken Farber	Abeyon
Jorge A. Ferrer, MD, MBA, FAMIA	Department of Veterans Affairs
Jaime Garcia	Department of Homeland Security
Timothy George	Maven Group LLC
Lesly Goh	World Bank
Todd Hager	Macro Solutions
David Hernandez	Excella
Gabriel Hidalgo	National Institute of Dental and Craniofacial Research
Joyce Hunter	Vulcan Enterprises, LLC
Rodney I. Johnson	Department of Housing and Urban Development
Gail Kalbfleisch	Department of Veterans Affairs
June W. Lau	National Institute of Standards and Technology
Katherine Livis	Livis Consulting
Orlando Lopez	National Institutes of Health
Brendan Mahoney	General Services Administration

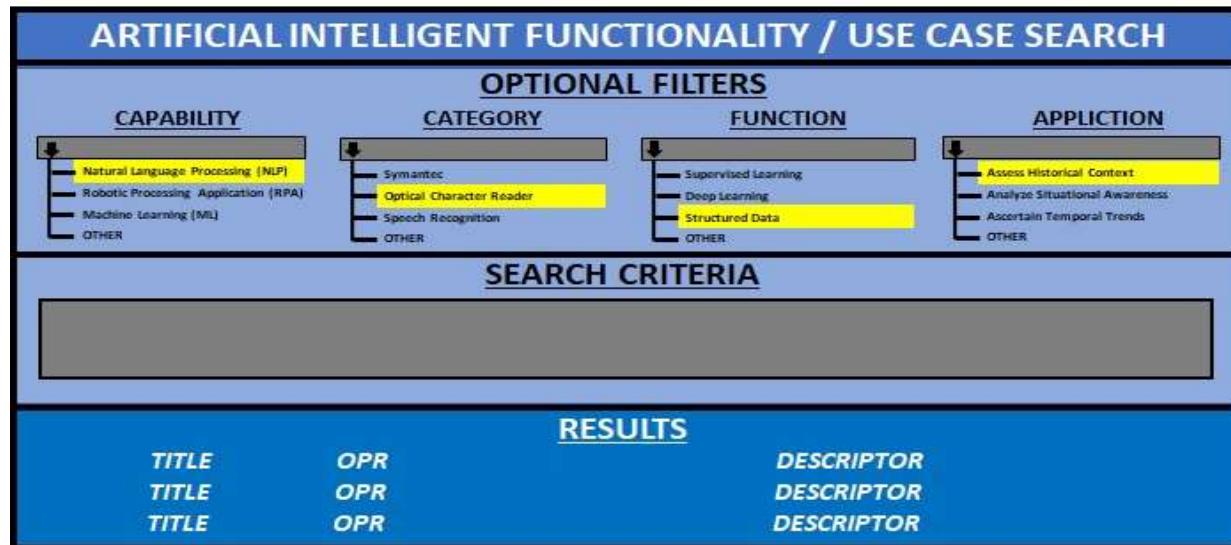
Aaron Margolis	Harper Paige, LLC
David Maron	National Artificial Intelligence Institute, Veterans Affairs
Brian Muolo	General Services Administration
Mallesh Murugesan	Abeyon
Sreenivasa Mutthe	Department of Veterans Affairs
Madhavi Nookala	Department of Veterans Affairs
Dennis Papula	Department of Health and Human Services
Stephen Pereira	Department of Transportation
Matthew Diamond, MD, PhD	Food and Drug Administration
Eric Popiel	Office of Personnel Management
Sanjeev Pulapaka	REI Systems
Sridhar Rajagopalan	Alpha Omega Integration
Mike Rice	CornerStone, LLC
James Rolfes	U.S. Consumer Product Safety Commission
Jennifer Rostami	General Services Administration
Akanksha Sharma	General Services Administration
Sherri Sokol	Defense Information Systems Agency
Eric Steinberg	Internal Revenue Service
Juanita C. Stewart	Department of Defense
Nevin Taylor	National Artificial Intelligence Institute, Veterans Affairs
Raj Tiwari	IEEE
Michael Torres	Office of Personnel Management
Jim Tunnessen	Voice of America
Mitchell D. Winans	Internal Revenue Service
Marc Wine	Department of Veterans Affairs
Robert Wurhman	General Services Administration
Swathi Young	TechNotch Solution

## APPENDICES

### Appendix A - AI Functionality Template

As ascribed in this Playbook, leveraging the AI capabilities to solve the most perplexing problems and challenges capitalizes on the operational benefits that this technology can provide. The following is a template that can be used to identify the multitude of means to utilize AI. This is designed to register the use cases so that others can appreciate the various ways of applying AI technology. It serves as a card catalog upon which to search functionality, capabilities, and application of the creative use of AI.

ARTIFICIAL INTELLIGENT FUNCTIONALITY / USE CASE SEARCH			
<u>OPTIONAL FILTERS</u>			
CAPABILITY	CATEGORY	FUNCTION	APPLICATION
<ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> Natural Language Processing (NLP)</li> <li><input type="checkbox"/> Robotic Processing Application (RPA)</li> <li><input type="checkbox"/> Machine Learning (ML)</li> <li><input type="checkbox"/> OTHER</li> </ul>	<ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> Symantec</li> <li><input checked="" type="checkbox"/> Optical Character Reader</li> <li><input type="checkbox"/> Speech Recognition</li> <li><input type="checkbox"/> OTHER</li> </ul>	<ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> Supervised Learning</li> <li><input checked="" type="checkbox"/> Deep Learning</li> <li><input checked="" type="checkbox"/> Structured Data</li> <li><input type="checkbox"/> OTHER</li> </ul>	<ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> Assess Historical Context</li> <li><input checked="" type="checkbox"/> Analyze Situational Awareness</li> <li><input checked="" type="checkbox"/> Ascertain Temporal Trends</li> <li><input type="checkbox"/> OTHER</li> </ul>
<u>SEARCH CRITERIA</u>			
<input style="width: 100%; height: 100px;" type="text"/>			
<u>RESULTS</u>			
<b>TITLE</b>	<b>OPR</b>	<b>DESCRIPTOR</b>	
<b>TITLE</b>	<b>OPR</b>	<b>DESCRIPTOR</b>	
<b>TITLE</b>	<b>OPR</b>	<b>DESCRIPTOR</b>	
<input style="width: 100%; height: 50px;" type="text"/>			
<u>RESULTS / VALUE PROPOSITION:</u>			
<input style="width: 100%; height: 50px;" type="text"/>			



#### Use case template/questionnaire:

- What question(s) do you believe AI can help you solve?
- Tell us about your data. Are they very well structured? Well-structured data would be driver's license records at a DMV, poorly structured data would be if you Googled "cat." This goes to the heart of the question about how much data you need. With well-structured data, you need less. With poorly structured data, you need a lot. The internet has lots and lots of pictures of cats, for example, so poorly structured data in this case.
- How does your data move from point of acquisition (point A) to the place where they are deposited (point B)? Are all your datasets in one place?
- What infrastructure/software components/people are required to get data from A to B?
- Have you examined possible sources of bias?
- Has your organization adapted well to the AI deployment?
- Did AI help you solve your problem? Why or why not?

## Appendix B – Playbook Navigation

In this appendix, you will find valuable information to help you understand the framework laid out in this Playbook and a series of questions to guide you and help you kick start your AI development journey.

### Framework Flow

The Playbook introduces a framework made out of several phases connected to one another sequentially to guide you through key activities that will help you leverage AI technologies to tackle your use case.

Each phase is connected to a decision gate before automatically going to the next phase. This decision gate helps you determine if you should:

- **Go to the next phase** – using the outputs generated by the phase  $n$ , you determine that there is enough value to keep going to the next phase.
- **Stop** – using the outputs generated by the phase  $n$ , you determine that AI does not bring enough value at this time.
- **Iterate** – using the outputs generated by the phase  $n$ , you determine that more work is needed or data from previous phases need to be adjusted.

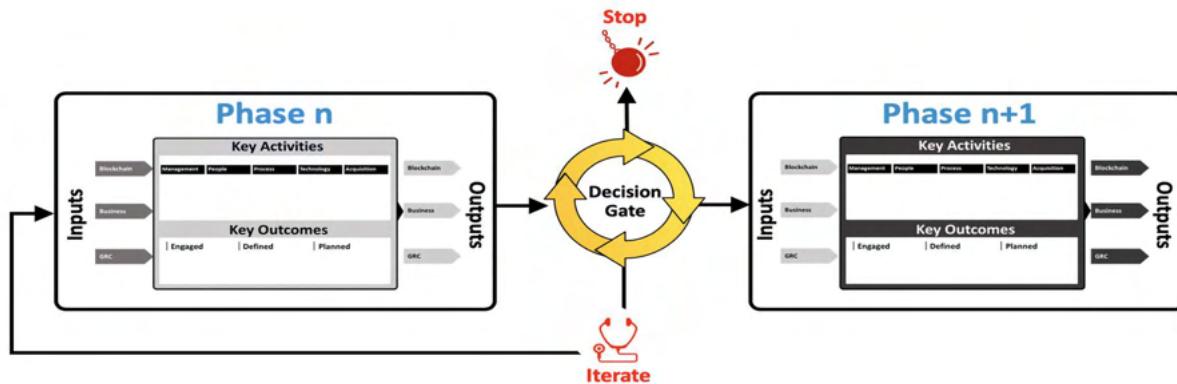


Figure 8: Playbook phase to phase flow

In addition, each phase is made up of key activities and outcomes to help you work through aspects needed to achieve your use case objectives leveraging AI technologies. The phase has inputs necessary for the key activities and generated outputs that will be needed for the decision gate. These outputs will also be used as inputs for the next phase.

## REFERENCES

<sup>1</sup> ACT-IAC Artificial Intelligence/Machine Learning Primer. Published March 12, 2019.

<https://www.actiac.org/act-iac-white-paper-artificial-intelligence-machine-learning-primer>

<sup>2</sup> GSA's Office of Shared Solutions and Performance Improvement (OSSPI), Office of Government-wide Policy, Modernization and Migration Management (M3) framework

<https://ussm.gsa.gov/m3/>

<sup>3</sup> General Services Administration (GSA) Emerging Citizen Technology Atlas

<https://emerging.digital.gov/>

<sup>4</sup> Vadlamudi, P. September 20, 2019. Adobe Obtains New FedRAMP Authorizations, Enhancing Its Digital Government Services in the Cloud <https://theblog.adobe.com/adobe-obtains-new-fedramp-authorizations-enhancing-its-digital-government-services-in-the-cloud/>

<sup>5</sup> National Institute for Standards and Technology (NIST) Privacy Impact Assessment

<https://www.nist.gov/system/files/documents/2017/05/09/NIST-TIP-PIA-Consolidated.pdf>

<sup>6</sup> General Data Protection Regulation (GDPR) <https://gdpr-info.eu/>

<sup>7</sup> U.S. Department of Health & Human Services Health Information Portability and Accountability Act (HIPAA) <https://www.hhs.gov/hipaa/index.html>

<sup>8</sup> GSA Shared Management Offices <https://www.gsa.gov/shared-services/shared-services-qsmo>

<sup>9</sup> Office of Management and Budget OMB M-19-13 March 20, 2019. Category Management: Making Smarter Use of Common Contract Solutions and Practices

<https://www.whitehouse.gov/wp-content/uploads/2019/03/M-19-13.pdf>

<sup>10</sup> GSA's Office of Shared Solutions and Performance Improvement (OSSPI), Office of Government-wide Policy, Modernization and Migration Management (M3) framework

<https://ussm.gsa.gov/m3/>

<sup>11</sup> Best-in-Class (BIC) solution designated by the Office of Management and Budget

<https://www.gsa.gov/buying-selling/category-management/bestinclass>

<sup>12</sup> General Services Administration Federal Acquisition Regulation (FAR)

<https://www.acquisition.gov/>

<sup>13</sup> National Institute for Standards and Technology (NIST) FIPS 140-2 Security Requirements for Cryptographic Modules <https://csrc.nist.gov/publications/detail/fips/140/2/final>

<sup>14</sup> Federal Acquisition Regulation, Part 7 - Acquisition Planning

<https://www.acquisition.gov/content/part-7-acquisition-planning>

<sup>15</sup> Federal Acquisition Regulation, Part 31 - Contract Cost Principles and Procedures

<https://www.acquisition.gov/content/part-31-contract-cost-principles-and-procedures>

<sup>16</sup> Federal Acquisition Regulation, Part 37.6 – Service Contracting

<https://www.acquisition.gov/content/part-37-service-contracting#i1077388>

<sup>17</sup> Federal Acquisition Regulation, Part 6, Competition Requirements

<https://www.acquisition.gov/content/part-6-competition-requirements>

<sup>18</sup> Federal Acquisition Regulation, Part 14, Sealed Bidding

<https://www.acquisition.gov/content/part-14-sealed-bidding>

<sup>19</sup> Federal Acquisition Regulation, Part 4, Administrative and Information Matters

<https://www.acquisition.gov/content/part-4-administrative-and-information-matters>

<sup>20</sup> Federal Acquisition Regulation, Part 30 Cost Accounting Standards Administration

<https://www.acquisition.gov/content/part-30-cost-accounting-standards-administration>

<sup>21</sup> Federal Acquisition Regulation, Part 42 Contract Administration and Audit Services

<https://www.acquisition.gov/content/part-42-contract-administration-and-audit-services>

<sup>22</sup> Federal Acquisition Regulation, Part 43, Contract Modifications

<https://www.acquisition.gov/content/part-43-contract-modifications>

<sup>23</sup> Federal Acquisition Institute. 2018. New FAC Specialization Focuses on Digital Services

<https://www.fai.gov/announcements/new-fac-specialization-focuses-digital-services>

<sup>24</sup> Office of Personnel Management (OPM). 2018. Federal Workforce Priorities Report (FWPR)

<https://www.opm.gov/policy-data-oversight/human-capital-management/federal-workforce-priorities-report/2018-federal-workforce-priorities-report.pdf>

Office of Personnel Management (OPM). Assessment & Evaluation COMPETENCY GAP

ANALYSIS. <https://www.opm.gov/services-for-agencies/assessment-evaluation/competency-gap-analysis/>

<sup>25</sup> Andresen, Scott L. (2002). John McCarthy: Father of AI. *IEEE Intelligent Systems*,

September/October 2002 (17), 84-85. DOI Bookmark: 10.1109/MIS.2002.1039837

<https://www.computer.org/csl/magazine/ex/2002/05/x5084/13rRUxE04ph>

# Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing

Inioluwa Deborah Raji\*

Partnership on AI

deb@partnershiponai.org

Margaret Mitchell

Google

Jamila Smith-Loud

Google

Andrew Smart\*

Google

andrewsmart@google.com

Timnit Gebru

Google

Daniel Theron

Google

Rebecca N. White

Google

Ben Hutchinson

Google

Parker Barnes

Google

## ABSTRACT

Rising concern for the societal implications of artificial intelligence systems has inspired a wave of academic and journalistic literature in which deployed systems are audited for harm by investigators from outside the organizations deploying the algorithms. However, it remains challenging for practitioners to identify the harmful repercussions of their own systems prior to deployment, and, once deployed, emergent issues can become difficult or impossible to trace back to their source.

In this paper, we introduce a framework for algorithmic auditing that supports artificial intelligence system development end-to-end, to be applied throughout the internal organization development lifecycle. Each stage of the audit yields a set of documents that together form an overall audit report, drawing on an organization's values or principles to assess the fit of decisions made throughout the process. The proposed auditing framework is intended to contribute to closing the *accountability gap* in the development and deployment of large-scale artificial intelligence systems by embedding a robust process to ensure audit integrity.

## CCS CONCEPTS

- Social and professional topics → System management; Technology audits;
- Software and its engineering → Software development process management.

## KEYWORDS

Algorithmic audits, machine learning, accountability, responsible innovation

\*Both authors contributed equally to this paper. This work was done by Inioluwa Deborah Raji as a fellow at Partnership on AI (PAI), of which Google, Inc. is a partner. This should not be interpreted as reflecting the official position of PAI as a whole, or any of its partner organizations.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

FAT\* '20, January 27–30, 2020, Barcelona, Spain

© 2020 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6936-7/20/02.

<https://doi.org/10.1145/3351095.3372873>

## ACM Reference Format:

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Conference on Fairness, Accountability, and Transparency (FAT\* '20)*, January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3351095.3372873>

## 1 INTRODUCTION

With the increased access to artificial intelligence (AI) development tools and Internet-sourced datasets, corporations, nonprofits and governments are deploying AI systems at an unprecedented pace, often in massive-scale production systems impacting millions if not billions of users [1]. In the midst of this widespread deployment, however, come valid concerns about the effectiveness of these automated systems for the full scope of users, and especially a critique of systems that have the propensity to replicate, reinforce or amplify harmful existing social biases [8, 37, 62]. External audits are designed to identify these risks from outside the system and serve as accountability measures for these deployed models. However, such audits tend to be conducted after model deployment, when the system has already negatively impacted users [26, 51].

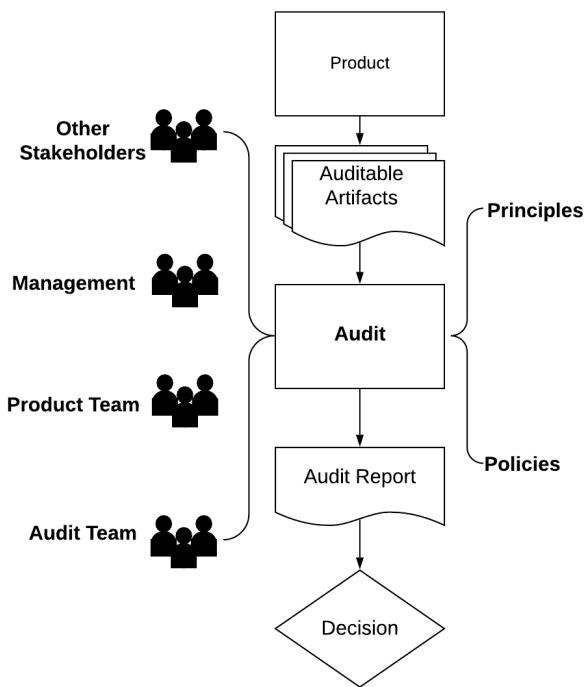
In this paper, we present internal algorithmic audits as a mechanism to check that the engineering processes involved in AI system creation and deployment meet declared ethical expectations and standards, such as organizational AI principles. The audit process is necessarily boring, slow, meticulous and methodical—antithetical to the typical rapid development pace for AI technology. However, it is critical to slow down as algorithms continue to be deployed in increasingly high-stakes domains. By considering historical examples across industries, we make the case that such audits can be leveraged to anticipate potential negative consequences before they occur, in addition to providing decision support to design mitigations, more clearly defining and monitoring potentially adverse outcomes, and anticipating harmful feedback loops and system-level risks [20]. Executed by a dedicated team of organization employees, internal audits operate within the product development context and can inform the ultimate decision to abandon the development of AI technology when the risks outweigh the benefits (see Figure 1).

Inspired from the practices and artifacts of several disciplines, we go further to develop SМАСТР, a defined internal audit framework meant to guide practical implementations. Our framework strives to establish interdisciplinarity as a default in audit and engineering processes while providing the much needed structure to support the conscious development of AI systems.

## 2 GOVERNANCE, ACCOUNTABILITY AND AUDITS

We use *accountability* to mean the state of being responsible or answerable for a system, its behavior and its potential impacts [38]. Although algorithms themselves cannot be held accountable as they are not moral or legal agents [7], the organizations designing and deploying algorithms can through *governance* structures. Proposed standard ISO 37000 defines this structure as "the system by which the whole organization is directed, controlled and held accountable to achieve its core purpose over the long term."<sup>1</sup> If the responsible development of artificial intelligence is a core purpose of organizations creating AI, then a governance system by which the whole organization is held accountable should be established.

<sup>1</sup><https://committee.iso.org/sites/tc309/home/projects/ongoing/ongoing-1.html>



**Figure 1: High-level overview of the context of an internal algorithmic audit. The audit is conducted during product development and prior to launch. The audit team leads the product team, management and other stakeholders in contributing to the audit. Policies and principles, including internal and external ethical expectations, also feed into the audit to set the standard for performance.**

In environmental studies, Lynch and Veland [45] introduced the concept of *urgent governance*, distinguishing between *auditing* for system reliability vs societal harm. For example, a power plant can be consistently productive while causing harm to the environment through pollution [42]. Similarly, an AI system can be found technically reliable and functional through a traditional engineering quality assurance pipeline without meeting declared ethical expectations. A separate governance structure is necessary for the evaluation of these systems for ethical compliance. This evaluation can be embedded in the established quality assurance workflow but serves a different purpose, evaluating and optimizing for a different goal centered on social benefits and values rather than typical performance metrics such as accuracy or profit [39]. Although concerns about reliability are related, and although practices for testing production AI systems are established for industry practitioners [4], issues involving social impact, downstream effects in critical domains, and ethics and fairness concerns are not typically covered by concepts such as technical debt and reliability engineering.

### 2.1 What is an audit?

*Audits* are tools for interrogating complex processes, often to determine whether they comply with company policy, industry standards or regulations [43]. The IEEE standard for software development defines an audit as "an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures" [32]. Building from methods of external auditing in investigative journalism and research [17, 62, 65], algorithmic auditing has started to become similar in spirit to the well-established practice of bug bounties, where external hackers are paid for finding vulnerabilities and bugs in released software [46]. These audits, modeled after intervention strategies in information security and finance [62], have significantly increased public awareness of algorithmic accountability.

An external audit of automated facial analysis systems exposed high disparities in error rates among darker-skinned women and lighter-skinned men [8], showing how structural racism and sexism can be encoded and reinforced through AI systems. [8] reveals *interaction failures*, in which the production and deployment of an AI system interacts with unjust social structures to contribute to biased predictions, as Safiya Noble has described [54]. Such findings demonstrate the need for companies to understand the social and power dynamics of their deployed systems' environments, and record such insights to manage their products' impact.

### 2.2 AI Principles as Customized Ethical Standards

According to Mittelstadt [49], at least 63 public-private initiatives have produced statements describing high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI. Important values such as ensuring AI technologies are subject to human direction and control, and avoiding the creation or reinforcement of unfair bias, have been included in many organizations' ethical charters. However, the AI industry lacks proven methods to translate principles into practice [49], and AI principles have been criticized for being vague and providing

little to no means of accountability [27, 82]. Nevertheless, such principles are becoming common methods to define the ethical priorities of an organization and thus the operational goals for which to aim [34, 83]. Thus, in the absence of more formalized and universal standards, they can be used as a North Star to guide the evaluation of the development lifecycle, and internal audits can investigate alignment with declared AI principles prior to model deployment. We propose a framing of risk analyses centered on the failure to achieve AI principle objectives, outlining an audit practice that can begin translating ethical principles into practice.

### 2.3 Audit Integrity and Procedural Justice

Audit results are at times approached with skepticism since they are reliant on and vulnerable to human judgment. To establish the integrity of the audit itself as an independently valid result, the audit must adhere to the proper execution of an established audit process. This is a repeatedly observed phenomenon in tax compliance auditing, where several international surveys of tax compliance demonstrate that a fixed and vetted tax audit methodology is one of the most effective strategies to convince companies to respect audit results and pay their full taxes [22, 53].

Procedural justice implies the legitimacy of an outcome due to the admission of a fair and thorough process. Establishing procedural justice to increase compliance is thus a motivating factor for establishing common and robust frameworks through which independent audits can demonstrate adherence to standards. In addition, audit integrity is best established when auditors themselves live up to an ethical standard, vetted by adherence to an expected code of conduct or norm in how the audit is to be conducted. In finance, for example, it became clear that any sense of dishonesty or non-transparency in audit methodology would lead audit targets to dismiss rather than act on results [66].

### 2.4 The Internal Audit

External auditing, in which companies are accountable to a third party [62], are fundamentally limited by lack of access to internal processes at the audited organizations. Although external audits conducted by credible experts are less affected by organization-internal considerations, external auditors can only access model outputs, for example by using an API [65]. Auditors do not have access to intermediate models or training data, which are often protected as trade secrets [9]. Internal auditors' direct access to systems can thus help extend traditional external auditing paradigms by incorporating additional information typically unavailable for external evaluations to reveal previously unidentifiable risks.

The goals of an internal audit are similar to quality assurance, with the objective to enrich, update or validate the risk analysis for product deployment. Internal audits aim to evaluate how well the product candidate, once in real-world operation, will fit the expected system behaviour encoded in standards.

A modification in objective from a post-deployment audit to pre-deployment audit applied throughout the development process enables proactive ethical intervention methods, rather than simply informing reactive measures only implementable after deployment, as is the case with a purely external approach. Because there is an increased level of system access in an internal audit, identified

gaps in performance or processes can be mapped to sociotechnical considerations that should be addressed through joint efforts with product teams. As the audit results can lead to ambiguous conclusions, it is critical to identify key stakeholders and decision makers who can drive appropriate responses to audit outcomes.

Additionally, with an internal audit, because auditors are employees of the organization and communicate their findings primarily to an internal audience, there is opportunity to leverage these audit outcomes for recommendations of structural organizational changes needed to make the entire engineering development process auditable and aligned with ethical standards. Ultimately, internal audits complement external accountability, generating artifacts or transparent information [70] that third parties can use for external auditing, or even end-user communication. Internal audits can thus enable review and scrutiny from additional stakeholders, by enforcing transparency through stricter reporting requirements.

## 3 LESSONS FROM AUDITING PRACTICES IN OTHER INDUSTRIES

Improving the governance of artificial intelligence development is intended to reduce the risks posed by new technology. While not without faults, safety-critical and regulated industries such as aerospace and medicine have long traditions of auditable processes and design controls that have dramatically improved safety [77, 81].

### 3.1 Aerospace

Globally, there is one commercial airline accident per two million flights [63]. This remarkable safety record is the result of a joint and concerted effort over many years by aircraft and engine manufacturers, airlines, governments, regulatory bodies, and other industry stakeholders [63]. As modern avionic systems have increased in size and complexity (for example, the Boeing 787 software is estimated at 13 million lines of code [35]), the standard 1-in-1,000,000,000 per use hour maximum failure probability for critical aerospace systems remains an underappreciated engineering marvel [19].

However, as the recent Boeing 737 MAX accidents indicate, safety is never finished, and the qualitative impact of failures cannot be ignored—even one accident can impact the lives of many and is rightfully acknowledged as a catastrophic tragedy. Complex systems tend to drift toward unsafe conditions unless constant vigilance is maintained [42]. It is the sum of the tiny probabilities of individual events that matters in complex systems—if this grows without bound, the probability of catastrophe goes to one. The *Borel-Cantelli Lemmas* are formalizations of this statistical phenomenon [13], which means that we can never be satisfied with safety standards. Additionally, standards can be compromised if competing business interests take precedence. Because the non-zero risk of failure grows over time, without continuous active measures being developed to mitigate risk, disaster becomes inevitable [29].

**3.1.1 Design checklists.** Checklists are simple tools for assisting designers in having a more informed view of important questions, edge cases and failures [30]. Checklists are widely used in aerospace for their proven ability to improve safety and designs. There are several cautions about using checklists during the development of complex software, such as the risk of blind application, the broader

context and nuanced interrelated concerns are not considered. However, a checklist can be beneficial. It is good practice to avoid yes/no questions to reduce the risk that the checklist becomes a box-ticking activity, for example by asking designers and engineers to describe their processes for assessing ethical risk. Checklist use should also be related to real-world failures and higher-level system hazards.

**3.1.2 Traceability.** Another key concept from aerospace and safety-critical software engineering is *traceability*—which is concerned with the relationships between product requirements, their sources and system design. This practice is familiar to the software industry in requirements engineering [2]. However, in AI research, it can often be difficult to trace the provenance of large datasets or to interpret the meaning of model weights—to say nothing of the challenge of understanding how these might relate to system requirements. Additionally, as the complexity of sociotechnical systems is rapidly increasing, and as the speed and complexity of large-scale artificial intelligence systems increase, new approaches are necessary to understand risk [42].

**3.1.3 Failure Modes and Effects Analysis.** Finally, a standard tool in safety engineering is a *Failure Modes and Effects Analysis* (FMEA), methodical and systematic risk management approach that examines a proposed design or technology for foreseeable failures [72]. The main purpose of a FMEA is to define, identify and eliminate potential failures or problems in different products, designs, systems and services. Prior to conducting a FMEA, known issues with a proposed technology should be thoroughly mapped through a literature review and by collecting and documenting the experiences of the product designers, engineers and managers. Further, the risk exercise is based on known issues with relevant datasets and models, information that can be gathered from interviews and from extant technical documentation.

FMEAs can help designers improve or upgrade their products to reduce risk of failure. They can also help decision makers formulate corresponding preventive measures or improve reactive strategies in the event of post-launch failure. FMEAs are widely used in many fields including aerospace, chemical engineering, design, mechanical engineering and medical devices. To our knowledge, however, the FMEA method has not been applied to examine ethical risks in production-scale artificial intelligence models or products.

## 3.2 Medical devices

Internal and external quality assurance audits are a daily occurrence in the pharmaceutical and medical device industry. Audit document trails are as important as the drug products and devices themselves. The history of quality assurance audits in medical devices dates from several medical disasters in which devices, such as infusion pumps and autoinjectors, failed or were used improperly [80].

**3.2.1 Design Controls.** For medical devices, the stages of product development are strictly defined. In fact, federal law (Code of Federal Regulations Title 21) mandates that medical-device makers establish and maintain “design control” procedures to ensure that design requirements are met and designs and development processes are auditable. Practically speaking, design controls are a documented method of ensuring that the end product matches the intended use, and that potential risks from using the technology

have been anticipated and mitigated [77]. The purpose is to ensure that anticipated risks related to the use of technology are driven down to the lowest degree that is reasonably practicable.

**3.2.2 Intended Use.** Medical-device makers must maintain procedures to ensure that design requirements meet the “intended use” of the device. The intended use of a “device” (or, increasingly in medicine, an algorithm—see [60] for more) determines the level of design control required: for example, a tongue depressor (a simple piece of wood) is the lowest class of risk (Class I), while a deep brain implant would be the highest (Class III). The intended use of a tongue depressor could be “to displace the tongue to facilitate examination of the surrounding organs and tissues”, differentiating a tongue depressor from a Popsicle stick. This may be important when considering an algorithm that can be used to identify cats or to identify tumors; depending on its intended use, the same algorithm might have drastically different risk profiles, and additional risks arise from unintended uses of the technology.

**3.2.3 Design History File.** For products classified as medical devices, at every stage of the development process, device makers must document the design input, output, review, verification, validation, transfer and changes—the design control process (section 3.2.1). Evidence that medical device designers and manufacturers have followed design controls must be kept in a design history file (DHF), which must be an accurate representation and documentation of the product and its development process. Included in the DHF is an extensive risk assessment and hazard analysis, which must be continuously updated as new risks are discovered. Companies also proactively maintain “post-market surveillance” for any issues that may arise with safety of a medical device.

**3.2.4 Structural Vulnerability.** In medicine there is a deep acknowledgement of socially determinant factors in healthcare access and effectiveness, and an awareness of the social biases influencing the dynamic of prescriptions and treatments. This widespread acknowledgement led to the framework of operationalizing structural vulnerability in healthcare contexts, and effectively the design of an assessment tool to record the anticipated social conditions surrounding a particular remedy or medical recommendation [61]. Artificial intelligence models are equally subject to social influence and social impact, and undergoing such assessments on more holistic and population- or environment-based considerations is relevant to algorithmic auditing.

## 3.3 Finance

As automated accounting systems started to appear in the 1950s, corporate auditors continued to rely on manual procedures to audit “around the computer”. In the 1970s, the Equity Funding Corporation scandal and the passage of the Foreign Corrupt Practices Act spurred companies to more thoroughly integrate internal controls throughout their accounting systems. This heightened the need to audit these systems directly. The 2002 Sarbanes-Oxley Act introduced sweeping changes to the profession in demanding greater focus on financial reporting and fraud detection [10].

Financial auditing had to play catch-up as the complexity and automation of financial business practices became too unwieldy to manage manually. Stakeholders in large companies and government

regulators desired a way to hold companies accountable. Concerns among regulators and shareholders that the managers in large financial firms would squander profits from newly created financial instruments prompted the development of financial audits [74].

Additionally, as financial transactions and markets became more automated, abstract and opaque, threats to social and economic values were answered increasingly with audits. But financial auditing lagged behind the process of technology-enabled financialization of markets and firms.

**3.3.1 Audit Infrastructure.** In general, internal financial audits seek assurance that the organization has a formal governance process that is operating as intended: values and goals are established and communicated, the accomplishment of goals is monitored, accountability is ensured and values are preserved. Further, internal audits seek to find out whether significant risks within the organization are being managed and controlled to an acceptable level [71].

Internal financial auditors typically have unfettered access to necessary information, people, records and outsourced operations across the organization. IIA Performance Standard 2300, Performing the Engagement [55], states that internal auditors should identify, analyze, evaluate and record sufficient information to achieve the audit objectives. The head of internal audit determines how internal auditors carry out their work and the level of evidence required to support their conclusions.

### 3.4 Discussion and Challenges

The lessons from other industries above are a useful guide toward building internal accountability to society as a stakeholder. Yet, there are many novel and unique aspects of artificial intelligence development that present urgent research challenges to overcome.

Current software development practice in general, and artificial intelligence development in particular, does not typically follow the *waterfall* or verification-and-validation approach [16]. These approaches are still used, in combination with agile methods, in the above-mentioned industries because they are much more documentation-oriented, auditable and requirements-driven. Agile artificial intelligence development is much faster and iterative, and thus presents a challenge to auditability. However, applying agile methodologies to internal audits themselves is a current topic of research in the internal audit profession.<sup>2</sup>

Most internal audit functions outside of heavily regulated industries tend to take a risk-based approach. They work with product teams to ask "what could go wrong" at each step of a process and use that to build a risk register [59]. This allows risks to rise to the surface in a way that is informed by the people who know these processes and systems the best. Internal audits can also leverage relevant experts from within the company to facilitate such discussions and provide additional insight on potential risks [3].

Large-scale production AI systems are extraordinarily complex, and a critical line of future research relates to addressing the interaction of highly complex coupled sociotechnical systems. Moreover, there is a dynamic complex interaction between users as sources of data, data collection, and model training and updating. Additionally, governance processes based solely on risk have been criticized for

<sup>2</sup><https://deloitte.wsj.com/riskandcompliance/2018/08/06/mind-over-matter-implementing-agile-internal-audit/>

being unable to anticipate the most profound impacts from technological innovation, such as the financial crisis in 2008, in which big data and algorithms played a large role [52, 54, 57].

With artificial intelligence systems it can be difficult to trace model output back to requirements because these may not be explicitly documented, and issues may only become apparent once systems are released. However, from an ethical and moral perspective it is incumbent on producers of artificial intelligence systems to anticipate ethics-related failures before launch. However, as [58] and [31] point out, the design, prototyping and maintenance of AI systems raises many unique challenges not commonly faced with other kinds of intelligent systems or computing systems more broadly. For example, *data entanglement* results from the fact that artificial intelligence is a tool that mixes data sources together. As Scully et al. point out, artificial intelligence models create entanglement and make the isolation of improvements effectively impossible [67], which they call *Change Anything Change Everything*. We suggest that by having explicit documentation about the purpose, data, and model space, potential hazards could be identified earlier in the development process.

Selbst and Baracas argue that "one must seek explanations of the process behind a model's development, not just explanations of the model itself" [68]. As a relatively young community focused on fairness, accountability, and transparency in AI, we have some indication of the system culture requirements needed to normalize, for example, an adequately thorough documentation procedure and guidelines [24, 48]. Still, we lack the formalization of a standard model development template or practice, or process guidelines for when and in which contexts it is appropriate to implement certain recommendations. In these cases, internal auditors can work with engineering teams to construct the missing documentation to assess practices against the scope of the audit. Improving documentation can then be a remediation for future work.

Also, as AI is at times considered a "general purpose technology" with multiple and dual uses [78], the lack of reliable standardization poses significant challenges to governance efforts. This challenge is compounded by increasing customization and variability of what an AI product development lifecycle looks like depending on the anticipated context of deployment or industry.

We thus combine learnings from prior practice in adjacent industries while recognizing the uniqueness of the commercial AI industry to identify key opportunities for internal auditing in our specific context. We do so in a way that is appropriate to the requirements of an AI system.

## 4 SMACTR: AN INTERNAL AUDIT FRAMEWORK

We now outline the components of an initial internal audit framework, which can be framed as encompassing five distinct stages—Scoping, Mapping, Artifact Collection, Testing and Reflection (SMACTR)—all of which have their own set of documentation requirements and account for a different level of the analysis of a system. Figure 2 illustrates the full set of artifacts recommended for each stage.

To illustrate the utility of this framework, we contextualize our descriptions with the hypothetical example of Company X Inc.,

Scoping	Mapping	Artifact Collection	Testing	Reflection	Post-Audit
Define Audit Scope	Stakeholder Buy-In	Audit Checklist	Review Documentation	Remediation Plan	Go / No-Go Decisions
Product Requirements Document (PRD)	Conduct Interviews	Model Cards	Adversarial Testing	Design History File (ADHF)	Design Mitigations
AI Principles	Stakeholder Map	Datasheets	Ethical Risk Analysis Chart		Track Implementation
Use Case Ethics Review	Interview Transcripts			Summary Report	
Social Impact Assessment	Failure modes and effects analysis (FMEA)				

**Figure 2: Overview of Internal Audit Framework.** Gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.

a large multinational software engineering consulting firm, specializing in developing custom AI solutions for a diverse range of clients. We imagine this company has designated five AI principles, paraphrased from the most commonly identified AI principles in a current online English survey [34]—“Transparency”, “Justice, Fairness & Non-Discrimination”, “Safety & Non-Maleficence”, “Responsibility & Accountability” and “Privacy”. We also assume that the corporate structure of Company X is typical of any technical consultancy, and design our stakeholder map by this assumption.

Company X has decided to pilot the SMACTR internal audit framework to fulfill a corporate mandate towards responsible innovation practice, accommodate external accountability and operationalize internal consistency with respect to its identified AI principles. The fictional company thus pilots the audit framework on two hypothetical client projects.

The first (hypothetical) client wishes to develop a child abuse screening tool similar to that of the real cases extensively studied and reported on [11, 14, 15, 21, 25, 36]. This complex case intersects heavily with applications in high-risk scenarios with dire consequences. This scenario demonstrates how, for algorithms interfacing with high-risk contexts, a structured framework can allow for the careful consideration of all the possibilities and risks with taking on the project, and the extent of its understood social impact.

The second invented client is Happy-Go-Lucky, Inc., an imagined photo service company looking for a smile detection algorithm to automatically trigger the cameras in their installed physical photo booths. In this scenario, the worst case is a lack of customer satisfaction—the stakes are low and the situation seems relatively straightforward. This scenario demonstrates how in even seemingly simple and benign cases, ethical consideration of system deployment can reveal underlying issues to be addressed prior to deployment, especially when we contextualize the model within the setting of the product and deployment environment.

An end-to-end worked example of the audit framework is available as supplementary material to this paper for the Happy-Go-Lucky, Inc. client case. This includes demonstrative templates of all recommended documentation, with the exception of specific process files such as any experimental results, interview transcripts,

a design history file and the summary report. Workable templates can also be accessed as an online resource [here](#).

#### 4.1 The Governance Process

To design our audit procedure, we suggest complementing formal risk assessment methodologies with ideas from responsible innovation, which stresses four key dimensions: *anticipation*, *reflexivity*, *inclusion* and *responsiveness* [73], as well as system-theoretic concepts that help grapple with increasing complexity and coupling of artificial intelligence systems with the external world [42]. Risk-based assessments can be limited in their ability to capture social and ethical stakes, and they should be complemented by anticipatory questions such as, “what if...?”. The aim is to increase ethical foresight through systematic thinking about the larger sociotechnical system in which a product will be deployed [50]. There are also intersections between this framework and just effective product development theory [5], as many of the components of audit design refocus the product development process to prioritize the user and their ultimate well-being, resulting in a more effective product performance outcome.

At a minimum, the internal audit process should enable critical reflections on the potential impact of a system, serving as internal education and training on ethical awareness in addition to leaving what we refer to as a “transparency trail” of documentation at each step of the development cycle (see Figure 2). To shift the process into an actionable mechanism for accountability, we present a validated and transparently outlined procedure that auditors can commit to. The thoroughness of our described process will hopefully engage the trust of audit targets to act on and acknowledge post-audit recommendations for engineering practices in alignment with prescribed AI principles.

This process primarily addresses how to conduct internal audits, providing guidance for those that have already deemed an audit necessary but would like to further define the scope and execution details. Though not covered here, an equally important process is determining what systems to audit and why. Each industry has a way to judge what requires a full audit, but that process is discretionary and dependent on a range of contextual factors pertinent to the industry, the organization, audit team resourcing, and the case

at hand. Risk prioritization and the necessary variance in scrutiny is a separately interesting and rich research topic on its own. The process outlined below can be applied in full or in a lighter-weight formulation, depending on the level of assessment desired.

## 4.2 The Scoping Stage

For both clients, a product or request document is provided to specify the requirements and expectations of the product or feature. The goal of the scoping stage is to clarify the objective of the audit by reviewing the motivations and intended impact of the investigated system, and confirming the principles and values meant to guide product development. This is the stage in which the risk analysis begins by mapping out intended use cases and identifying analogous deployments either within the organization or from competitors or adjacent industries. The goal is to anticipate areas to investigate as potential sources of harm and social impact. At this stage, interaction with the system should be minimal.

In the case of the smile-triggered phone booth, a smile detection model is required, providing a simple product, with not a broad scope of considerations as the potential for harm does not go much beyond inconvenience or customer exclusion and dissatisfaction. For the child abuse detection product, there are many more approaches to solving the issue and many more options for how the model interacts with the broader system. The use case itself involves many ethical considerations, as an ineffective model may result in serious consequences like death or family separation.

The key artifacts developed by the auditors from this stage include an ethical review of the system use case and a social impact assessment. Pre-requisite documents from the product and engineering team should be a declaration or confirmation statement of ethical objectives, standards and AI principles. The product team should also provide a Product Requirements Document (PRD), or project proposal from the initial planning of the audited product.

**4.2.1 Artifact: Ethical Review of System Use Case.** When a potential AI system is in the development pipeline, it should be reviewed with a series of questions that first and foremost check to see, at a high level, whether the technology aligns with a set of ethical values or principles. This can take the form of an ethical review that considers the technology from a responsible innovation perspective by asking who is likely to be impacted and how.

Importantly, we stress standpoint diversity in this process. **Algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values.** Thus it is not always possible for individual technology workers to identify or assess their own biases or faulty assumptions [33]. For this reason, a critical range of viewpoints is included in the review process. The essential inclusion of independent domain experts and marginalized groups in the ethical review process "has the potential to lead to more rigorous critical reflection because their experiences will often be precisely those that are most needed in identifying problematic background assumptions and revealing limitations with research questions, models, or methodologies" [33]. Another method to elicit implicit biases or motivated cognition [40] is to ask people to reflect on their preliminary assessment and then ask whether they might have reason to regret the

action later on. This can shed light on how our position in society biases our assumptions and ways of knowing [18].

An internal ethics review board that includes a diversity of voices should review proposed projects and document its views. Internal ethics review boards are common in biomedical research, and the purpose of these boards is to ensure that the rights, safety, and well-being of all human subjects involved in medical research are protected [56]. Similarly, the purpose of an ethics review board for AI systems includes safeguarding human rights, safety, and well-being of those potentially impacted.

**4.2.2 Artifact: Social Impact Assessment.** A social impact assessment should inform the ethical review. Social impact assessments are commonly defined as a method to analyze and mitigate the unintended social consequences, both positive and negative, that occur when a new development, program, or policy engages with human populations and communities [79]. In it, we describe how the use of an artificial intelligence system might change people's ways of life, their culture, their community, their political systems, their environment, their health and well-being, their personal and property rights, and their experiences (positive or negative) [79].

The social impact assessment includes two primary steps: an assessment of the severity of the risks, and an identification of the relevant social, economic, and cultural impacts and harms that an artificial intelligence system applied in context may create. The severity of risk is the degree to which the specific context of the use case is assessed to determine the degree in which potential harms may be amplified. The severity assessment proceeds from the analysis of impacts and harms to give a sense of the relative severity of the harms and impacts depending on the sensitivity, constraints, and context of the use case.

## 4.3 The Mapping Stage

The mapping stage is not a step in which testing is actively done, but rather a review of what is already in place and the perspectives involved in the audited system. This is also the time to map internal stakeholders, identify key collaborators for the execution of the audit, and orchestrate the appropriate stakeholder buy-in required for execution. At this stage, the FMEA (Section 3.1.3) should begin and risks should be prioritized for later testing.

As Company X is a consultancy, this stage mainly requires identifying the stakeholders across product and engineering teams anchored to this particular client project, and recording the nature of their involvement and contribution. This enables an internal record of individual accountability with respect to participation towards the final outcome, and enables the trace of relevant contacts for future inquiry.

For the child abuse detection algorithm, the initial identification of failure modes reveals the high stakes of the application, and immediate threats to the "Safety & Non-Maleficence" principle. False positives overwhelm staff and may lead to the separation of families that could have recovered. False negatives may result in a dead or injured child that could have been rescued. For the smile detector, failures disproportionately impact those with alternative emotional expressions—those with autism, different cultural norms on the formality of smiling, or different expectations for the photograph who are then excluded from the product by design.

The key artifacts from this stage include a stakeholder map and collaborator contact list, a system map of the product development lifecycle, and the engineering system overview, especially in cases where multiple models inform the end product. Additionally, this stage includes a design history file review of all existing documentation of the development process or historical artifacts on past versions of the product. Finally, it includes a report or interview transcripts on key findings from internal ethnographic fieldwork involving the stakeholders and engineers.

**4.3.1 Artifact: Stakeholder Map.** Who was involved in the system audit and collaborators in the execution of the audit should be outlined. Clarifying participant dynamics ensures a more transparent representation of the provided information, giving further context to the intended interpretation of the final audit report.

**4.3.2 Artifact: Ethnographic Field Study.** As Leveson points out, bottom-up decentralized decision making can lead to failures in complex sociotechnical systems [42]. Each local decision may be correct in the limited context in which it was made, but can lead to problems when these decisions and organizational behaviors interact. With modern large-scale artificial intelligence projects and API development, it can be difficult to gain a shared understanding at the right level of system description to understand how local decisions, such as the choice of dataset or model architecture, will impact final system behavior.

Therefore, ethnography-inspired fieldwork methodology based on how audits are conducted in other industries, such as finance [74] and healthcare [64] is useful to get a deeper and qualitative understanding of the engineering and product development process. As in internal financial auditing, access to key people in the organization is important. This access involves semi-structured interviews with a range of individuals close to the development process and documentation gathering to gain an understanding of possible gaps that need to be examined more closely.

Traditional metrics for artificial intelligence like loss may conceal fairness concerns, social impact risks or abstraction errors [69]. A key challenge is to assess how the numerical metrics specified in the design of an artificial intelligence system reflect or conform with these values. Metrics and measurement are important parts of the auditing process, but should not become aims and ends in themselves when weighing whether an algorithmic system under audit is ethically acceptable for release. Taking metrics measured in isolation risks recapitulating the abstraction error that [69] point out, "To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore to make a category error, or as we posit here, an abstraction error." What we consider data is already an interpretation, highly subjective and contested [23]. Metrics must be understood in relation to the engineering context in which they were developed and the social context into which they will be deployed. During the interviews, auditors should capture and pay attention to what falls outside the measurements and metrics, and to render explicit the assumptions and values the metrics apprehend [75]. For example, the decision about whether to prioritize the false positive rate over false negative rate (precision/recall) is a question about values and cannot be answered without stating the values of the organization, team or even engineer within the given development context.

## 4.4 The Artifact Collection Stage

Note that the collection of these artifacts advances adherence to the declared AI principles of the organization on "Responsibility & Accountability" and "Transparency".

In this stage, we identify and collect all the required documentation from the product development process, in order to prioritize opportunities for testing. Often this implies a record of data and model dynamics though application-based systems can include other product development artifacts such as design documents and reviews, in addition to systems architecture diagrams and other implementation planning documents and retrospectives.

At times documentation can be distributed across different teams and stakeholders, or is missing altogether. In certain cases, the auditor is in a position to enforce retroactive documentation requirements on the product team, or craft documents themselves.

The model card for the smile detection model is the template model card from the original paper [48]. A hypothetical datasheet for this system is filled out using studies on the CelebA dataset, with which the smile detector is built [44, 47]. In the model card, we identify potential for misuse if smiling is confused for positive affect. From the datasheet for the CelebA dataset, we see that although the provided binary gender labels seem balanced for this dataset (58.1% female, 42% male), other demographic details are quite skewed (77.8% aged 0-45, 22.1% aged over 46 and 14.2% lighter-skinned, 85.8% darker-skinned)[47].

The key artifact from auditors during this stage is the audit checklist, one method of verifying that all documentation pre-requisites are provided in order to commence the audit. Those pre-requisites can include model and data transparency documentation.

**4.4.1 Artifact: Design Checklist.** This checklist is a method of taking inventory of all the expected documentation to have been generated from the product development cycle. It ensures that the full scope of expected product processes and that the corresponding documentation required to be completed before the audit review can begin are finished. This is also a procedural evaluation of the development process for the system, to ensure that appropriate actions were pursued throughout system development ahead of the evaluation of the final system outcome.

**4.4.2 Artifacts: Datasheets and Model Cards.** Two recent standards can be leveraged to create auditable documentation, model cards and datasheets [24, 48]. Both model cards and datasheets are important tools toward making algorithmic development and the algorithms themselves more auditable, with the aim of anticipating risks and harms with using artificial intelligence systems. Ideally, these artifacts should be developed and/or collected by product stakeholders during the course of system development.

To clarify the intended use cases of artificial intelligence models and minimize their usage in contexts for which they are not well suited, Mitchell et al. recommend that released models be accompanied by documentation detailing their performance characteristics [48], called a *model card*. This should include information about how the model was built, what assumptions were made during development, and what type of model behavior might be experienced by different cultural, demographic or phenotypic groups. A

model card is also extremely useful for internal development purposes to make clear to stakeholders details about trained models that are included in larger software pipelines, which are parts of internal organizational dynamics, which are then parts of larger sociotechnical logics and processes. A robust model card is key to documenting the intended use of the model as well as information about the evaluation data, model scope and risks, and what might be affecting model performance.

Model cards are intended to complement "Datasheets for Datasets" [24]. Datasheets for machine learning datasets are derived by analogy from the electronics hardware industry, where a datasheet for an electronics component describes its operating characteristics, test results, and recommended uses. A critical part of the datasheet covers the data collection process. This set of questions are intended to provide consumers of the dataset with the information they need to make informed decisions about using the dataset: what mechanisms or procedures were used to collect the data? Was any ethical review process conducted? Does the dataset relate to people?

This documentation feeds into the auditors' assessment process.

## 4.5 The Testing Stage

This stage is where the majority of the auditing team's testing activity is done—when the auditors execute a series of tests to gauge the compliance of the system with the prioritized ethical values of the organization. Auditors engage with the system in various ways, and produce a series of artifacts to demonstrate the performance of the analyzed system at the time of the audit. Additionally, auditors review the documentation collected from the previous stage and begin to make assessments of the likelihood of system failures to comply with declared principles.

High variability in approach is likely during this stage, as the tests that need to be executed change dramatically depending on organizational and system context. Testing should be based on a risk prioritization from the FMEA.

For the smile detector, we might employ counterfactual adversarial examples designed to confuse the model and find problematic failure modes derived from the FMEA. For the child prediction model, we test performance on a selection of diverse user profiles. These profiles can also be treated for variables that correlate with vulnerable groups to test whether the model has learned biased associations with race or SES.

For the ethical risk analysis chart, we look at the principles and realize that there are immediate risks to the "Privacy" principle—with one case involving juvenile data, which is sensitive, and the other involving face data, a biometric. This is also when it becomes clear that in the smiling booth case, there is disproportionate performance for certain underrepresented user subgroups, thus jeopardizing the "Justice, Fairness & Non-Discrimination" principle.

The main artifacts from this stage of the auditing process are the results of tests such as adversarial probing of the system and an ethical risk analysis chart.

**4.5.1 Artifact: Adversarial Testing.** Adversarial testing is a common approach to finding vulnerabilities in both pre-release and post-launch technology, for example in privacy and security testing [6]. In general, adversarial testing attempts to simulate what a hostile actor might do to gain access to a system, or to push the limits of

the system into edge case or unstable behavior to elicit very-low probability but high-severity failures.

In this process, direct non-statistical testing uses tailored inputs to the model to see if they result in undesirable outputs. These inputs can be motivated by an intersectional analysis, for example where an ML system might produce unfair outputs based on demographic and phenotypic groups that might combine in non-additive ways to produce harm, or over time recapitulate harmful stereotypes or reinforce unjust social dynamics (for example, in the form of opportunity denial). This is distinct from adversarially attacking a model with human-imperceptible pixel manipulations to trick the model into misidentifying previously learned outputs [28], but these approaches can be complementary. This approach is more generally defined—encompassing a range of input options to try in an active attempt to fool the system and incite identified failure modes from the FMEA.

Internal adversarial testing prior to launch can reveal unexpected product failures before they can impact the real world. Additionally, proactive adversarial testing of already-launched products can be a best practice for lifecycle management of released systems. The FMEA should be updated with these results, and the relative changes to risks assessed.

**4.5.2 Artifact: Ethical Risk Analysis Chart.** The ethical risk analysis chart considers the combination of the likelihood of a failure and the severity of a failure to define the importance of the risk. Highly likely and dangerous risks are considered the most high-priority threats. Each risk is assigned a severity indication of "high", "mid" and "low" depending on their combination of these features.

Failure likelihood is estimated by considering the occurrence of certain failures during the adversarial testing of the system and the severity of the risk is identified in earlier stages, from informative processes such as the social impact assessment and ethnographic interviews.

## 4.6 The Reflection Stage

This phase of the audit is the more reflective stage, when the results of the tests at the execution stage are analyzed in juxtaposition with the ethical expectations clarified in the audit scoping. Auditors update and formalize the final risk analysis in the context of test results, outlining specific principles that may be jeopardized by the AI system upon deployment. This phase will reflect on product decisions and design recommendations that could be made following the audit results.

Additionally, key artifacts at this stage may include a mitigation plan or action plan, jointly developed by the audit and engineering teams, that outlines prioritized risks and test failures that the engineering team is in a position to mitigate for future deployments or for a future version of the audited system.

For the smile detection algorithm, the decision could be to train a new version of the model on more diverse data before considering deployment, and add more samples of underrepresented populations in CelebA to the training data. It could be decided that the use case does not necessarily define affect, but treats smiling as a favourable photo pose. Design choices for other parts of the product outside the model should be considered—for instance, an opt-in functionality with user permissions required on the screen before

applying the model-controlled function, and the default being that the model-controlled trigger is disabled. There could also be an included disclaimer on privacy, assuring users of safe practices for face data storage and consent. Once these conditions are met, Company X could be confident to greenlight developing this product for the client.

For the child abuse detection model—this is a more complex decision. Given the ethical considerations involved, the project may be stalled or even cancelled, requiring further inquiry into the ethics of the use case, and the capability of the team to complete the mitigation plan required to deploy an algorithm in such a high risk scenario.

**4.6.1 Artifact: Algorithmic Use-related Risk Analysis and FMEA.** The risk analysis should be informed by the social impact assessment and known issues with similar models. Following Leveson's work on safety engineering [42], we stress that careful attention must be paid to the distinction between the *designers' mental models* of the artificial intelligence system and the *user's mental model*. The designers' mental models are an idealization of the artificial intelligence system before the model is released. Significant differences exist between this ideal model and how the actual system will behave or be used once deployed. This may be due to many factors, such as distributional drift [41] where the training and test set distributions differ from the real-world distribution, or intentional or unintentional misuse of the model for purposes other than those for which it was designed. Reasonable and foreseeable misuse of the model should be anticipated by the designer. Therefore, the *user's mental model* of the system should be anticipated and taken into consideration. Large gaps between the *intended* and *actual* uses of algorithms have been found in contexts such as criminal justice and web journalism [12].

This adds complexity to anticipated hazards and risks, nevertheless these should be documented where possible. Christin points out “the importance of studying the practices, uses, and implementations surrounding algorithmic technologies. Intellectually, this involves establishing new exchanges between literatures that may not usually interact, such as critical data studies, the sociology of work, and organizational analysis”. We propose that known use-related issues with deployed systems be taken into account during the design stage. The format of the risk analysis can be variable depending on context, and there are many valuable templates to be found in *Failure Modes and Effects Analysis* (Section 3.1.3) framing and other risk analysis tools in finance and medical deployments.

**4.6.2 Artifact: Remediation and Risk Mitigation Plan.** After the audit is completed and findings are presented to the leadership and product teams, it is important to develop a plan for remediating these problems. The goal is to drive down the risk of ethical concerns or potential negative social impacts to the extent reasonably practicable. This plan can be reviewed by the audit team and leadership to better inform deployment decisions.

For the concerns raised in any audit against ethical values, a technical team will want to know: what is the threshold for acceptable performance? If auditors discover, for example, unequal classifier performance across subgroups, how close to parity is necessary to say the classifier is acceptable? In safety engineering, a risk threshold is usually defined under which the risk is considered

tolerable. Though a challenging problem, similar standards could be established and developed in the ethics space as well.

**4.6.3 Artifact: Algorithmic Design History File.** Inspired by the concept of the design history file from the medical device industry [77], we propose an algorithmic design history file (ADHF) which would collect all the documentation from the activities outlined above related to the development of the algorithm. It should point to the documents necessary to demonstrate that the product or model was developed in accordance with an organization's ethical values, and that the benefits of the product outweigh any risks identified in the risk analysis process.

This design history file would form the basis of the final audit report, which is a written evaluation by the organization's audit team. The ADHF should assist with an audit trail, enabling the reconstruction of key decisions and events during the development of the product. The algorithmic report would then be a distillation and summary of the ADHF.

**4.6.4 Artifact: Algorithmic Audit Summary Report.** The report aggregates all key audit artifacts, technical analyses and documentation, putting this in one accessible location for review. This audit report should be compared qualitatively and quantitatively to the expectations outlined in the given ethical objectives and any corresponding engineering requirements.

## 5 LIMITATIONS OF INTERNAL AUDITS

Internal auditors necessarily share an organizational interest with the target of the audit. While it is important to maintain an independent and objective viewpoint during the execution of an audit, we acknowledge that this is challenging. The audit is never isolated from the practices and people conducting the audit, just as artificial intelligence systems are not independent of their developers or of the larger sociotechnical system. Audits are not unified or monolithic processes with an objective "view from nowhere", but must be understood as a "patchwork of coupled procedures, tools and calculative processes" [74]. To avoid audits becoming simply acts of reputation management for an organization, the auditors should be mindful of their own and the organizations' biases and viewpoints. Although long-standing internal auditing practices for quality assurance in the financial, aviation, chemical, food, and pharmaceutical industries have been shown to be an effective means of controlling risk in these industries [76], the regulatory dynamics in these industries suggest that internal audits are only one important aspect of a broader system of required quality checks and balances.

## 6 CONCLUSION

AI has the potential to benefit the whole of society, however there is currently an inequitable risk distribution such that those who already face patterns of structural vulnerability or bias disproportionately bear the costs and harms of many of these systems. Fairness, justice and ethics require that those bearing these risks are given due attention and that organizations that build and deploy artificial intelligence systems internalize and proactively address these social risks as well, being seriously held to account for system compliance to declared ethical principles.

## REFERENCES

- [1] Omar Y Al-Jarrah, Paul D Yoo, Sami Muhaidat, George K Karagiannidis, and Kamal Taha. 2015. Efficient machine learning for big data: A review. *Big Data Research* 2, 3 (2015), 87–93.
- [2] Amel Bennaceur, Thein Than Tun, Yijun Yu, and Bashar Nuseibeh. 2019. Requirements Engineering. In *Handbook of Software Engineering*. Springer, 51–92.
- [3] Li Bing, Akintola Akintoye, Peter J Edwards, and Cliff Hardcastle. 2005. The allocation of risk in PPP/PFI construction projects in the UK. *International Journal of project management* 23, 1 (2005), 25–35.
- [4] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D Sculley. 2017. The ml test score: A rubric for ml production readiness and technical debt reduction. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1123–1132.
- [5] Shona L Brown and Kathleen M Eisenhardt. 1995. Product development: Past research, present findings, and future directions. *Academy of management review* 20, 2 (1995), 343–378.
- [6] Chad Brubaker, Suman Jana, Baishakhi Ray, Sarfraz Khurshid, and Vitaly Shmatikov. 2014. Using Frankencerts for Automated Adversarial Testing of Certificate Validation. In *In SSL/TLS Implementations,â€ IEEE Symposium on Security and Privacy*. Citeseer.
- [7] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25, 3 (2017), 273–291.
- [8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency*, 77–91.
- [9] Jenna Burrell. 2016. How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.
- [10] Paul Eric Byrnes, Abdullah Al-Awadhi, Benita Gullvist, Helen Brown-Liburd, Ryan Teeter, J Donald Warren Jr, and Miklos Vassarhelyi. 2018. Evolution of Auditing: From the Traditional Approach to the Future Audit 1. In *Continuous Auditing: Theory and Application*. Emerald Publishing Limited, 285–297.
- [11] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability, and Transparency*, 134–148.
- [12] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017), 2053951717718855.
- [13] Kai Lai Chung and Paul Erdős. 1952. On the application of the Borel-Cantelli lemma. *Trans. Amer. Math. Soc.* 72, 1 (1952), 179–186.
- [14] Rachel Courtland. 2018. Bias detectives: the researchers striving to make algorithms fair. *Nature* 558, 7710 (2018), 357–357.
- [15] Stephanie Cuccaro-Alamin, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review* 79 (2017), 291–298.
- [16] Michael A Cusumano and Stanley A Smith. 1995. Beyond the waterfall: Software development at Microsoft. (1995).
- [17] Nicholas Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. (2014).
- [18] Roel Dobbe, Sarah Dean, Thomas Gilbert, and Nitin Kohli. 2018. A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. *arXiv preprint arXiv:1807.00553* (2018).
- [19] Kevin Driscoll, Brendan Hall, Håkan Sivencrona, and Phil Zumsteg. 2003. Byzantine fault tolerance, from theory to reality. In *International Conference on Computer Safety, Reliability, and Security*. Springer, 235–248.
- [20] Danielle Ensign, Sorella Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847* (2017).
- [21] Virginia Eubanks. 2018. A child abuse prediction model fails poor families. *Wired Magazine* (2018).
- [22] Sellywati Mohd Faizal, Mohd Rizal Palil, Ruhanita Maelah, and Rosiati Ramli. 2017. Perception on justice, trust and tax compliance behavior in Malaysia. *Kasetsart Journal of Social Sciences* 38, 3 (2017), 226–232.
- [23] Jonathan Furner. 2016. “Data”: The data. In *Information Cultures in the Digital Age*. Springer, 287–306.
- [24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).
- [25] Jeremy Goldhaber-Fiebert and Lea Prince. 2019. Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny County’s Child Welfare Office. *Pittsburgh: Allegheny County.[Google Scholar]* (2019).
- [26] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 90–99.
- [27] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
- [28] Shixiang Gu and Luca Rigazio. 2014. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068* (2014).
- [29] John Haigh. 2012. *Probability: A very short introduction*. Vol. 310. Oxford University Press.
- [30] Brendan Hall and Kevin Driscoll. 2014. Distributed System Design Checklist. (2014).
- [31] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2018. Improving fairness in machine learning systems: What do industry practitioners need? *arXiv preprint arXiv:1812.05239* (2018).
- [32] IEEE. 2008. IEEE Standard for Software Reviews and Audits. *IEEE Std 1028-2008* (Aug 2008), 1–53. <https://doi.org/10.1109/IEEESTD.2008.4601584>
- [33] Kristen Intemann. 2010. 25 years of feminist empiricism and standpoint theory: Where are we now? *Hypatia* 25, 4 (2010), 778–796.
- [34] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. Artificial Intelligence: the global landscape of ethics guidelines. *arXiv preprint arXiv:1906.11668* (2019).
- [35] Paul A Judas and Lorraine E Prokop. 2011. A historical compilation of software metrics with applicability to NASA’s Orion spacecraft flight software sizing. *Innovations in Systems and Software Engineering* 7, 3 (2011), 161–170.
- [36] Emily Keddell. 2019. Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice. *Social Sciences* 8, 10 (2019), 281.
- [37] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).
- [38] Nitin Kohli, Renata Barreto, and Joshua A Kroll. 2018. Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems. In *1st Conference on Fairness, Accountability, and Transparency*. New York, NY, USA, 7.
- [39] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.
- [40] Arie W Kruglanski. 1996. Motivated social cognition: Principles of the interface. (1996).
- [41] Joel Lehman. 2019. Evolutionary Computation and AI Safety: Research Problems Impeding Routine and Safe Real-world Application of Evolution. *arXiv preprint arXiv:1906.10189* (2019).
- [42] Nancy Leveson. 2011. *Engineering a safer world: Systems thinking applied to safety*. MIT press.
- [43] Jie Liu. 2012. The enterprise risk management and the risk oriented internal audit. *Ibusiness* 4, 03 (2012), 287.
- [44] Zhiwei Liu, Ping Luo, Xiaogang Wang, and Xiaou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.
- [45] Amanda H Lynch and Siri Veland. 2018. *Urgency in the Anthropocene*. MIT Press.
- [46] Thomas Maillart, Mingyi Zhao, Jens Grossklags, and John Chuang. 2017. Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. *Journal of Cybersecurity* 3, 2 (2017), 81–90.
- [47] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. *arXiv preprint arXiv:1901.10436* (2019).
- [48] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.
- [49] Brent Mittelstadt. 2019. AI Ethics: Too Principled to Fail? *SSRN* (2019).
- [50] Brent Daniel Mittelstadt and Luciano Floridi. 2016. The ethics of big data: current and foreseeable issues in biomedical contexts. *Science and engineering ethics* 22, 2 (2016), 303–341.
- [51] Laura Moy. 2019. How Police Technology Aggravates Racial Inequity: A Taxonomy of Problems and a Path Forward. *Available at SSRN 3340898* (2019).
- [52] Fabian Muniesa, Marc Lenglet, et al. 2013. Responsible innovation in finance: directions and implications. *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*. Wiley, London (2013), 185–198.
- [53] Kristina Murphy. 2003. Procedural justice and tax compliance. *Australian Journal of Social Issues (Australian Council of Social Service)* 38, 3 (2003).
- [54] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.
- [55] Institute of Internal Auditors. Research Foundation and Institute of Internal Auditors. 2007. *The Professional Practices Framework*. Inst of Internal Auditors.
- [56] General Assembly of the World Medical Association et al. 2014. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *The Journal of the American College of Dentists* 81, 3 (2014), 14.
- [57] Cathy O’neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.
- [58] Charles Parker. 2012. Unexpected challenges in large scale machine learning. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*. ACM, 1–6.

- [59] Fiona D Patterson and Kevin Nealey. 2002. A risk register database system to aid the management of project risk. *International Journal of Project Management* 20, 5 (2002), 365–374.
- [60] W Price and II Nicholson. 2017. Regulating black-box medicine. *Mich. L. Rev.* 116 (2017), 421.
- [61] James Quesada, Laurie Kain Hart, and Philippe Bourgois. 2011. Structural vulnerability and health: Latino migrant laborers in the United States. *Medical anthropology* 30, 4 (2011), 339–362.
- [62] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*.
- [63] Clarence Rodrigues and Stephen Cusick. 2011. *Commercial aviation safety* 5/e. McGraw Hill Professional.
- [64] G Sirgo Rodriguez, M Olona Cabases, MC Martin Delgado, F Esteban Reboll, A Pobo Peris, M Bodí Saera, et al. 2014. Audits in real time for safety in critical care: definition and pilot study. *Medicina intensiva* 38, 8 (2014), 473–482.
- [65] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014).
- [66] David Satava, Cam Caldwell, and Linda Richards. 2006. Ethics and the auditing culture: Rethinking the foundation of accounting and auditing. *Journal of Business Ethics* 64, 3 (2006), 271–284.
- [67] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. 2014. Machine learning: The high interest credit card of technical debt. (2014).
- [68] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.
- [69] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 59–68.
- [70] Hetan Shah. 2018. Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018), 20170362.
- [71] Dominic SB Soh and Nonna Martinov-Bennie. 2011. The internal audit function: Perceptions of internal audit roles, effectiveness and evaluation. *Managerial Auditing Journal* 26, 7 (2011), 605–622.
- [72] Diomidis H Stamatis. 2003. *Failure mode and effect analysis: FMEA from theory to execution*. ASQ Quality press.
- [73] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42, 9 (2013), 1568–1580.
- [74] Alexander Styhre. 2015. *The financialization of the firm: Managerial and social implications*. Edward Elgar Publishing.
- [75] Alexander Styhre. 2018. The unfinished business of governance: towards new governance regimes. In *The Unfinished Business of Governance*. Edward Elgar Publishing.
- [76] JohnK Taylor. 2018. *Quality assurance of chemical measurements*. Routledge.
- [77] Marie B Teixeira, Marie Teixeira, and Richard Bradley. 2013. *Design controls for the medical device industry*. CRC press.
- [78] Manuel Trajtenberg. 2018. *AI as the next GPT: a Political-Economy Perspective*. Technical Report. National Bureau of Economic Research.
- [79] Frank Vanclay. 2003. International principles for social impact assessment. *Impact assessment and project appraisal* 21, 1 (2003), 5–12.
- [80] Tim Vanderveen. 2005. Averting highest-risk errors is first priority. *Patient Safety and Quality Healthcare* 2 (2005), 16–21.
- [81] Ajit Kumar Verma, Srividya Ajit, Durga Rao Karanki, et al. 2010. *Reliability and safety engineering*. Vol. 43. Springer.
- [82] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the AAAI/ACM Conference on AI Ethics and Society, Honolulu, HI, USA*, 27–28.
- [83] Yi Zeng, Enmeng Lu, and Cunqing Huangfu. 2018. Linking Artificial Intelligence Principles. *arXiv preprint arXiv:1812.04814* (2018).



*Unlocking Public Sector AI*

# AI Procurement in a Box: AI Government Procurement Guidelines

TOOLKIT  
JUNE 2020

# Contents

- 3 What is artificial intelligence (AI)?
- 4 Why do we need guidelines for public procurement of AI?
- 5 How were these guidelines developed?
- 6 How to use the guidelines
- 10 Guidelines overview
- 13 Detailed explanation of guidelines
- 27 Acknowledgements
- 29 Endnotes



1

# What is artificial intelligence (AI)?

AI has been formally defined as “technologies [that] aim to reproduce or surpass abilities (in computational systems) that would require ‘intelligence’ if humans were to perform them. These include: learning and adaptation; sensory understanding and interaction; reasoning and planning; optimization of procedures and parameters; autonomy; and creativity.”<sup>1</sup>

New AI approaches developed in the past decade, particularly the use of deep-learning neural networks, have dramatically advanced the capability of AI to recognize complex patterns, optimize for specific outcomes and make automated decisions. Doing this requires massive amounts of relevant data, a strong algorithm, a narrow domain and a concrete goal, and can result in dramatic improvements in reliability, efficiency and productivity.



## 2

# Why do we need guidelines for public procurement of AI?

Governments are increasingly seeking to capture the opportunities offered by AI to improve public-sector productivity and the provision of services to the public, and to stimulate the economy. AI holds the potential to vastly improve government operations and meet the needs of citizens in new ways, ranging from traffic management to healthcare delivery to processing tax forms. However, governments often lack experience in acquiring modern AI solutions and many public institutions are cautious about harnessing this powerful technology. Guidelines for public procurement can help in a number of ways.

First, government and the general public have justified concerns over bias, privacy, accountability, transparency and overall complexity. New examples are emerging of negative consequences arising from the use of AI in areas such as criminal sentencing, law enforcement and even employment opportunities. As citizens increasingly demand the same level of service from their governments as they do from innovative private-sector companies, public officials will be required not only to identify the specific benefits AI can bring, but also to understand the negative outcomes that can be generated.

Governments do not have the latitude of using the inscrutable “black box” algorithms that increasingly characterize AI deployed by industry. Without clear guidance on how to ensure accountability, transparency and explainability, governments may fail in their responsibility to meet public expectations of both expert and democratic oversight of algorithmic decision-making and may inadvertently create new risks or harms.

Governments rely on the expertise, and previously developed models, of technology providers and may lack the necessary skills to fully understand or trace algorithmic causality. Technology providers understand these challenges and look to governments to create clarity and predictability about how to manage them, starting in the procurement process. While companies are generally wary of stricter guidelines for government procurement, common-sense frameworks can help governments overcome reluctance to procure complex new technologies and actually open new markets for companies. Transparent guidelines will permit both established companies and new entrants to the AI space to compete on a level playing field for government contracts.

Second, AI procurement can build on a foundation of previous efforts to improve the effectiveness and efficiency of government technology procurement or be integrated into existing efforts. These may include legislation or policy measures such as frameworks or model contracts.

Established principles of good government technology procurement may take on added significance in AI procurement. For example, many governments already ensure that procurement efforts are run by multidisciplinary teams. Experience has shown that a lack of diversity in AI teams and positions of leadership has correlated with inadvertent harms or discrimination to vulnerable minority groups and protected classes. Given government’s role in upholding inclusion, an added emphasis on a multidisciplinary approach and diversity may be necessary in AI procurement.

**“New examples are emerging of negative consequences arising from the use of AI in areas such as criminal sentencing, law enforcement and even employment opportunities.”**

Some of the elements highlighted in the guidelines might already be evaluated in existing governance approaches but are not brought together holistically for decision-making. Closer working relationships between different teams should simplify the review of governance processes of AI systems even if they happen throughout different governance bodies and should integrate them in a strategy for AI adoption.

Third, as noted, AI has advanced rapidly in recent years, spurring further research and applications. New uses of AI that are of interest to governments will continue to emerge and will bring with them both benefits and risks. It is important that governments prepare for this future now by investing in building responsible practices for how they procure AI.

Finally, government procurement rules and purchasing practices often have a strong influence on markets, particularly in their early stages of development. As industry debates setting its own standards on these technologies, the government's moral authority and credibility can help set a baseline for these discussions.

Overall, the guidelines aim to guide all parties involved in the procurement life cycle – policy officials, procurement officers, data scientists, technology providers and their leaders – towards the overarching goal of safeguarding public benefit and well-being.

3

## How were these guidelines developed?

The guidelines were developed by the World Economic Forum Centre for the Fourth Industrial Revolution, in consultation with a multistakeholder community. Project fellows from the UK Government's Office for AI, Deloitte and Salesforce worked with Forum staff, and in partnership with Splunk-convened workshops with appropriate representatives from government, academia, civil society and the private sector to explore key issues and co-design responses.

4

# How to use the guidelines

The guidelines provide fundamental considerations that a government should address before acquiring and deploying AI solutions and services. They apply once it has been determined that the solution needed for a problem could be AI-driven. The guidelines are not intended as a silver bullet for solving all public sector AI-adoption challenges, but by influencing how new AI solutions are procured, they can set government use and adoption of AI on a better path.

**Specifically, the guidelines will help:**

- Policy officials to accelerate attainment of their policy goals
- Procurement officials and commercial teams to develop AI-related requests for proposals and to manage procurement processes

- Data practitioners (e.g. statisticians, data scientists, digital and technology experts) to safeguard public benefit and identify and manage potential risks
- AI-solutions providers to better understand the core expectations for government AI projects and to align their proposals with emerging standards for public procurement

The guidelines consist of 10 high-level recommendations, ordered roughly sequentially in terms of their relevance to the cumulative process of procurement, each containing:

- Multiple principles relating to each guideline
- Explanatory text elaborating on the thinking and substance underlying each principle

**⌚ This increases the risks and sensitivities about AI deployment in many use cases.**

It is important to approach AI procurement proportionality and not all guidelines may apply to the same extent to all procurement decisions. This is also why it is crucial to conduct an initial AI impact assessment and then act appropriately and proportional.

Important issues that can drive your decision whether to add additional ethical criteria to consider within your procurement approach, can fall within the following categories, many of which are closely interlinked. Note that this is not an exhaustive list of issues that need to be considered nor does it give you the answers whether your AI project might be more or less risky but it highlights key areas that need to be investigated further, particular in a public sector context.

### Key variables to consider in a risk assessment:

#### Data:

- **Data sensitivity** – The more sensitive the data that you are using within the AI system is, the more checks you should be building in. You need to closely consider if the data could be re-identified or give away any personal information.
- **Data quality** – The less sure you are about the quality of your data, the better it is to build in additional assurances to avoid bias and de-risk the project. Ensuring the representativeness of the data set might be difficult to ensure and qualitative measures might need to be taken. It is important to consider specific societal bias that could be reflected in the data for public sector use cases.
- **Data consent** – If meaningful personal data consent in the context that you are planning to use an AI-driven solution is not clear, the project is considered riskier. Also ensure that you are not inferring consent to a certain use of the data that does not comply with the original use case.

#### Field of use:

- **Public scrutiny** – If the project is within a sector of intense public scrutiny because of privacy concerns, legal concerns, interest and/or frequent litigation, the stakes are also higher. Fields, among others, such as health, social assistance, employment, financial services, insurance, the criminal justice systems, immigration, access and mobility, or decisions about permits and licences are examples of areas of applications that demand further considerations.

#### Socioeconomic impact:

- **Stakeholders involved** – The higher the impact on individuals, businesses, and

communities, the more important it gets to thoroughly consider AI ethics and scrutinize the application of AI.

- **Scope of impact** – It is important to consider factors such as how many people are impacted; how high the impact is and how high the likelihood of impact is. The risk also increases when decisions of the systems are linked to groups of people that are particularly vulnerable.

#### Financial consequences for agency and individuals:

- **Scope of financial impact** – The higher the potential financial consequences, the more you should address all areas linked to AI specific considerations.
- **Types of financial impact** – The financial consequences can be diverse and include monetary aspects as well as the access to credit, economic opportunities, schooling or training, insurance and certifications.

#### Impact of the AI system on your processes, employees and core businesses:

- **Core functions impact** – If the AI system is central to the core function of the agency, you should take on a more mandated approach to not only mitigate technical risks but also for reputational risk. The more tech dependence you create the riskier.
- **Business functions impact** – Consider whether you are replacing a business function rather than just improving and adding to the status quo, this might also impact your decision on how much to scrutinize the procurement process.
- **Job loss** – the more processes are automated, the more job losses can be expected. This increases the risks and sensitivities about AI deployment in many use cases.
- **Human in the loop** – The less checks and balances you have in place, the more risk. You should focus on adding explainability, interpretability and mindful friction to your AI deployment.

#### Example of tools that are already used within the public sector and the risks attached to this adoption:

- **Spam-filters in email programs** – designed to detect and block unwanted emails. Have the least risk prone use of AI in the public sector but can lead to discrimination if certain email addresses are blocked. However, “human in the loop” is usually included at various junctions so that the program isn’t

making decisions completely on its own, thus easily mitigating risk.

- **AI in cybersecurity solutions** – designed to protect networks, programs, and data from attack, damage, or unauthorized access. At first sight less prone to risks related to AI ethics, but we need to closely consider how the system is used in practice. If AI is used to better predict threats or identify cyber security risks, thus in a supporting function rather than making specific decisions, this use case seems to have a lower risk profile and thus would demand a less stringent approach to the implementation of all parts of the guidelines.
- **Chatbots** – designed to converse with people via voice interfaces or text messages. If they mainly provide information back to you and make it easier to sort through a large amount of data, rather than driving decisions, the use case seems to be less risk prone. But since they will likely be built into key processes and will have citizens interact with them, it is advised to follow the AI procurement guidelines to support those purchases.
- **Fraud detection** – designed to detect, prevent and manage fraudulent patterns in the data. Well tested use case of AI in the public sector, allows departments to make more effective enforcement decisions but the risk can be high if data quality is poor and if vulnerable groups are over proportionally targeted. False positive can also have high financial consequences and data sensitivity can be high depending on the use case. Hence, AI procurement guidelines should be followed.
- **AI in policing or social services** – designed to support and/or drive decisions in fields such as law enforcement, crime prevention, public safety, children welfare, social programs. The use of AI in those fields involves large risks as policy decisions are built into those systems and socioeconomic impacts are high. These use cases need to be put under particular scrutiny and procurement decisions need to follow very clear rules that include system testing, ethical considerations and a great focus on data governance. AI procurement guidelines should be closely followed.
- **AI in HR** – designed to take on key HR tasks including hiring, retaining talent, training, benefits and employee satisfaction. Employment decisions have high stakes with critical

consequences for individuals, organizations and society. Algorithms can make predictions in ways that disadvantage certain groups. Hence, concerns about AI algorithms bias and discrimination are particularly heightened, further complicated by labour and anti-discrimination laws. Finally, unique aspects of the human resources setting, including small datasets, complex social interactions, data privacy concerns and the need for accountability pose challenges and require close procurement guidelines governance.

#### Examples on how to do this:

1. [AI risk assessment tool](#): The tool aims to help you decide on a proportional approach to AI procurement. It sets out examples for decision criteria to include in a risk assessment of any potential solutions that contain AI capabilities. The tool outlines some of key questions you should consider when deciding your procurement strategy, considering what questions to ask in your RFP and assessing a solution.
2. [Alan Turing Institute, Understanding artificial intelligence ethics and safety](#): This guide is an end-to-end guidance on how to apply principles of AI ethics and safety to the design and implementation of algorithmic systems in the public sector. The ethical platform includes; a list of values that orient you in deliberating about the ethical permissibility and impact of a prospective AI project; a set of principles that all members of your project delivery team should be well-acquainted with and a framework that operationalizes these values and principles in an end-to-end workflow governance model.
3. [Canadian directive on automated decision-making](#): The Canadian government has developed a risk-based approach to AI adoption in the public sector which divides the AI systems in different levels. The four factors used to determine the risk-level are impact on: the rights of individuals or communities, the health or well-being of individuals or communities, the economic interests of individuals, entities, or communities and the ongoing sustainability of an ecosystem. Based on the risk-level, the guide provides insights on how to best approach AI procurement from a proportionality view and to what extent each requirement should be applied.  
Please refer to figure 1.

FIGURE 1

## Canadian Directive on Automated Decision-Making

Level	Description
01	<p>The decision will likely have little to no impact on:</p> <ul style="list-style-type: none"> <li>– The rights of individuals or communities.</li> <li>– The health or well-being of individuals or communities.</li> <li>– The economic interests of individuals, entities, or communities.</li> <li>– The ongoing sustainability of an ecosystem.</li> </ul> <p>Level 01 decisions will often lead to impacts that are reversible and brief.</p>
02	<p>The decision will likely have moderate impacts on:</p> <ul style="list-style-type: none"> <li>– The rights of individuals or communities.</li> <li>– The health or well-being of individuals or communities.</li> <li>– The economic interests of individuals, entities, or communities.</li> <li>– The ongoing sustainability of an ecosystem.</li> </ul> <p>Level 02 decisions will often lead to impacts that are likely reversible and short-term.</p>
03	<p>The decision will likely have high impacts on:</p> <ul style="list-style-type: none"> <li>– The rights of individuals or communities.</li> <li>– The health or well-being of individuals or communities.</li> <li>– The economic interests of individuals, entities, or communities.</li> <li>– The ongoing sustainability of an ecosystem.</li> </ul> <p>Level 03 decisions will often lead to impacts that can be difficult to reverse, and are ongoing.</p>
04	<p>The decision will likely have very high impacts on:</p> <ul style="list-style-type: none"> <li>– The rights of individuals or communities.</li> <li>– The health or well-being of individuals or communities.</li> <li>– The economic interests of individuals, entities, or communities.</li> <li>– The ongoing sustainability of an ecosystem.</li> </ul> <p>Level 04 decisions will often lead to impacts that are irreversible, and are perpetual.</p>

As the technological sophistication and government use of AI evolves, the guidelines should be updated to reflect new learning and leading practices. This is a living document that is intended to integrate feedback from practitioners over time. Much of that feedback will come from two sources: the project's community of subject matter experts, and the pilots to be held with the UK, the United Arab Emirates, Colombia and other partner governments. We also

welcome feedback from other stakeholders and the general public. If you wish to provide feedback, please share via email: [AI@weforum.org](mailto:AI@weforum.org).

Ultimately, the goal is that these guidelines will enable governments and international bodies to set the right policies, protocols and perhaps even standards to facilitate effective, responsible and ethical public use of AI.

5

# Guidelines overview

What are the key considerations when starting a procurement process, writing a request for proposal (RFP), and evaluating RFP responses?



## Guideline

## Principles

01

Use procurement processes that focus not on prescribing a specific solution but rather on outlining problems and opportunities, and allow room for iteration.

- a. Make use of innovative procurement processes to acquire AI systems.
- b. Focus on developing a clear problem statement, rather than on detailing specifications of a solution.
- c. Support an iterative approach to product development.

02

Define the public benefit of using AI while assessing risks.

- a. Set out clearly in your RFP why you consider AI to be relevant to the problem and be open to alternative technical solutions.
- b. Explain in your RFP that public benefit is a main driver of your decision-making process when assessing proposals.
- c. Conduct an initial AI risk and impact assessment before starting the procurement process, ensure that your interim findings inform the RFP, and revisit the assessment at decision points.

03

Align your procurement with relevant existing governmental strategies and contribute to their further improvement.

- a. Consult relevant governmental initiatives such as AI national strategies, innovation and/or industrial strategies, and guidance documents informing public policy about emerging technologies.
- b. Collaborate with other relevant government bodies and institutions to share insights and learn from each other.

04

Incorporate potentially relevant legislation and codes of practice in your RFP.

- a. Conduct a review of relevant legislation, rights, administrative rules and other relevant norms that govern the types of data and kinds of applications in scope for the project and reference them in the RFP.
- b. Take into consideration the appropriate confidentiality, trade-secret protection, and data-privacy best practices that may be relevant to the deployment of the AI systems.

05

Articulate the technical and administrative feasibility of accessing relevant data.

- a. Ensure that you have proper data governance mechanisms in place from the start of the procurement process.
- b. Assess whether relevant data will be available for the project.
- c. Define if and how you will share data with the vendor(s) for the procurement initiative and the subsequent project.
- d. Ensure that you have the required access to data used and produced by the vendor(s) solution.

## Guideline

## Principles

06

Highlight the technical and ethical limitations of intended uses of data to avoid issues such as historical data bias.

- a. Consider the susceptibility of data that could be in scope and if usage of the data is fair.
- b. Highlight known limitations (e.g. quality) of the data in the RFP and require tenderers to describe their strategies on how to address these shortcomings. Have a plan for addressing relevant limitations that you may have missed.

07

Work with a diverse, multidisciplinary team.

- a. Develop ideas and make decisions throughout the procurement process in a multidisciplinary team.
- b. Require the successful bidder(s) to assemble a team with the right skill set.

08

Focus throughout the procurement process on mechanisms of algorithmic accountability and of transparency norms.

- a. Promote a culture of accountability across AI-powered solutions.
- b. Ensure that AI decision-making is as transparent as possible.
- c. Explore mechanisms to enable interpretability of the algorithms internally and externally as a means of establishing accountability and contestability.

09

Implement a process for the continued engagement of the AI provider with the acquiring entity for knowledge transfer and long-term risk assessment.

- a. Consider during the procurement process that acquiring a tool that includes AI is not a one-time decision; testing the application over its lifespan is crucial.
- b. Ask the AI provider to ensure that knowledge transfer and training are part of the engagement.
- c. Ask the AI provider for insights on how to manage the appropriate use of the application by non-specialists.

10

Create the conditions for a level and fair playing field among AI solution providers.

- a. Reach out in various ways to a wide variety of AI solution providers.
- b. Engage vendors early and frequently throughout the process.
- c. Ensure interoperability of AI solutions and require open licensing terms to avoid vendor lock-in.

## 6

# Detailed explanation of guidelines

## 6.1

### Use procurement processes that focus not on prescribing a specific solution, but rather on outlining problems and opportunities and allow room for iteration.

#### Why is it important?

To acquire the AI systems that best address the challenge you want to address and encourage responsible innovation.

#### a. Make use of innovative procurement processes to acquire AI systems.

- Innovation-oriented procurement procedures provide opportunities to accelerate the adoption of new technologies such as AI systems, to promote innovation and to support secondary policy criteria such as support for small and medium-sized enterprises and the ethical development of AI.
- For example, these processes support early market engagement, enable you to go to market in different stages and can include the use of proofs of concept. These provide the opportunity to test the technologies on your problem area before making a final buying decision. Innovative public procurement processes that include practices such as detailing challenging problems, organizing technology contests, providing opportunities for demonstrators, and giving newly established providers the opportunity to compete for public-sector contracts, have the potential to boost innovation and help new companies

become established. This market-making role also encourages small enterprises with new ideas and reduces the risks for new technology start-ups.

- By strategically choosing the procurement approach depending on the nature of the challenge that you mean to address, these processes could include, for example:
  - Agile procurement processes that allow you to go to market in different stages and can include proofs of concept to test the technologies before the final purchase.
  - Challenge-based procurement processes that have vendors compete against each other based on their AI skills and include an evaluation of the technologies applied to the challenges they mean to address.
- Innovation partnerships that enable the procurement of technologies that cannot be delivered by the current options available to the market.
- Dynamic purchasing systems – procedures currently used mainly for products commonly available on the market – can accelerate uptake of technologies that are rapidly developing. As

## **● Encouraging collaboration between different bidders.**

a procurement tool, it is similar in some ways to an electronic framework agreement but, as new suppliers can join at any time, this allows newly established firms to participate in the framework agreements when they meet the set criteria.

- AI procurement frameworks that prescribe the terms and conditions applying to any subsequent contract and allow the pre-vetting of providers against a set of predefined criteria that can include ethical requirements.
- When making use of novel approaches to procuring emerging technologies you should also focus on best practices that have been shown to increase the supplier base of smaller and innovative suppliers, which is important for fast-developing markets such as AI. These practices include, but are not limited to:
  - Setting out and following a detailed procurement timeline at the start of the campaign.
  - Breaking down large proposals into smaller work components.
  - Encouraging collaboration between different bidders.

### **b. Focus on developing a clear problem statement, rather than on detailing the specifications of a solution.**

- AI technologies are developing rapidly, with new technologies and products constantly being introduced to the market. By focusing on describing the challenges and/or opportunities that you want to address and drawing on the expertise of technology partners, you can better

decipher what technology is most appropriate for the issue at hand. By focusing on the challenge and/or opportunity, you might also discover a higher-priority issue, or realize you were focusing on a symptom rather than the root cause.

- Beyond playing to each stakeholder's strength, this approach has two added benefits. First, it demands and promotes early market engagement, which we explain in further detail in Guideline 10. Second, it makes it easier for newer AI service providers (such as start-ups) to participate, as the government will not be focused on a specific product. Nurturing an emerging AI ecosystem is a key economic investment in the future.

### **c. Support an iterative approach to product development.**

- AI-powered solutions differ significantly from other technology tools in their unique ability to learn and adapt through ongoing, periodic training with new data. Therefore, the procurement process should allow room for iteration, while ensuring a robust, fair and transparent evaluation and decision process.
- For example, a phased challenge-based procurement could serve to evaluate different competitors' minimum viable products (MVPs) during phase one of procurement, with only the winner going on to develop the full solution. This building and testing in phases within the procurement cycle facilitates informed decision-making, innovation and transparency. It also provides you with relevant information to conduct meaningful impact assessments and evaluate risks.

**FIGURE 2 Visual to depict the challenge-based procurement process used by the UK GovTech Catalyst challenge**



01 | Eligible government organizations



02 | Submit eligible problems they need to be resolved



03 | Experts and GovTech Steering Group review and provide shortlist of 15



04 | Private companies offer answers



05 | Five companies receive up to £50,000 each for prototyping in 12 weeks



06 | Top two receive up to £500,000 each/develop product in 12 months



07 | All products available to public sector to buy

## 6.2

### Define the public benefit of using AI while assessing risks.

What do you expect such a system to achieve and be capable of, and what are the types of failure and harm that must be avoided?

#### Why is it important?

Defining the public benefit goal provides an anchor for the overall project and procurement process that the AI is intended to achieve. AI also brings new and specific risks that must be identified and managed as early as the procurement phase of the project.

**a. Set out clearly in your RFP why you consider AI to be relevant to the problem and be open to alternative technical solutions.**

- In most circumstances, you should refer to the need for an AI solution in your invitation to tender only if there is strong indication that the technology will address the problem that you are trying to solve. A need for the acquisition of an AI system should arise through analysis of policy challenges and alternatives, and be compared to other potential courses of action when the AI project does not have a clear research and innovation focus. If, during the evaluation of the tender responses, it becomes evident that another solution that doesn't incorporate AI is better able to address the problem, you should make the decision to follow this alternative delivery path.
- Assess whether AI could be part of a solution to your problem, before starting the procurement process. If you lack the capabilities in your team to carry out this assessment, you should seek these from elsewhere in your organization or relevant professional network (e.g. academia, trusted vendors) and make the consultation and collaboration with appropriate stakeholders a priority. For this assessment, it is crucial to engage a multistakeholder community to define and test a clear policy problem statement and reflect the findings in the RFP.
- Pre-market engagement is also often essential in helping you to describe your problem and narrow down the tasks that AI may be able to assist with. This will help you better communicate to potential suppliers what you are asking for and why, as well as highlighting where the gaps are. Documenting user need and challenges to the best of your ability is crucial, since the success of the project also depends on how well AI system providers understand the problem.

**b. Explain in your RFP that public benefit is a main driver of your decision-making process when assessing proposals.**

- When setting out the requirements in the RFP, you should consider explicitly referring to public benefit as well as user needs. When determining user needs, public servants should be confident

that they are acting in the public benefit. With regard to AI systems, the public benefit extends beyond value for money and also includes considerations about transparency of the decision-making process and other factors that are included in these guidelines.

- In practice this requires you, for example, to specify success and failure criteria in the context of public benefit: What do you expect such a system to achieve and be capable of, and what are the types of failure and harm that must be avoided? Conducting an impact assessment will help you to set these issues out. Refer to Guideline 7 for additional information on adding ethical requirements to the RFP.
- c. Conduct an initial AI risk and impact assessment even before starting the procurement process, ensure that your interim findings inform the RFP, and revisit the assessment at decision points.**
  - To better understand the potential impacts of the use of AI and to mitigate the risks, it is important to start an assessment in a systematic manner before the acquisition of an AI system and to make sure that the findings also inform your commercial strategy. There will be different considerations depending on which policy challenges you are trying to solve and which potential application of AI could help to address this challenge. Without knowing which AI system you will acquire, it is not possible to conduct a whole assessment.
  - An initial assessment should outline user needs and affected communities, as well as potential risks such as inaccuracy and bias of the AI system. It should also include some consideration of scenarios involving unintended consequences. The initial assessment should make you think about strategies to address these potential impacts, including but not limited to citizen panels, transparency reports and testing on differentially private or synthetic datasets. Associated risks and their respective mitigation strategies must be recognized by a suitable risk owner with decision-making power and should include a go/no-go decision.
  - In your invitation to tender, you should consider asking potential suppliers to identify risks and explain how they would mitigate them. This can give you valuable information regarding how careful each tenderer is and how aware they are of potential risks. Where you identified significant risks in your initial assessments, you should specifically require tenderers to set out how they would address those.

- Data protection impact assessments and equality impact assessments can provide a useful starting point for assessing potential unintended consequences. In assessing these, you should consider how the use of these systems, such as semi-automated or solely automated decisions, interact with mechanisms of oversight, review and other safeguards. We developed a high-level risk assessment, which allows you to make a more informed decision about your approach, and introduced the concept of a proportional approach to AI procurement. See the [AI risk assessment tool in the workbook](#). For other examples of risk assessment questionnaires for automated decision-making, refer to the government of Canada's [Directive on Automated Decision Making](#), and the framework on [Algorithmic Impact Assessments](#) from AI Now.
- In addition to the above, there should be systematic and continuous risk monitoring during every stage of the AI solution's life cycle, from design to post-implementation maintenance. AI solution providers can do this by identifying, drafting mitigations for and reporting risks through a project management function, which is central to the implementation (refer to Guideline 9 for more information on how to consider life-cycle management during the procurement process). The impact assessment should be revisited where necessary (e.g. in the event of significant changes to the opportunity statement).

## BOX 2

### Example of human rights assessment from Google Cloud

Google Cloud launched a Celebrity Recognition tool to a select set of media and entertainment customers to help them identify and label celebrities in professionally produced content, such as movies and sporting events. From the beginning of the product development process, they engaged in a human rights impact assessment (HRIA) and internal AI principles reviews. In partnership with BSR, a human rights non-profit organization, and using the UN's Guiding Principles on Business and Human Rights as a framework, the team considered potential impacts throughout numerous dimensions including privacy, discrimination, freedom of expression and many others. Aspects such as consultation with potentially affected stakeholders, dialogue with independent expert resources and paying particular attention to those at heightened risk of vulnerability or marginalization were part of the methodology. Their input played

an essential role in shaping the API's capabilities and the policies established around them.<sup>2</sup>

#### Some mitigation strategies adopted after this initial human rights risk assessment:

- Creation of “Service Specific Terms” that customers need to agree with. These limit the range of content upon which the API can be used and that address issues such as copyright, hate speech, child rights, surveillance and censorship.
- Adoption of a narrow definition of celebrity that respects the principle of informed consent by only including those that have actively and deliberately sought a role in public life.
- Creation of an “opt-out” option for celebrities not wanting to be included in Google’s celebrity database.

FIGURE 3

### Visual of the SDLC stages, with sample AI risk assessment question for each stage.

SDLC stage	Sample AI risk mitigation considerations
01 Requirements gathering and analysis	<ul style="list-style-type: none"> <li>– Is the use of AI/ML necessary for the desired outcome?</li> <li>– Should AI/ML even be discussed at this stage?</li> </ul>
02 Design	<ul style="list-style-type: none"> <li>– Do we have consent to use the data sources required by the solution?</li> <li>– Do we fully understand the implications of using external data, models or solutions?</li> </ul>
03 Implementation and coding	<ul style="list-style-type: none"> <li>– Do we have the right skills or domain expertise to develop the solution?</li> <li>– Does the development process protect data confidentiality and integrity?</li> </ul>
04 Testing	<ul style="list-style-type: none"> <li>– What level and type of bias is acceptable in the solution?</li> <li>– Do the acceptance criteria set appropriate levels of accuracy to ensure the model performance is satisfactory?</li> </ul>
05 Deployment	<ul style="list-style-type: none"> <li>– Have users received adequate training to ensure they understand the output of the system?</li> <li>– Is it transparent to users how the solution is deriving an output?</li> </ul>
06 Maintenance	<ul style="list-style-type: none"> <li>– Do the system administrators know what metrics to examine to validate that models are operating as expected?</li> <li>– Is there a clear process for updating or refining models using new data?</li> </ul>

## 6.3

### Aim to include your procurement within a strategy for AI adoption across government and learn from others.

Many countries are currently in the process of drafting and releasing national AI strategies, and some have already published theirs.

#### Why is it important?

To ensure that you use procurement strategically to support efforts on AI development and deployment, and to spread the knowledge of the public application of an emerging technology.

#### a. Consult relevant AI national strategy initiatives and guidance documents from ministries and departments informing public policy of emerging technologies.

- Many countries are currently in the process of drafting and releasing national AI strategies, and some have already published theirs. Prior to commencing an AI rollout, evaluate how your pursuit of acquiring an AI system aligns to your country's overall strategy.
- This allows you to include secondary policy aims in your strategic procurement and potentially make use of economies of scale by pooling the demand for AI systems. An added

benefit of aligning to a national AI strategy is that there may be special support for initiatives that align to the strategy, such as access to additional experts.

#### b. Consult with government agencies that have looked into procuring AI solutions, irrespective of the outcome of the procurement efforts.

- To improve your practices and share your experiences, you could actively seek out collaboration across departments and fields of expertise. You could also share knowledge and feedback via expert communities, such as the digital-buying community, professional networks or meet-ups.
- Within your department it can be helpful to set up platforms and networks that allow for the exchange of information, experiences and best practices about the purchasing of AI-powered solutions.

## 6.4

### Ensure that legislation and codes of practice are incorporated in the RFP.

#### Why is this important?

Conforming with existing laws and regulations ensures compliance; incorporating codes of practices supports the standardization of norms; and surveying the relevant rules enables better policy-making in a dynamic innovation technology ecosystem.

#### a. Conduct a review of relevant legislation, rights, administrative rules and other relevant norms that govern the types of data and kinds of applications in scope for the project.

- Conduct a review of relevant legislation, human and civil rights, administrative rules, and other relevant norms that govern the types of data and kinds of applications connected to the problem being addressed and solutions being proposed. Clarify the appropriate adjudicative and administrative jurisdictions within the domestic government in relation to conflicts of laws concerning the data. Depending on the problem being addressed in the invitation to tender, existing laws and regulations relevant to that government function may already have some rules on the use, processing, transfer etc. of data. Incorporate those rules and norms into the RFP by referring to the originating laws and regulations.

– When identifying the relevant rules, sources should include not only formal law, but also industry best practices, trade organization consensus guidelines and other forms of norm-setting mechanisms of soft law. For example, freedom of information laws<sup>3</sup> establish rules about what needs to be made available to the public, and data ethics frameworks guide the design of appropriate data use in government and the wider public sector.

#### b. Take into consideration the appropriate confidentiality, trade secret protection and data privacy best practices that may be relevant to the deployment of the AI solutions.

- To meaningfully evaluate proposed AI solutions, government officials must strike the right balance between preserving accountability through transparency and reassuring vendors that the trade secrets associated with their products and services, as well as their business confidentiality, will not be compromised. Information about government processes should be open by default, with the limits of disclosure justified in exceptional circumstances such as export controls, national security or ongoing criminal investigations.

- In those circumstances where confidentiality and trade-secrecy protection can be justified in light of public-interest considerations, investigate the possibilities of facilitating transparency through partial disclosure, limited review options and other means of enhancing public trust.

## 6.5

### **Articulate the technical feasibility and governance considerations of obtaining relevant data.**

#### **Why is this important?**

Availability of relevant data is a prerequisite for any AI solution, so time should not be spent discussing AI procurement if no data will be available. In addition, access to data should be granted only after careful consideration by the data-governing party(ies).

#### **a. Ensure that you have proper data-governance mechanisms in place from the start of the procurement process.**

- Set out a data-governance approach from the start of the procurement process. Given the importance and complexity of data governance, it is almost mandatory to implement sound data-governance processes before engaging in transformative AI projects. Governance needs to cover all data activities related to the proposed project, such as granting data access to project members, moving/storing data in other locations for analysis, and reviewing data consent (the purposes for which we are authorized to use the data).
- Data governance, and all other aspects of an AI initiative, require a multidisciplinary team. Refer to Guideline 7 for more information on multidisciplinary teams.
- In the absence of a data-governance framework, ensure that it is clear who is accountable (who is responsible for data management during the procurement process and the subsequent project).

#### **b. Assess whether relevant data will be available for the project.**

- Data is crucial for modern-day AI tools. You should determine, at a high level, data availability before starting your procurement process. This entails developing an understanding of what data might be required for the project. The idea is not to assess all possible data sources, but to build general awareness of data sources of potential interest. Data documentation, using data dictionaries, for example, is helpful when trying to build a high-level understanding of data assets.

- In cases where data is not available for the use case in mind, you may be able to find data through third parties, for example, vendors, partners or data brokers. If data is not available through any channel, engage skilled data scientists (for example, through vendors) to assess whether the use case can be addressed at all in a data-driven manner.

#### **c. Define if and how you will share data with the vendor(s) for the procurement initiative and the subsequent project.**

- Depending on the sensitivity of your project and data, it is worth considering the release of data to providers during procurement so that bidders can craft a response to the RFP that is tailored to your needs, with assumptions, timelines and fees that match your situation as closely as possible. This improves the quality of RFP responses you receive.
- If you are releasing data that is sensitive and not meant for public consumption, consider releasing only a sample, so that vendors have a clear idea of what the data enables them to do without having access to all of it. When doing this, make sure that you provide a sample that is representative of the overall dataset. Otherwise, vendors might make erroneous assumptions that can impact the quality of bids and consequently the integrity of the project.
- Create and document the appropriate data-sharing conditions. For example:
- Minimum requirements for the environment where the vendor will host the data (e.g. enterprise laptop that meets the vendor's standards for their sensitive data).
- Data consent form signed by the vendor's lead for the project, stating that the data will be used exclusively for the pursuit and for no other purpose. It should be clear to vendors that while in possession of the data they are not allowed to use the data for any purpose other than that specified in the RFP.
- Date for data deletion (e.g. immediately upon submission of the vendor's RFP response). In

 **Data is crucial for modern-day AI tools.**

<p>no circumstances should governments allow vendors to keep data after the procurement process, or after the conclusion of the project for successful bidders.</p> <ul style="list-style-type: none"> <li>- Confirmation of deletion of all data (e.g. written confirmation of deletion signed and submitted by the vendor's lead for the project).</li> <li>- There are many anonymization techniques to help safeguard data privacy, including data aggregation, masking and synthetic data.<sup>4</sup> Keep in mind, however, that you must manage anonymized data as carefully as the original data, since it may inadvertently expose important insights. RFPs should encourage innovative technological approaches, such as those mentioned above, that make less intrusive use of data or that achieve the same or similar outcomes with less sensitive datasets.</li> <li>- Certain vendors may have data that is complementary to the initiative, and it is in your best interest to consider using this data. It is important to have a framework that gives guidance regarding the circumstances under which it is reasonable to accept data from a vendor. Decision criteria could include: <ul style="list-style-type: none"> <li>- Vendor: some vendors could be pre-qualified as accepted data providers, be considered more trustworthy as a result of their previous track record as existing suppliers or have a strong reputation related to their data assets.</li> <li>- Domain: some domains – such as health, justice and immigration – are very sensitive. Use of third-party data in these domains requires careful scrutiny before it is accepted.</li> <li>- Data precedence and integrity: before using any third-party data, the government should have a clear understanding of how the data was collected, the governance processes employed to ensure its integrity, and whether the third party offering the data is legally allowed to commercialize it for the RFP.</li> </ul> </li> </ul>	<p><b>d. Ensure that you have the required access to data used and produced by the vendor(s) solution.</b></p> <ol style="list-style-type: none"> <li>1. Access and control of data used and produced by AI models is critical in monitoring, assessing and rectifying performance.</li> <li>2. You must ensure that you have access to raw input, processed/combined and enriched data produced by the vendor(s) AI models. This should also include third party and open source data, particularly if there is the chance that these will not be available/maintained on a long-term basis.</li> <li>3. Dependent on the solution(s) proposed the vendor(s) may not be willing or able to provide full access to all data (e.g. to protect IP for SaaS or COTS solution): <ul style="list-style-type: none"> <li>- Access to data should be provided with as wide a scope as possible. The supplier should be able to clearly articulate the reason for restricted sharing and this should be limited to only relevant areas not a blanket justification (e.g. commercially sensitive training sets do not preclude sharing enriched model outputs).</li> <li>- You should ensure that, where restricted access is justified, the supplier provides relevant, up-to-date and representative sampled data sets. Ideally these will be constructed from operational/live data.</li> </ul> </li> <li>4. Data ownership should be clearly articulated by the supplier: <ol style="list-style-type: none"> <li>4a. You should aim for contractual ownership of the data on a persistent basis.</li> <li>4b. As a minimum enriched data produced by the AI model(s) should be under “shared ownership” with access rights to all remaining data.</li> <li>4c. Ideally key data sets should be available for your internal teams to use learn and develop enhanced/new systems and approaches.</li> </ol> </li> </ol>
---	--

FIGURE 4

**Sample data governance framework**

Deloitte's data governance framework enables organizations to be specific in terms of what goals will be prioritized, what capabilities will be deployed and what results are expected



## 6.6

### Highlight the technical and ethical limitations of using the data to avoid issues such as bias.

Though available, legal to use and proportionate to need, there may be limitations to data (e.g. data bias) that make an AI approach inappropriate, unreliable or misleading.

#### Why is this important?

Though available, legal to use and proportionate to need, there may be limitations to data (e.g. data bias) that make an AI approach inappropriate, unreliable or misleading.

**a. Consider the susceptibility of data that could be in scope and whether usage of the data is fair.**

- As important as data protection is, not all data is sensitive (e.g. open-government data is freely accessible online). All data, sensitive or not, must have its integrity safeguarded, but it is not necessary to keep non-sensitive data behind closed doors. It is important to assess the privacy needs of different datasets to determine the right level of protection. Normally, personally identifiable information (PII), such as financial and health data, is considered extremely sensitive. The RFP needs to reflect data governance requirements for both the procurement process and the project that are in accordance with the nature of the data.
- Select data that fits criteria of fairness. For example, the data should be representative of the population that the AI solution will address, as well as being reasonably recent.<sup>5</sup>

**b. Highlight known limitations of the data (e.g. quality) in your RFP and require tenderers to describe their strategies on how to address these shortcomings. Have a plan for addressing relevant limitations that you may have missed.**

Considerations when deciding if a source of data is suitable include:<sup>6</sup>

- Representativeness (whether the data accurately represents the segment of the population in scope for the AI solution)
- Provenance (including how and why the data was collected)
- Gaps in data quality (e.g. many values missing from a particular data element)
- Bias present in the data (if it is not representative of the population to which the algorithm will be applied)
- Lack of clarity in metadata (for example, confusing or vague data element names)
- Check data completeness, representativeness and accuracy of potential sources before starting the procurement process. Articulate data quality observations and the apparent limitations and, if possible, share those insights through the RFP. Bidders must be aware of these data considerations during the procurement process or, in cases where data is sensitive, the selected provider(s) must be made aware after the contract has been awarded.
- If you do not have the right skills or means to comprehensively check for possible limitations of your data, provide vendors with guiding insights into the high-level state of the data and its origin,<sup>7</sup> so that they can draft adequate proposals. Also, ensure the RFP's data requirements include performing a comprehensive data quality assessment and, if required, development of mitigation strategies for low-quality data.

## 6.7

### Work with a diverse, multidisciplinary team.

#### Why is this important?

Developing and fulfilling a proper AI RFP will require a diverse team that understands the interdependent disciplines that AI covers, including: domain expertise (e.g. healthcare, transportation), systems and data engineering, model development (e.g. deep learning) and visualization/information design, among others.

**a. Develop ideas and make decisions throughout the procurement process in a diverse and multidisciplinary team.**

- Develop an understanding of the skills that are needed to effectively acquire and maintain an AI-powered solution, before starting the procurement process.
- Assemble multidisciplinary teams that specialize in designing, procuring, evaluating and operationalizing AI systems. These multidisciplinary teams should include expertise in: policy from the domain (e.g. justice) in which the AI solution will be applied, machine learning/data science, data engineering, technology (software and hardware), procurement, ethics and human rights.<sup>8</sup>

- Ensure that you have a diverse team. This should include people from different genders, ethnicities, socioeconomic backgrounds, disabilities and sexualities. You should also make sure that there is a mix of perspectives and points of view. This ensures that problems and solutions are tackled from different angles and helps to mitigate bias.
- This is important when it comes to evaluating tender responses. You need to be certain that you have the right expertise in your team to compare AI-driven solutions. Technical, business as well as legal and ethical experts are needed to score the different bids. You can integrate processes in your procurement decision to ensure that a multidisciplinary evaluation is mandatory. If expertise is lacking within your team, you can reach out to pools or professional networks within your organization or across the civil service.

Note that many value-laden decisions will likely be made during development (i.e. post-procurement), and it is essential that your team maintains the skills, or at the very least access to expertise, to ensure that all important decisions and trade-offs are made or overseen internally, rather than exclusively by a contractor or vendor.

**b. Require the successful bidder(s) to assemble a team with the right skill set.**

- As part of your requirements, ensure bidders provide evidence of the skills and qualifications of the proposed project resources who will develop and deploy the AI solution.<sup>9</sup> This should be part of the RFP response and it should be one of your decision criteria when evaluating the proposals.

## 6.8

### Focus throughout the procurement process on mechanisms of accountability and transparency norms.

#### Why is this important?

To build public trust in the legitimacy of the AI system, the procurement process should enable accountability in understanding how the AI solution works, so that it can be evaluated independently and thus promote a culture of responsibility over the AI solution life cycle.

**a. Promote a culture of accountability across AI-powered solutions.**

- Public institutions cannot rely on black-box algorithms to justify decisions that affect individual and collective citizens' rights, especially with the increased understanding about algorithmic bias and its discriminatory effects on access to public resources. There will be different considerations depending on the use case and application of AI that you are aiming to acquire, and you should plan to work with the supplier to explain the application for external scrutiny, ensuring your approach can be held to account. These considerations should link to the risk and impact assessment described in Guideline 2. Under certain scenarios, you could consider making it a requirement for providers to allow independent audit(s) of their solutions. This can help prevent or mitigate unintended outcomes.
- Providers and public officials should incorporate risk analysis for the unexpected and unintended effects of AI-powered solutions, within the limits prescribed by the law, and specify their respective responsibilities in the contract. Note that the laws and standards for assigning accountability may differ according to jurisdiction. For example, the Canadian federal

government's Directive on Automated Decision-Making requires the associate deputy minister of the respective federal entity to sign off on an algorithmic impact assessment (AIA) as part of an AI project.

- Consider how applicable accountability requirements in law, such as freedom of information legislation and data- protection logging requirements, will be implemented throughout the project life cycle.

**b. Ensure that AI decision-making is as transparent as possible.**

- Encourage transparency of AI decision-making (i.e. the decisions and/or insights generated by AI). One way to do this is to encourage the use of explainable AI. You can also make it a requirement for the bidder to provide the required training and knowledge transfer to your team, even making your team part of the AI-implementation journey. Finally, you can ask for documentation that provides information about the algorithm (e.g. data used for training, whether the model is based on supervised, unsupervised or reinforcement learning, or any known biases).
- Documentation is especially important when the algorithm is a pre-packaged solution that the bidder will bring to the project, as opposed to an algorithm that will be built and/or customized as part of the upcoming project. Finally, you can also ask bidders to provide information on their model-building methodology, including how they select variables, build samples (where applicable) and validate the model. Be aware, however, that algorithm-building is an iterative

process and that it depends on creativity as much as it does on science.

- Documentation provided by a bidder will give you directional awareness of their practices and methods; it will not give you a step-by-step guide that details exactly what would be done during the project, as the exact process will invariably shift from project to project to meet the needs of each scenario.

**c. Explore mechanisms to enable interpretability of the algorithms internally and externally as a means of establishing accountability and contestability.**

- With AI solutions that make decisions affecting people's rights and benefits, it is less important to know exactly how a machine-learning model has arrived at a result if we can show logical steps to achieving the outcome. In

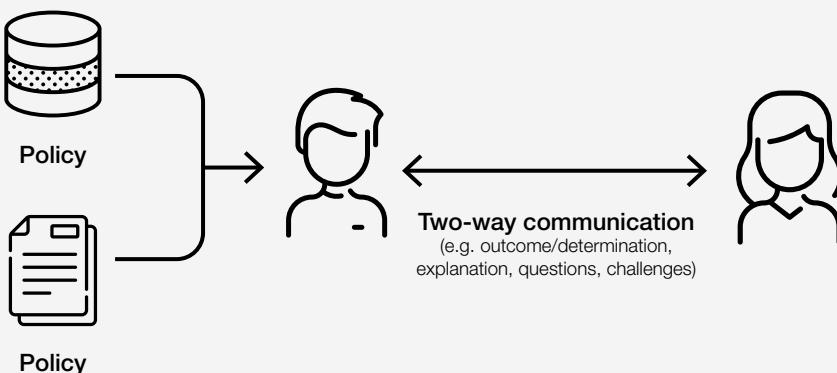
other words, the ability to know how and why a model performed in the way it did is a more appropriate means of evaluating transparency in the context of AI. For example, this might include what training data was used, which variables have contributed most to a result, and the types of audit and assurance the model went through in relation to systemic issues such as discrimination and fairness. This should be set out as documentation needed by your supplier.

- It is also important to consider the potential tension between explainability and accuracy of AI when acquiring AI solutions. Classic statistical techniques such as decision-tree models are easier to explain but might have less predictive power, whereas more complex models, such as neural networks, have high predictive power but are considered to be black boxes. Given these challenges you should think carefully about.

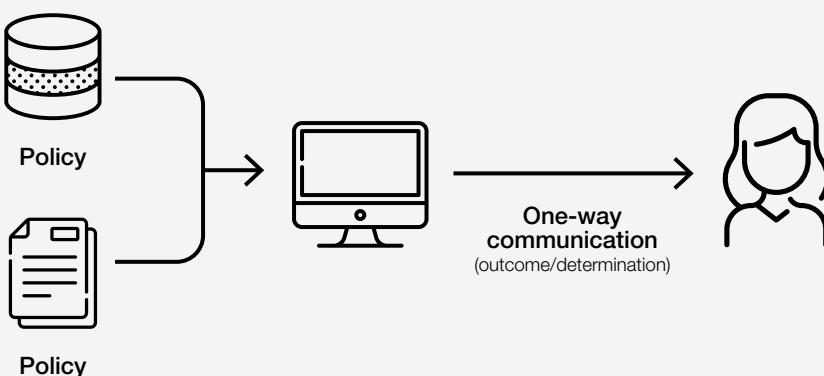
FIGURE 5

**Diagram to explain what is meant by a “black box” algorithm and why they’re an issue**

In a traditional model where a service provider interfaces with the service recipient, the recipient can communicate back and forth with the service provider regarding an outcome and/or determination. The recipient can ask questions regarding a decision and challenge an outcome.



With a fully automated system that uses a technique such as neural networks, the service recipient cannot expect to understand the outcome. This is because certain algorithms, such as neural networks, are very accurate but do not explain their path to a decision.



**BOX 3 | Sample type of documentation to ask for: [Google Model Cards](#)**

Machine learning models are often distributed without a clear understanding of how they function. For example, under what conditions does the model perform best and most consistently? Does it have blind spots? If so, where? Model cards address that issue by providing information about a model's performance and limitations. These "cards" are short documents accompanying trained machine learning models that provide benchmarked

evaluation in a variety of conditions. They are aimed at experts and non-experts alike. Developers can use them to make better decisions about what models to use for what purpose and how to deploy AI responsibly. For journalists and industry analysts, they might provide insights that make it easier to explain complex technology to a general audience and they might even help advocacy groups better understand the impact of AI on their communities.

**FIGURE 6 | Model card example**

Cards content	Sample information provided
<b>01</b> Overview of the model	<ul style="list-style-type: none"><li>- Simple description of the model</li><li>- Input data and output of the model</li><li>- Model architecture used</li></ul>
<b>02</b> Limitations	<ul style="list-style-type: none"><li>- Factors that might degrade the model's performance</li><li>- Situations in which the model might perform less than optimally</li></ul>
<b>03</b> Performance	<ul style="list-style-type: none"><li>- Model's performance on various evaluation datasets drawn from different sources than the training data</li></ul>

**BOX 4 | Solution to address explainability: example from [Google Cloud AI Explainability](#)**

The most useful models are often the most explainable, as they are the most trusted. Cloud AI Explanations help developers and enterprises understand why their AI model made the decisions it did by quantifying how each data factor contributes to the output. They can use this information to improve the models or share useful insights with their end users. The What-If tool, an interactive

visual interface, also allows users to investigate model behaviour by using dataset visualization to explain performance. AI Platform users can develop a deeper understanding of how their models work under different scenarios and build rich visualizations to explain model performance to business users and other stakeholders.

## 6.9

### Implement a process for the continued engagement of the AI provider with the acquiring entity for knowledge transfer and long-term risk assessment.

#### Why is this important?

The functionality and consequences of AI systems may not be apparent in the procurement process and often become evident only over the duration of its application, requiring extended communication and information-sharing between the procuring entity and the system developer.

For AI systems in the public sector, sustainable and ongoing evaluation methods and means of providing feedback on the data model are crucial to ensure that the tool's use remains ethical. You should make clear in your RFP that this should be considered by the provider and discussed as part of the procurement process.

#### a. Consider during the procurement process that acquiring a tool that includes AI is not a one-time decision; testing the application over its lifespan is crucial.

- The tool will need support during its life cycle. Knowing where to go for that support and how much support is available will be vital for getting the most out of any tool. Accepting the potential impact of any support gaps or employing outside expertise both come at a cost. This should be factored in when purchasing an intelligent tool.
- Consider the implementation of a process-based governance framework that provides a template for the integration of the norms, values and principles that inform the procedures and protocols organizing the project workflow.
- Testing the model on an ongoing basis is necessary to maintain its accuracy. An inaccurate model can result in erroneous decisions and affect users of public services.

Therefore, you should establish with the provider how the efficacy of the model will be monitored once deployed.

#### b. Ask the AI provider for knowledge transfer and training to be part of the engagement.

- Make knowledge transfer a requirement under the RFP. Evaluate the thoroughness and logic of the knowledge-transfer plan to ensure that government resources will be able to use the tool appropriately on their own once the project is finalized.
- Set out clearly your expectations for project documentation. Ensure that maintenance and

auditing of the AI solution would be possible by entities independent of the vendor.

#### c. Ask the AI provider for insights on how to manage the appropriate use of the application by non-specialists.

- Operational or service staff must have enough knowledge or training on the solution to understand how to use it and successfully exploit its outputs. You should address the need for enough training and support to avoid the misuse of AI applications with the AI provider. The application must make it easy to report any suspected unauthorized behaviour to the relevant authority(ies) within and/or outside the organization. Enable end-to-end auditability with a process log that gathers the data across the modelling, training, testing, verifying and implementation phases of the project life cycle. Such a log will allow for the variable accessibility and presentation of information with different users in mind to achieve interpretable and justifiable AI.

#### d. Make ethical considerations part of your evaluation criteria for proposals.

- There are robust ethical practices that you should require suppliers to demonstrate when providing AI solutions. Leading AI-solution providers have begun to create internal frameworks for the ethical design, development and deployment of AI, which cover processes to ensure accountability over algorithms, avoiding outputs of analysis that could result in unfair and/or biased decision-making, designing for reproducibility, testing the model under a range of conditions and defining acceptable model performance. Bidders should be able not only to describe their approach to the above, but also to provide examples of projects, complete with client references, where these considerations have been followed.<sup>10</sup>
- Make comprehensive, transparent algorithm assessment one of the requirements in the proposal and, if applicable, state minimum performance metrics that the model must meet. If possible, work with bidders to determine what the thresholds should be. As part of testing the model, you should work with the provider to establish how often you need to update the model with new data. Testing over the lifespan of the model for suitability and accuracy is highly important, especially when the AI is supporting critical services.

 **The tool will need support during its life cycle.**

 **Testing the model on an ongoing basis is necessary to maintain its accuracy.**

## 6.10

### Create the conditions for a level and fair playing field among AI solution providers.

#### Why is this important?

Government spending can be used to create a fair, competitive market, which leads to better AI systems. In addition, early engagement with AI vendors can result in more relevant responses, increasing the probability of success for the procurement and the subsequent project.

While AI systems generate new challenges that you need to reflect within the requirements and procurement approach, you must be proportionate in your approach and not impose any unnecessary burdens that would deter a wide diversity of suppliers, including small and medium sized enterprises (SMEs), Voluntary, Community and Social Enterprise (VCSE) suppliers and those owned by under-represented groups, from competing for public contracts.

#### a. Reach out in various ways to a wide variety of AI solution providers.

- Given the rapidly developing landscape of AI service providers, largely comprising smaller enterprises such as start-ups, consider non-traditional methods of market engagement to attract AI solution providers. For example, explain the needs that lead to the proposal through in-person presentations, webinars, information sessions at co-working spaces and/or online platforms such as LinkedIn or Twitter.
- Consider reaching out to non-traditional stakeholders, such as research institutes and academia. In some cases, these may have the right skills to be part of an AI implementation, and in all cases, they can act as advisers.<sup>11</sup>
- You should ensure that you have taken action to attract a wide diversity of suppliers to bid such SMEs, VCSEs and other under-represented businesses. You should test your approach to ensure it will not deter bidders or create unnecessary burdens on them either during the bidding process or during contract delivery. You must be proportionate in your approach.
- Keep in mind that successfully designing and deploying AI in organizations as big and complex as public agencies requires much more than technical expertise. It requires experience in change management, familiarity with public organizations, and the ability to manage complex projects.

#### b. Engage vendors early and frequently throughout the process.

- Market engagement is a process; it takes place prior to procurement and aims to identify

potential bidders and/ or solutions, build capacity in the market to address challenges and opportunities, and inform the design of the procurement and contract.

- Early engagement between government and industry is vital to a successful AI purchasing campaign. Early supplier engagement can help to determine the scope and feasibility of the RFP and, in turn, the most appropriate way to design and structure the requirements, increasing the likelihood that the winning bidder will meet your needs at a competitive cost. Ways to engage vendors early include having vendors provide inputs on possible evaluation criteria for the RFP, and hosting vendors to walk them through the RFP. Approaches like this are already being deployed in Canada, for example, and greatly help government and the private sector increase the effectiveness of procurement.
- To mitigate any risks that could be associated with market engagement (e.g. commercial confidentiality, protection of intellectual property [IP], fettering discretion of tender process), be sure to broadly advertise the engagement opportunity, allow all interested parties to participate, ensure that there is adequate time for responses and reasonable time for bidder selection and, where appropriate, that RFP responses can be marked as confidential.
- Ensure interoperability of AI solutions and require open licencing terms to avoid vendor lock-in.
- Consider strategies to avoid vendor lock-in, particularly in relation to black-box algorithms. These practices could involve the use of open standards, royalty-free licensing and public domain publication terms.
- During the design and deployment of the AI solution, it is likely that either a new algorithm will be designed, or an existing one will be tailored (e.g. retrained through your data). It is therefore useful to consider whether your department should own that IP and how it would control it. The arrangements should be mutually beneficial and fair, and require royalty-free licencing when adopting a system that includes IP controlled by a vendor.
- In order to preserve access to systems that become obsolete, ensure the ability to reverse-engineer the system to allow for maintenance of the AI solution independent of the vendor.

# Acknowledgements

The World Economic Forum's Unlocking Public Sector Artificial Intelligence project, in collaboration with the Government of the United Kingdom, Deloitte Consulting and Splunk is a global, multistakeholder and cross-disciplinary initiative intended to help shape the public sector's adoption of AI, and emerging technologies in general, around the world. The project has engaged leaders from

private companies, governments, civil society organizations and academia to understand public-sector procurement of AI technology, identify challenges and define principles to guide responsible and ethical procurement. The opinions expressed herein may not correspond with the opinions of all members and organizations involved in the project.

## Lead authors:

### **Sabine Gerdon**

Artificial Intelligence and Machine Learning Fellow, World Economic Forum, Seconded from the Office for Artificial Intelligence, Government of the United Kingdom

### **Eddan Katz**

Project Lead, World Economic Forum

### **Emilie LeGrand**

McGill University Integrated Management Student Fellow

### **Gordon Morrison**

Director of EMEA Government Affairs, Splunk Inc.

### **Julián Torres Santeli**

Artificial Intelligence and Machine Learning Fellow, World Economic Forum, Seconded from Deloitte Canada's AI practice

We would like to thank our Unlocking Public-Sector AI project community as well as the following contributors for their insights:

### **Rashid Alahmedi**

Senior Specialist Technology and Solutions, Dubai Electricity and Water Authority

### **Greg Ainslie-Malik**

Machine Learning Architect, Splunk Inc.

### **Jesus Alvarez-Pinera**

Head of Data, Food Standards Agency

### **Shelby Austin**

Managing Partner, Growth and Investments and Omnia AI, Deloitte

### **Yousef Al-Barkawie**

Partner, Analytics and Cognitive Middle East Leader, Deloitte

### **Neil Barlow**

Head of Vehicle Policy and Engineering, Driver and Vehicle Standards Agency

### **Kathy Baxter**

Architect, Ethical AI Practice, Salesforce

### **Lorena Cano**

Digital Trade Fellow, World Economic Forum from Inter-American Development Bank

### **Ashley Casovan**

Executive Director, AI Global

### **Michael Costigan**

Artificial Intelligence and Machine Learning Fellow, World Economic Forum from Salesforce

### **Sue Daley**

Associate Director, techUK

### **Nihar Dalmia**

Government and Public Sector AI leader for Deloitte Canada, Deloitte

### **Gourav Dhiman**

Business Development Manager, XLPAT

### **Cosmina Dorobantu**

Deputy Director of Public Policy Programme, The Alan Turing Institute

### **Leslie Harper**

Senior Sector Specialist, Inter-American Development Bank

### **James Hodge**

Chief Technical Adviser, Splunk Inc.

### **Hamad Karam**

Senior Specialist Artificial Intelligence, Dubai Electricity and Water Authority

### **Andrew Kim**

Head of AI Policy, Google Cloud

<b>Steven Knight</b> AI Lead, Food Standards Agency	<b>Nada Al-Saeed</b> Data Policy Fellow, World Economic Forum from Bahrain Economic Development Board
<b>Benjamin Leich</b> Economic Adviser, Better Regulation Executive	<b>Komal Sharma Talwar</b> Director, XLPAT and TT Consultants
<b>Katherine Mayes</b> Programme Manager, techUK	<b>Leonard Stein</b> Senior Strategic Adviser, Splunk Inc.
<b>Maha Mofeed</b> Chief Corporate Officer, Bahrain Economic Development Board	<b>Jitin Talwar</b> Founder, XLPAT and TT Consultants
<b>Valesca Molinari</b> Automotive and Autonomous Mobility Fellow, World Economic Forum from Baker McKenzie	<b>Sandeep Singh Kohli</b> Co-founder, XLPAT
<b>Mariam Al Muhairi</b> Head, Centre for the Fourth Industrial Revolution United Arab Emirates	<b>Ahmad Al Tawallbeh</b> Specialist Artificial Intelligence, Dubai Electricity and Water Authority
<b>Khalid Al Mutawa</b> Director, Bahrain Information and eGovernment Authority	<b>Abbey Thornhill</b> Assistant Economist, Better Regulation Executive
<b>Brandie Nonnbecke</b> Founding Director, CITRIS Policy Lab	<b>Adrian Weller</b> Programme Director for AI, The Alan Turing Institute
<b>Arwa Al Qassim</b> AI Lead, Centre for the Fourth Industrial Revolution United Arab Emirates	<b>Mark Woods</b> Director, Technology and Innovation, Splunk Inc.
<b>Ana Rollan</b> Artificial Intelligence and Machine Learning Fellow, World Economic Forum from BBVA	<b>Tim Woodbury</b> Director of State and Local Government Affairs, Splunk Inc.

Thank you also to the teams in the UK from the Defence Science and Technology Laboratory, the Department for Transport, the Home Office Accelerated Capability Environment and local governments that supported the user testing and piloting. The steering and working group from the Department of Digital, Culture, Media and Sport, the Government Digital Service, the Cabinet Office, the Crown Commercial Service and the Centre for Data Ethics and Innovation has been instrumental to progressing this work, in particular:

<b>Sue Bateman</b> Deputy Director for Policy and Innovation, Government Digital Service	<b>Stephen Hennigan</b> Deputy Head of Office for Artificial Intelligence, United Kingdom Government
<b>Oliver Buckley</b> Executive Director, Centre for Data Ethics and Innovation	<b>Sana Khareghani</b> Head of Office for Artificial Intelligence, United Kingdom Government

Thank you to everyone who contributed through interviews, workshops and discussions in the last 18 months in Dalian, Dubai, London, Manama, San Francisco, Tianjin, Toronto and Washington DC.

## 8

# Endnotes

1. Definition from the Engineering and Physical Science Research Council, a UK government research funding body.
2. For the complete results, see <https://services.google.com/fh/files-bsr-google-cr-api-hria-executive-summary.pdf>
3. For an up-to-date list of freedom of information laws around the world, see [https://en.wikipedia.org/wiki/Freedom\\_of\\_information\\_laws\\_by\\_country](https://en.wikipedia.org/wiki/Freedom_of_information_laws_by_country) (link as of 29.05.2020).
4. For more information on data anonymization, refer to: “Guide to basic data anonymisation techniques”, Personal Data Protection Commission, Singapore. 25 January 2018.
5. For more information on fairness during data selection, refer to: “Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector”, section “Data fairness”, David Leslie, the Alan Turing Institute.
6. For more information on data selection criteria, refer to: “Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector”, section “Data fairness”, David Leslie, the Alan Turing Institute.
7. For example, summary statistics such as number of rows present, number of missing values for each data field, description of how the data is collected and processed.
8. For more information on the domain and technical skills required to deliver an AI engagement, refer to: “Searching for superstars isn’t the answer. How organizations can build world-class analytics teams that deliver results”, Deloitte.
9. ibid.
10. AI ethics is a deep and evolving field, and there are various publications on the matter, including those listed below. Refer to these sources for a full background on the topic.
  - “OECD principles on artificial intelligence”, Organizations for Economic Co-operation and Development
  - “Ethics guidelines for trustworthy AI”, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission
  - Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector”, section “Data fairness”, David Leslie, the Alan Turing Institute.
  - “For a meaningful artificial intelligence. Towards a French and European strategy”, Cédric Villani
11. Examples of organizations include the Alan Turing Institute in the UK and the Vector Institute, MILA, and the Alberta Machine Intelligence Institute in Canada.



---

COMMITTED TO  
IMPROVING THE STATE  
OF THE WORLD

---

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

---

World Economic Forum  
91–93 route de la Capite  
CH-1223 Cologny/Geneva  
Switzerland

Tel.: +41 (0) 22 869 1212  
Fax: +41 (0) 22 786 2744  
[contact@weforum.org](mailto:contact@weforum.org)  
[www.weforum.org](http://www.weforum.org)



*Unlocking Public Sector AI*

# AI Procurement in a Box: Workbook

TOOLKIT  
JUNE 2020

# Contents

- 3** A: AI risk assessment tool
- 9** B: User manual
- 18** C: AI specification and evaluation tool
- 30** D: How to kick-start the implementation of the guidelines
- 32** E: Case studies
- 51** Acknowledgements
- 53** Endnotes



A

# AI risk assessment tool





## A.1 Overview

This document sets out example decision criteria for conducting an artificial intelligence (AI) project risk assessment. An assessment of the potential risks involved in any solution that contains AI elements should be conducted as part of the planning phase of an AI procurement. This can also be a

useful basis to develop a proportionate approach to AI procurement. It is important to approach AI procurement proportionally because not all guidelines as well as issues explored in guidelines apply to all procurement decisions in the public sector.

### Purpose of this tool

The following table outlines some of the key questions you should consider when deciding your procurement strategy, choosing what requirements to include in your request for proposal (RFP) and assessing a

solution. These questions have also been mapped to the issues that were set out in the *guidelines for AI procurement* document under the risk assessment header in the how to use the guide section.

### How to use this tool

All these questions are designed to be answered with a yes or a no. Note that the list is not exhaustive and you should consider additional risks that are specific to your organization. For some of the questions below it might also be useful to consult the risk-based approach to AI adoption from the

Canadian public sector, which divides AI systems developed at different levels. These categorizations provide insights into how to best approach AI procurement from a proportionality view and will help govern some of the decision-making.

**Q1**

**Is the solution intended for use in an area of public interest?**

If the project is within an area of intense public scrutiny (e.g. because of privacy concerns), interest and/or frequent litigation, then additional controls may be required. Fields such as health, social assistance, access and mobility, or decisions about permits and licences are examples of areas of applications that demand further consideration.

The higher the impact on individuals, businesses and communities, the more important it becomes to thoroughly consider AI ethics. The risk also increases when decisions made by the systems are linked to groups of people that are particularly vulnerable.

**Q2**

**Does the data used or generated by the solution contain any biographical or sensitive information?**

The more sensitive the data used or generated within an AI system the greater the number of checks you should build in.

**Q3**

**Are you comfortable with the data being stored and processed in an externally hosted solution?**

Consider whether the data has any protective markings or handling requirements that necessitate storage on authority infrastructure, such as a fully managed data centre or within a private cloud environment.

If your organization has a cloud-first policy and the data is suitable, a SaaS solution may be appropriate.

**Q4**

**Do you need to understand the details of how the data is being processed?**

For low-risk applications it might be appropriate to consider solutions that provide limited insight into how the data is processed, but if the solution is intended for processing personal information (such as medical applications), it may be useful to know the details of how it's been processed to ensure the outcome can be explained.

**Q5**

**Do you need the results of the processing to be validated by a human or is an automated output acceptable?**

If the output of the solution is intended for making critical decisions about services that are provided directly to citizens, then validation of the output is necessary. Alternatively, if you are considering a solution for managing cloud infrastructure to ensure the performance of a given application it might be appropriate for this to be fully automated.

**Q6**

**Do you have the skills and knowledge to define and assess the performance of the solution?**

Depending on the levels of expertise within your organization you may need to rely more heavily on a supplier or vendor to curate the solution for you. In this case you should expect the supplier to provide more detailed information about how they manage the solution.

If you have strong organizational data science skills, however, you should be able to more easily set the performance parameters, which makes custom solutions more achievable.

**Q7**

**Are you confident that the data intended for use in the solution is of good quality?**

The less sure you are about the quality of your data, the better it is to build in additional assurances to avoid bias.

**Q8**

**Are you happy for the supplier or vendor to enrich the data with external information as part of the processing?**

Some solutions will use external data feeds to draw conclusions from your data, and the source and utility of this external data should be considered when assessing what is acceptable for your organization.

The following table links the issues set out in the *guidelines for AI procurement* document to the most relevant questions.

FIGURE 1 **Mapping guideline topics to the risk assessment tool**

Issue	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Data								
Field of use								
Socio-economic impact								
Financial consequences for agency and individuals								
Business function of the AI system								

## A.2 Essential requirements in a proportionate approach

The following table outlines how the answers to the questions relate to the requirements described in the workbook Part C AI Procurement Specification and Evaluation Tool. It highlights the most important requirements related to the risk assessment. Please note that this does not mean that other requirements aren't also essential.

FIGURE 2 How risk assessment relates to AI-specific RFP requirements

		Essential requirements	Additional requirements
Q1	Is the solution intended for use in an area of public interest?	1.1	If Yes: Add more weight to 1.1
Q2	Does the data used or generated by the solution contain any biographical or sensitive information?	4.2, 4.3	If Yes: 4.4
Q3	Are you comfortable with the data being stored and processed in an externally hosted solution?	2.2	If Yes: 3.1, 3.2
Q4	Do you need to understand the details of how the data is being processed?	1.4, 1.7, 4.1	If Yes: 1.2, 1.3, 1.5, 7.1, 8.1
Q5	Do you need the results of the processing to be validated by a human or is an automated output acceptable?	1.6, 2.3, 9.1, 9.2	
Q6	Do you have the skills and knowledge to define and assess the performance of the solution?	3.3, 6.1, 9.3	If No: 3.4, 5.2, 5.3, 4.5, 10.1
Q7	Are you confident that the data intended for use in the solution is of good quality?		If No: 4.1
Q8	Are you happy for the supplier or vendor to enrich the data with external information as part of the processing?		If Yes: 2.1

## A.3 Risk matrix

The risk matrix is designed to help the user determine their hosting and processing risks and what this means in terms of what types of solutions can be considered.

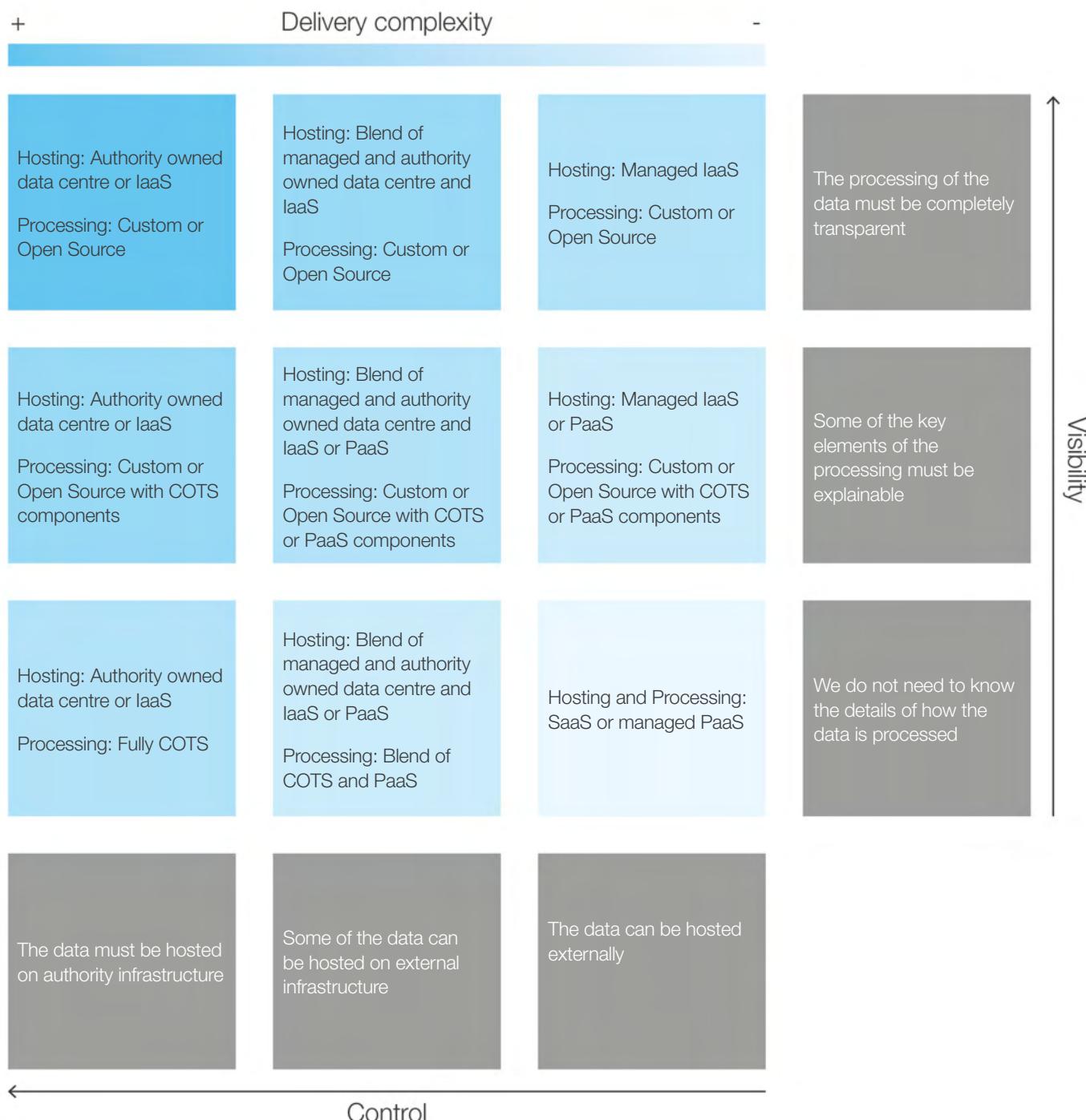
When considering the risks, you should:

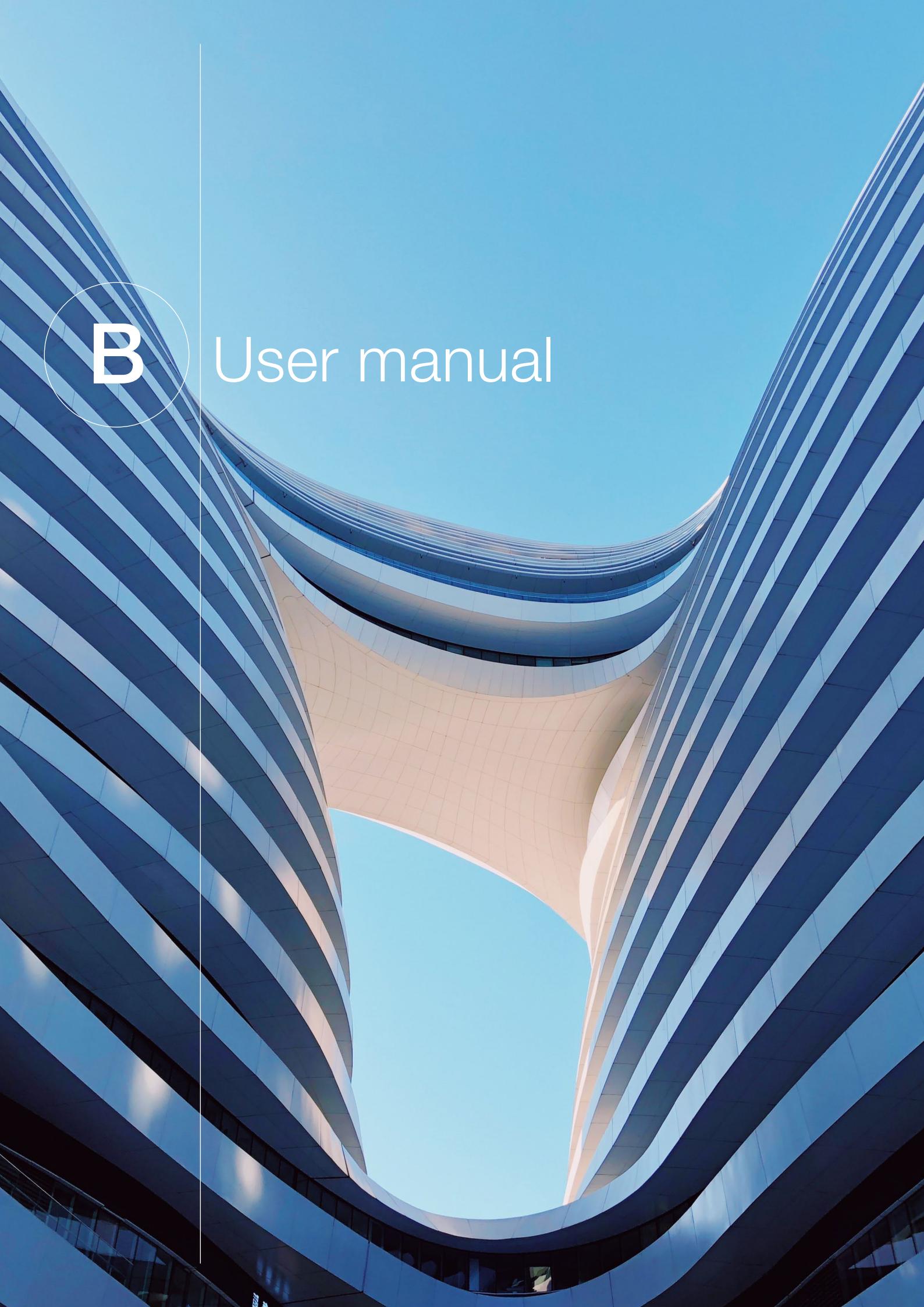
- For control (or hosting) risks: consider your answers to questions 2 and 3 above.
- For visibility (or processing) risks: consider your answers to questions 4, 5 and 6 above.

Depending on your control and visibility posture the diagram will help you determine what solutions may be appropriate. For example, if all of the data can be hosted externally and you do not need visibility of the processing a SaaS offering could be appropriate. Note that for any box you land on from a visibility and control perspective, solutions that fit types above and/or to the left would also be appropriate, but bring a higher delivery risk.

For clarity you can find definitions for Open Source<sup>1</sup>, COTS<sup>2</sup>, IaaS<sup>3</sup>, PaaS<sup>4</sup> and SaaS<sup>5</sup> from the links found in the endnotes section.

FIGURE 3 Risk matrix





B

# User manual

# Overview:

Key factors to consider when beginning the procurement process for an AI-enabled solution

This manual provides a set of questions that highlight the main considerations that users should be able to address when implementing the guidelines.

1

## Use procurement processes that focus not on prescribing a specific solution but rather on outlining problems and opportunities, and allow room for iteration.

### Purpose of this tool

The user manual should help users to work through the different guidelines and find out how they apply to the specific project that they are working on.

### How to use this tool

You can use the questions as a checklist at start of your procurement process.

#### 1a. Make use of innovative procurement processes to acquire AI systems.

- Does your agency have access to a procurement vehicle(s) developed specifically for innovative technologies, such as AI?
- Have you engaged peers who have leveraged this procurement vehicle(s) in the past, whether inside or outside your agency, to learn from their experience?
- Are you leveraging any special mechanisms made available by the procurement process, such as agile procurement, challenge-based procurement, and/or dynamic purchasing systems?
- Does the procurement vehicle allow the procurement team to evaluate responses within a reasonable amount of time, so as not to exclude potential participants?

#### 1b. Focus on developing a clear problem statement, rather than detailing the specifics of a solution.

- Do you have a clear, concise problem statement that focuses on the needs of a user (e.g. benefit applicants)?
- Have you phrased your problem in a way that is technology agnostic?
- Have you engaged a group of peers and market partners, preferably knowledgeable in human-centric design, to confirm that you are addressing the root cause of the problem, as opposed to a symptom?

#### 1c. Support an iterative approach to product development.

- Can you set expectations with providers through the RFP that the project must be delivered using an iterative (e.g. agile) approach?
- Can the problem be broken down into more manageable contracts and projects?

## Define the public benefit of using AI, while assessing risks.

**2a. Set out clearly in your RFP why you consider AI to be relevant to the problem and be open to alternative technical solutions.**

- Do you have strong indications that AI is applicable to the problem? (e.g. do you have large amounts of data you could use to derive insights that address the problem?)
- Can the problem be addressed through a technology/solution that is likely to be better understood by the resources who will be responsible for delivering and operating it?
- Have you engaged peers and vendors to confirm that AI is a good solution to the problem?

**2b. Explain in your RFP that public benefit is a main driver of your decision-making process when assessing proposals.**

- Have you identified the protected groups, whether internal or external, who would be affected by the decision-making of the AI solution?
- Have you identified the potential biases that could exist in the data, which could unfairly affect the protected groups previously identified?
- Have you engaged the parties who will be affected by the tool and obtained their inputs (e.g. by holding citizen panels)?
- Have you identified success and failure criteria for the solution from the perspective of the stakeholders who would be affected by the solution?

**2c. Conduct an initial AI risk and impact assessment before starting the procurement process, ensure that your interim findings inform the RFP and revisit the assessment at decision points.**

- Have you identified the high-level potential impacts, including unanticipated consequences, that a solution could have on stakeholders? For example, for an AI-driven unemployment solution, could eligible recipients be wrongfully denied the benefit?
- Have you documented these potential impacts, together with viable mitigation strategies?
- Has executive management signed off the impact assessment?
- Have you included the results of the impact assessment in the RFP and asked vendors to suggest mitigation strategies?

**3**

### **Align your procurement with relevant existing government strategies and contribute to their further improvement.**

#### **3a. Consult relevant government initiatives, such as AI national strategies, innovation and/or industrial strategies and guidance documents informing public policy about emerging technologies.**

- Have you identified relevant national strategies (e.g. AI strategy, digital strategy) and evaluated how your project can align?
- Have you identified and consulted on relevant policies and guidance frameworks, whether internal or external (e.g. innovation policies, technology policies, data policies and industry norms)?

#### **3b. Collaborate with other relevant government bodies and institutions to share insights and learn from each other.**

- Have you consulted peers, inside and outside your agency, who are specifically knowledgeable on govtech as well as the government's innovation and data policy agenda?
- Is there a public-sector community of practice or established body of knowledge that can be consulted for ideas on the solution and its potential benefits and risks?
- Have you consulted a repository of previous government AI projects for lessons learned?

**4**

### **Incorporate potentially relevant legislation and codes of practice in your RFP.**

#### **4a. Conduct a review of relevant legislation, rights, administrative rules and other relevant norms that govern the types of data and kinds of applications in scope for the project and reference them in the RFP.**

- Have you consulted legal experts to ensure that the RFP addresses any and all legislation that could be relevant (e.g. with regard to privacy, national security)?
- Have you investigated whether there are commonly accepted industry practices regarding data?
- If applicable, have you established the governing law of data in cases of cross-border data flows?
- Have you set expectations in the RFP that contestability (i.e. the ability for a user to appeal against a decision made by the AI tool) will be built into the tool?

**4b. Take into consideration the appropriate confidentiality, trade-secret protection and data-privacy best practices that may be relevant to the deployment of the AI systems.**

- Have you agreed on what is commercially valuable information with the vendor to ensure that confidentiality and intellectual property protection are preserved?
- Have you consulted the freedom of information policies that would govern the required disclosures of information to the public to ensure accountability?
- Will the transfer and processing of personally identifiable data in relation to the solution be consistent with data protection and domestic privacy laws?

**5**

## **Articulate the technical and administrative feasibility of accessing relevant data.**

**5a. Ensure that you have proper data governance mechanisms in place from the start of the procurement process.**

- How sensitive is the data that could be in scope? For example, could a solution potentially involve personally identifiable information (e.g., licence number, social insurance/security number, financial data, health data, etc.)?
- Are there processes in place to protect and manage data that could be used during the project?
- Are there processes in place to protect and manage data that could be used during the procurement process itself?
- Who will ultimately be accountable for the usage of data during the procurement process and the subsequent project (e.g. the Chief Data Officer, the data set's steward, etc.)?
- Is there an escalation mechanism for any procurement team members who may have a concern about potential data usage?

**5b. Assess whether relevant data will be available for the project.**

- Have you conducted a high level assessment to understand what data would be required to address the problem statement (e.g. necessary data sources or missing data)?
- Is the process to access this data understood, including identifying the data owner?
- Is there an understanding of how data would be accessed by the successful vendor(s) (e.g., onsite without leaving your data environment, remotely through VPN)?

**5c. Define if and how you will share data with the vendor(s) for the procurement initiative and the subsequent project.**

- Is there a case for sharing data with vendors (e.g. the benefits of sharing outweigh the risks)?

- If you have decided to share data, what mechanisms will you put in place to ensure the safety, confidentiality and privacy of the data?
- If you have decided to share data and you will be sharing a sample, how will you ensure the sample is representative of the users that will be affected by a possible solution?

**5d. Ensure that you have the required access to data used and produced by the AI system.**

- Have you asked for access to raw input, processed/combined and enriched data produced by the supplier(s) AI models?
- In case data sharing was not permitted, has the supplier been able to clearly articulate the reason for restricted sharing?
- Have you set out data ownership criteria for the AI system?

**6**

**Highlight the technical and ethical limitations of intended uses of data to avoid issues such as historical data bias.**

**6a. Consider the susceptibility of data that could be in scope and whether usage of the data is fair.**

- Would a solution use personally identifiable data, including but not limited to personal contact information, unique personal identifiers (e.g. licence number, social insurance/security number), financial data and/or health data?
- Would a solution use sensitive government data (e.g. military data)?
- What would be the impact of a data breach that could be in scope for the AI system?
- Does the data that could potentially be used for the project meet criteria for fairness, as specified in the guidelines?

**6b. Highlight known limitations (e.g. quality) of the data in the RFP and require those tendering to describe their strategies on how to address the shortcomings. Have a plan for addressing relevant limitations that you may have missed.**

- Does the team that owns and/or manages the data understand the data generation process?
- Have you consulted the data owner to obtain a high-level assessment of the integrity of the data?
- If data is of poor quality, have you considered alternative data sources, or consulted peers and/or market partners to seek advice on whether the data is usable and how much effort would be required to close the gaps?
- Is the data representative of the population to which the solution would apply or is the data biased? If biased, how will the bias(es) be addressed?

7

## Work with a diverse, multidisciplinary team.

### 7a. Develop ideas and make decisions throughout the procurement process in a multidisciplinary team.

- Do you have a clear understanding of the skills that will be required to conduct the procurement process, including those relevant to policy, procurement, data and AI?
- Have you put together a team that has the skill set needed to effectively acquire and maintain the AI solution?
- How do research and consultations develop an understanding of the impact on diverse stakeholders/stakeholder groups?
- Is your team diverse? Does it promote inclusion in its composition? At a minimum, do you meet domestic laws of anti-discrimination?

### 7b. Require the successful bidder(s) to assemble a team with the right skill set.

- Will you require the successful bidder to include in its team resources with understanding of the affected group(s)?
- Will you require the successful bidder to meaningfully engage with the affected group throughout the design process of the solution?
- Does the RFP evaluation criteria assign a score for team diversity?

8

## Throughout the procurement process focus on mechanisms of algorithmic accountability and transparency norms.

### 8a. Promote a culture of accountability throughout AI-powered solutions.

- Would the solution involve a human in the loop or would it be fully automated?
- Is the solution clearly understood by all stakeholders relevant to the RFP who would ultimately be accountable for the solution and its respective outcomes?
- Has an initial impact assessment for a possible solution been created as part of the procurement process, as well as been approved by the relevant stakeholders?

### 8b. Ensure that AI decision-making is as transparent as possible.

- Has an assessment been performed to gauge the necessary level of human oversight, given the sensitivity of the use case, the population affected by the solution and the data?
- Does the RFP ask the successful bidder(s) to create detailed user journey maps, including defining the level of information about the decision-making that the user would expect throughout the journey?
- Does the RFP ask the successful bidder(s) to provide users with an appeal mechanism when the user does not agree with an AI-driven outcome/determination?

- Does the RFP ask the successful bidder(s) to always inform users that they are interacting with a virtual agent, as opposed to a person?

**8c. Explore mechanisms to enable interpretability of the algorithms internally and externally as a means of establishing accountability and contestability.**

- Does the RFP require successful bidder(s) to provide documentation on the logic behind the algorithm, written in a way that can be understood by users with a limited knowledge of AI systems?
- Does the RFP require successful bidder(s) to provide detailed documentation of the solution and its processes?
- Does the RFP encourage successful bidder(s) to choose the least technically complex solution that will meet the requirements?

**9**

**Implement a process for the continued engagement of the AI provider with the acquiring entity for knowledge transfer and long-term risk assessment.**

**9a. Consider during the procurement process that acquiring a tool that includes AI is not a one-time decision; testing the application over its lifespan is crucial.**

- Has it been established whether the solution will be supported in-house or through a vendor? If through a vendor, will it be through the original vendor or a third party?
- Does the RFP require the successful bidder(s) to define how often the model should be updated to maintain the required performance?
- Does the RFP require the successful bidder(s) to agree to third-party solution audits and to provide the necessary level of access required for maintenance and support?
- Does the RFP ensure the necessary level of access, interoperability and data portability required for maintenance and support?
- Have you defined whether the optimal way to source the solution is through one or multiple contracts (e.g. through consideration of budget, risk management, access to skills)?

**9b. Ask the AI provider to ensure that knowledge transfer and training are part of the engagement.**

- Does the RFP require the successful bidder(s) to define how often and by whom the model should be updated to maintain the required performance?
- Does the RFP require the successful bidder(s) to define how they will team up with the public-sector authority to share insights into the technology and provide knowledge transfer?
- Does the RFP require the successful bidder(s) to provide thorough and holistic documentation about the solution?

**9c. Ask the AI provider for insights into how to manage the appropriate use of the application by non-specialists.**

- Does the RFP require the successful bidder(s) to provide training material and/or documentation sufficient for relevant non-technical staff to be able to effectively operate and govern the solution?
- Have you incorporated access control mechanisms to prevent unauthorized and unintended uses of the solution?

**9d. Make ethical considerations part of your evaluation criteria for proposals.**

- Does the RFP ask bidders to provide their own ethics framework for data and AI?
- Does the RFP require bidders to comply with existing government ethics standards, including those created specifically for AI?
- Does the RFP ask bidders to propose process and/or system metrics that reflect a consideration for ethical standards?
- Does the RFP's scoring assign non-trivial weight to ethics capabilities and experience shown by bidders?

**10**

**Create the conditions for a level and fair playing field among AI solution providers.**

**10a. Contact a variety of AI solution providers in various ways.**

- How could traditional and non-traditional partners, such as start-ups and academia, add value to the project?
- Have you actively sought new ways of market engagement, such as hosting a Q&A session, pre-RFP sessions to discuss the problem, supplier days, hackathons or co-working space presentations?

**10b. Engage vendors early and frequently throughout the process.**

- Have you validated the problem statement and your assumptions (e.g. user needs, applicability of AI) with potential partners?
- Have you defined a single point of contact for bidders who have questions and provided the relevant contact information?

**10c. Ensure interoperability of AI solutions and require open licencing terms to avoid vendor lock-in.**

- Does the RFP set expectations that tools used be open source and that open standards be leveraged as much as possible?
- Is there a clear understanding between vendors and the contracting agency regarding IP ownership of the project's deliverables?
- Does the solution involve technologies that contain patents or other intellectual property and if so is licencing available royalty-free?

C

# AI specification and evaluation tool

# Overview

This tool provides examples of requirements for civil servants to include in a request for proposal (RFP). It also highlights examples of robust AI systems development as well as deployment practices to look out for in the responses or discussions with suppliers. It is intended to be used during the procurement process in conjunction with the AI procurement guidelines as well as the risk

assessment that should allow for a proportionate approach to procurement. The key principle for AI procurement is to clearly describe the problem the contracting authority is aiming to address, focus on outcome-based criteria and not overspecify the AI system, ensuring that the most suitable system is purchased and to innovation is supported.

## Purpose of this tool

This document aims to provide you with an introduction on what to consider when evaluating AI systems during the procurement process. It gives examples of several questions that you can ask while procuring AI systems from suppliers in

categories such as intended use, accuracy of data, fairness and transparency of algorithmic-based decision flows, data security and effectiveness of the systems in meeting intended use.

## How to use this tool

You can consult this document while drafting RFPs and evaluating responses. To use this document effectively please refer to the AI risk assessment tool in the workbook to identify which AI systems and procurement considerations may be more relevant for your project and to assess your requirements.

This document does not aim to provide a recommendation for an exhaustive list of necessary requirements that suppliers need to respond to. It highlights issues that can be considered when setting out specifications in RFPs or evaluation responses in an iterative process. You might already

have robust processes in place for some of the issues mentioned below. These examples should not replace those processes, but rather introduce additional criteria to consider due to the complexity added by the AI system. The table below outlines how to use the document in more detail.

Note that the requirements and criteria in this document are for guidance purposes only. It is essential that you consider the importance of the requirements against your needs and tailor your questions and evaluation accordingly.

## 1

**Purpose:** The supplier understands the problem to be solved and the purpose and goals of the technical AI system

Sample specification	Key considerations to look out for in the answers
1.1 Describe the area of the problem space that is addressed by your AI system.	<ol style="list-style-type: none"> <li>1. Does the supplier articulate the part of your problem that is addressed by the AI system?</li> <li>2. Does the supplier recognize and describe any limitations of the AI system for the problem at hand?</li> <li>3. Is it made clear if the AI system is dependent on those AI elements being added?</li> <li>4. Can the supplier justify why use of AI/ML is the best approach to address the problem?</li> </ol>
1.2 Is your approach built on an existing AI system (Commercial Off the Shelf (COTS)) or will it be custom-made or a mix of the two?	<ol style="list-style-type: none"> <li>5. Does the supplier describe the elements of the AI system and where they originate?</li> </ol>
1.3 Describe what algorithms or techniques you anticipate the AI system to implement.	<ol style="list-style-type: none"> <li>6. Does the supplier explain the techniques applied in the AI system, including use of any algorithms and associated software libraries for the algorithms?</li> <li>7. Can the supplier explain how the system operates in an easy to understand way for various audiences?</li> </ol>
1.4 Describe the approach to ensuring that use of AI is necessary and proportionate in the AI system.	<ol style="list-style-type: none"> <li>8. Does the supplier explain the metrics and evaluation methods used and how they have impacted the selection of data that will be used in the proposed AI system?</li> <li>9. Can the supplier articulate potential risks of using the AI/ML solution and risk mitigation strategies?</li> </ol>
1.5 Describe how you have ensured that the AI system is proportional to the data available.	<ol style="list-style-type: none"> <li>10. Does the supplier explain how it will be ensured that data needs required to produce the intended outcome are considered proportional?</li> <li>11. Is the supplier capable of mitigating the data supply that they need from the operator?</li> <li>12. Does the supplier explain the need to access various data sets?</li> </ol>
1.6 Explain how all end users have been considered throughout the design and implementation process.	<ol style="list-style-type: none"> <li>13. Does the supplier describe how the proposed AI system supports transparency and explainability characteristics not just for the data subject, but the end user/operator as well?</li> <li>14. Does the supplier set out a plan that allows for user testing and an iterative design approach and risk mitigation?</li> </ol>
1.7 Explain how you will demonstrate accountability for the goals and outcomes of the AI system.	<ol style="list-style-type: none"> <li>1. Does the supplier describe the end user training they commit to deliver to ensure the ongoing health and maintenance of the AI system and outcomes?</li> <li>2. Is the supplier providing documentation detailing how the AI system can be configured or adapted if the results are not delivering the goals or the AI is not acting in an ethical or understandable manner?</li> </ol>

**Consent and control:** The developer will ensure that they have consent from the data subject before processing data or training an algorithm, and that human operators can control the outcome

**Sample specification**

2.1 Please provide evidence that you have considered the legal and ethical implications and gathered consent for processing and capturing the data throughout the full lifecycle of the AI system.

NOTE: criteria correspond to COTS AI system. Same criteria can, however, apply to tailored products (e.g. “The supplier provides information on what individuals will be told, when they will be made aware, what kind of consent will be needed from them, and what the procedures will be for gathering consent.”).

**Key considerations to look out for in the answers**

17. Can the supplier articulate how it was decided whose data to use or about whom to make inferences?
18. Is it clear that data subjects know that their data is being used or that inferences are being made about them?
19. Does the supplier provide information on what individuals were told, when they were made aware, what kind of consent was needed from them, and what the procedures were for gathering consent?
20. Does the supplier highlight potential risks to these individuals or groups and how the service output might interfere with individual rights?
21. In the case of risk identification, does the supplier describe how the risks are being handled or minimized?
22. Does the supplier describe how the rights of individuals who provided the data were safeguarded throughout the process?
23. Is it made clear whether individuals have the option to withdraw their data and opt out from inferences being made about them? If yes, what is the withdrawal procedure?

Suppliers should ensure that all raw input, processed, training and enriched data is accessible and usable in a timely manner for the public-sector authority, especially for monitoring and inspection. Ideally the suppliers process and data governance should make sure that persistent ownership and access to this data is granted to the public-sector authority, including third party and/or open source data sets.

2.2 Describe your approach for allowing access and control of the data within the AI system.

24. Does the supplier provide access to the AI model(s) input data, including any third party or open source data including mechanisms for controlling the flow of data?
25. Can the supplier provide access to all the AI-model(s) training data and when this is not feasible explain the process for providing a representative sample?
26. Can the supplier provide full access to the AI model(s) processed/combined and enriched data (i.e. key features, inferred scores/metrics) and when this is not feasible explain the process for providing a representative sample?
27. Does the supplier describe the level of contractual ownership that will be granted to the above data and for what period?

2.3 Describe the level of human decision-making at critical control points.

28. Does the supplier describe the approach to active monitoring to track user behaviour to identify irregular patterns that may indicate signs of unintended consequences?
29. Does the supplier mention operational bias reviews to track model inputs and outputs to identify irregularities that may indicate bias?
30. Does the supplier mention that they might retrain the model in agreement with the operator using new or more up-to-date data to account for changes in user behaviour?

**3****Privacy and cybersecurity:** The supplier will not introduce harm through unintended consequences or poor practice

Sample specification	Key considerations to look out for in the answers
<p>3.1 Describe your privacy and cybersecurity approach for the proposed AI system as well as how the data will be protected.</p> <p>NOTE: COTS and bespoke AI systems will have dependency on security controls managed by the authority.</p>	<p>31. Does the supplier deploy well-established techniques, security processes and standards to protect the data, for example, encryption and anonymization, where appropriate and feasible?</p> <p>32. Does the supplier describe how need-to-know principles for data access are applied and the decision criteria for allowing access to data and AI models?</p> <p>For legitimate and logical reasons, protected and or sensitive data may be required and processed by the AI system. Development teams should invest time in understanding the reasons why the data is sensitive and the impact on the data subjects in the event of a biased decision or data breach. Typically, AI systems must not be designed to be fully autonomous. Human operators or even data subjects should be able to intervene or interrupt in the event of incorrect or harmful decisions being made and/or be asked to confirm a processing phase or learning step before it commences.</p>
<p>3.2 Describe the potential threats to the system or AI system from external or internal adversaries.</p> <p>NOTE: Bespoke AI systems may have dependencies on authority risks, but should be able to describe risks that are specific to the AI system.</p>	<p>33. Does the supplier define how the system could be attacked or abused?</p> <p>34. Suppliers could:</p> <ul style="list-style-type: none"><li>- List applications or scenarios for which the service is unsuitable.</li><li>- Describe specific concerns and sensitive use cases and what procedures can be put in place to ensure that the service will not be used for these applications, or if the service needs to be used in a sensitive use case the precautions being taken to mitigate harm.</li><li>- Underline that they will verify AI model stability when exposed to sub-system compromise and/or outages.</li><li>- Describe how they are securing user or usage data.</li><li>- Identify if usage data from service operations is retained and stored.</li><li>- Ascertain how the data is being stored and for how long the data is stored.</li><li>- Mention how they will verify if enriched and/or inferred user or usage data is being shared outside the service and who has access to the data.</li><li>- Describe how the service checked for robustness against adversarial attacks, including once it is integrated/deployed at scale.</li><li>- Explain how robustness policies will be checked and the type of attacks considered.</li><li>- Propose a plan to handle any potential security breaches based on accepted industry best practice.</li></ul>
<p>3.3 Explain your test processes, including the specialist expertise used to assess the AI system.</p>	<p>35. Does the supplier provide evidence that the AI system has been tested and that AI domain experts were involved in the development, testing and deployment?</p> <p>36. Can the supplier describe how the AI model(s) will be monitored and checked to highlight potential malicious manipulation (internal and external)?</p>
<p>3.4 Please provide evidence of previous case studies of where the AI system has been implemented and how the output has been interpreted, highlighting best practice.</p>	<p>37. Does the supplier provide evidence of where the AI system has been used before?</p> <p>38. Can the supplier point to previous use cases that include description of how the output has been consumed, drawing out if any harm or negative impact on the end users or data subjects was introduced through misuse or misinterpretation?</p>

## Ethical considerations: Will the service or AI system be fair in its decision making and processing

Sample specification	Key considerations to look out for in the answers
4.1 What data limitations have you identified and what strategies will you implement to address these data limitations?  NOTE: this is applicable only when the authority has shared data with the supplier or when the supplier is using pre-trained models or their own data. Otherwise, this should be assessed during AI system design.	39. Can the supplier describe where they have missing or poor quality data? Are they able to identify potential risks that arise from missing or poor data and can they articulate how they are mitigating these risks?  Suppliers should be able to describe how data bias policies will be checked (with respect to known protected attributes), bias checking methods and results (e.g. disparate error rates throughout different groups).  Suppliers should also be aware of the personal or unconscious bias inherent in the development team and the human operators of the AI system and how it influences the output of the system. Bias may also be a legitimate input in certain problem sets or use cases, but unconscious or personal bias that undermine the correctness of the outcome or introduces harm must be avoided. There needs to be a focus on detecting unconscious or personal bias during the training and testing of the algorithm.  Given the needs to adapt processes to ensure fair treatment for persons with disabilities as employees and as service users and citizens accessing government information and services – suppliers must be required to demonstrate that the end-to-end process they are influencing or managing is non-discriminatory – it is important, but far from sufficient, to just address data bias.
4.2 How will you ensure that the AI system fits the requirements of data ethics frameworks and policies prior to going live?	40. Is the supplier able to demonstrate how data ethics principles referred to in the RFP are considered in designing, building and supporting their AI system?
4.3 Describe the approach to eliminate (or minimize) bias, ethical issues or other safety risks as a result of using the service.	41. Can the supplier describe the possible sources of bias or unfairness assessed and where they arise from – the data, the techniques being implemented or other sources?  42. Is there any mechanism for redress if individuals are negatively affected?
4.4 Describe the process for ensuring that the development team adopts an ethical mindset.	43. Does the supplier offer training or have an awareness process to ensure their team understands the potential impact of creating an AI system that produces an incorrect, biased or disproportional output?  44. Can the supplier describe how they educate their staff to understand and accept that individuals have unconscious bias and understand their responsibility for ensuring this does not affect the operation of the AI system?
5.5 Explain how the AI system will be tested during the life cycle to detect bias and the remediation steps if it is introduced.	45. Can the supplier describe bias policies models and bias checking procedures, as well as how they will monitor and verify results (e.g. disparate error rates throughout different groups) with a focus on controls for unacceptable bias and/or defined thresholds?  46. Does the supplier highlight life cycle considerations and maintenance of the AI system? Do these considerations include model validation processes to assess performance against defined tolerances and/or thresholds and demonstrate their ability to highlight other potentially less visible problems (i.e. overfitting)?

## 5

**Explainability:** Can the supplier adequately explain how the AI system functions to the affected consumer, data subject or operator

**Sample specification**

5.1 Describe the provisions in the AI system to ensure that the outputs are explainable and/or interpretable.

5.2 Would you allow independent, third party audit(s) of the AI system? If your answer is no, please explain.

5.3 Describe how you enable end-to-end auditability of the AI system.

**Key considerations to look out for in the answers**

- 47. Is the supplier able to define how their organization approaches ethics?
- 48. Is the supplier able to show how they aim to aid the explainability of their AI system (e.g. directly explainable algorithm, local explainability, explanations via examples)?
- 49. Can the supplier provide clear guidance and explanations on how the results of the AI process should be interpreted?
- 50. Does the supplier outline the target user of the explanations (AI expert, domain expert, general consumer etc.) and ask them to describe any human validation of the explainability of the algorithms?
- 51. Does the supplier highlight key parameters and inputs to their AI model(s) and how they affect the outputs (i.e. sensitivities)?
- 52. Is the supplier able to allow for external audits?
- 53. In the case that an external audit is not possible, justification must be provided.
- 54. Can the supplier describe what information is captured throughout the AI system and provide a taxonomy to describe the meaning of the information?
- 55. Is the supplier able to provide documentation related to the development and support of the AI system, for example, test reports, logs and quality criteria?

Sample specification	Key considerations to look out for in the answers
6.6 Explain how you will ensure the AI system or service does not drift from its intended purpose or outcome.	<p>56. As algorithms are learning continuously after they are developed it is possible for them to drift from the original concept and deliver different results. Providers can be assessed on their approach to the following:</p> <ul style="list-style-type: none"><li>– What is the expected performance on unseen data or data with different distributions?</li><li>– Does the system make updates to its behaviour based on newly ingested data?</li><li>– Is the new data uploaded by users? Is it generated by an automated process? Are the patterns in the data largely static or do they change over time?</li><li>– Are there any performance guarantees/bounds?</li><li>– Does the service have an automatic feedback/retraining loop or is there a human in the loop?</li><li>– How is the service tested and monitored for model or performance drift over time?</li><li>– Is the supplier providing performance drift monitoring KPIs that prompt retraining if there are any unexpected changes?</li><li>– How can the service be checked for correct, expected output when new data is added?</li><li>– Does the service allow for checking for differences between training and usage data?</li><li>– Does it deploy mechanisms to alert the user of the difference?</li><li>– Do you test the service periodically?</li><li>– Does the testing include bias or fairness related aspects?</li><li>– How has the value of the tested metrics evolved over time?</li></ul>

### Sample specification

7.1 Explain how your system or service conforms to specific international or local open interoperability standards or other relevant standards relating to cyber security, coding quality, safety, testing, accessibility and usability.

Examples are the IEEE standards as well as GDPR for personal identifiable information (PII).

### Key considerations to look out for in the answers

57. Does the supplier explain how the AI elements of the system or service operate with the following?
  - Required data storage/access requirements?
  - Operational monitoring/compliance tools?
  - Standard system elements, including COTS, Operation support systems (OSS) and/or custom?
58. Can the supplier demonstrate the range, velocity and veracity of data and features that can/will be provided for wider potential use/developments?
  - Detail interfaces (i.e. API) and integration dependencies (particularly OSS or custom elements)?
  - Provide an approach for future interoperability requirements?
59. Does the supplier include business continuity management measures such as documentation and access to key processes and algorithmic steps for the AI model(s), where these are not provided as part of the normal delivery of the AI system?

Sample specification	Key considerations to look out for in the answers
8.1 Describe the architecture of the AI system, including use of external COTS or open source elements and the function they provide in the AI system. This should consider the data used by each element of the AI system and how the output of that element was validated.	<p>60. If an AI system is based on an existing algorithm or will integrate with another functionality, the supplier should be able to describe the full nature of the system. For example, a COTS AI system could introduce unknown ethical risks if used improperly. Potential areas for consideration could be:</p> <ul style="list-style-type: none"><li>– Is the service or AI system based on COTS, OSS and/or legacy AI system(s)?</li><li>– Which datasets was the service trained on?</li><li>– Were there any quality assurance processes employed while the data was collected or before use?</li><li>– Were the datasets used for training built for purpose or were they repurposed/adapted?</li><li>– Were the datasets created specifically for the purpose of training the models offered by this service?</li><li>– Are the training datasets publicly available?</li><li>– For each dataset: Does the dataset have a datasheet or data statement?</li><li>– Did the service require any transformation of the data in addition to those provided in the datasheet?</li><li>– Was synthetic data used and how was this generated?</li><li>– How were the models trained and when were they last evaluated for correctness?</li><li>– How often are the models retrained or updated?</li><li>– Did you use any prior knowledge or reweight the data in any way before training?</li><li>– How is testing conducted by the service provider?</li><li>– Which datasets was the service tested on (e.g. links to datasets that were used for testing, along with corresponding datasheets)?</li><li>– Could these datasets be used for independent testing of the service? Did the data need to be changed or sampled before use?</li><li>– Please provide details on train, test and holdout data and what performance metrics were used (e.g. accuracy, error rates, AUC, precision/recall)?.</li></ul>

Sample specification	Key considerations to look out for in the answers
9.1 Explain how you will ensure the AI system or service does not drift from its intended purpose or outcome.	<p>61. Is the supplier able to provide information on any existing training courses or documentation they have available?</p> <p>62. Does the supplier include the creation of training materials as part of their offering bespoke AI systems?</p>
9.2 Explain how you will ensure usability for non-trained staff.	<p>63. Can the supplier describe the target user for the AI system, including expectations around their skills?</p> <p>64. Can the supplier articulate how users can be trained to use and understand the AI/ML solution being implemented?</p> <p>65. Can the supplier outline the types of skills required to support or use the AI system and the role types they would expect to see? For example, system admin, data scientist, end user.</p>
9.3 Explain how the AI system will be maintained, how its accuracy and integrity will be sustained over time, and whether third party providers could be engaged for these activities.	<p>66. Is the supplier able to describe the handover process in the case of a bespoke or COTS offering? This should detail:</p> <ul style="list-style-type: none"> <li>– Accuracy metrics and thresholds to ensure the integrity of the AI system.</li> <li>– Maintenance processes and activities.</li> <li>– Support contracts.</li> <li>– Suitability for third party support.</li> </ul> <p>67. Is the supplier able to provide a service agreement detailing the approach to AI in case the system is based on software as a service (SaaS)?</p> <p>68. Can the supplier demonstrate scale deployment considerations for their AI model(s) (e.g. limit to data coverage, minimum model training requirements, system processing time sensitivities, etc.)?</p>

**Sample specification**

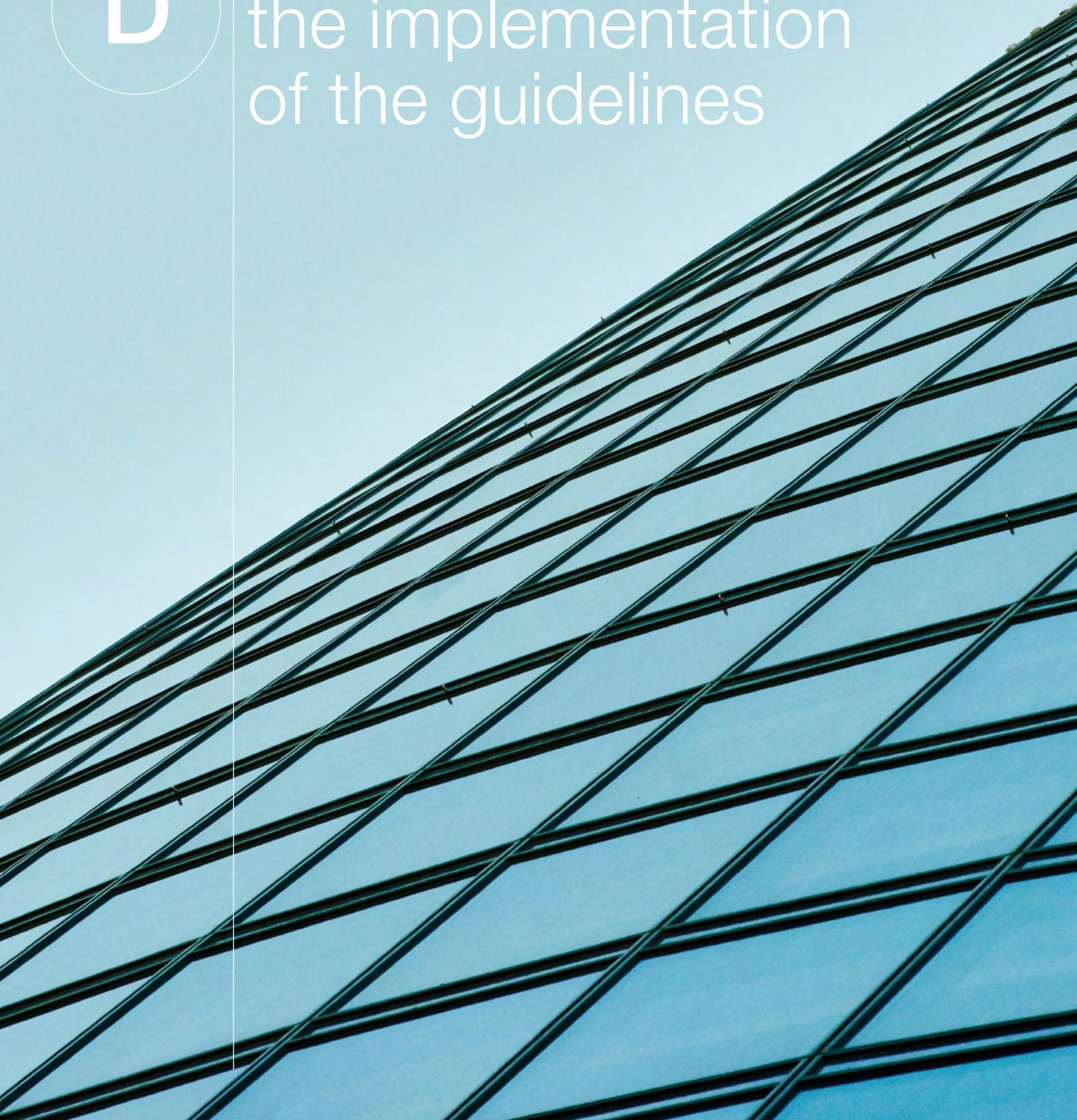
10.1 Can you demonstrate how you will assess the competencies, qualifications and diversity of the team that will develop and deploy the AI system?

**Key considerations to look out for in the answers**

69. Can the supplier outline how they are drawing on appropriate skills to be domain experts in the field of AI and in the area the AI system is to be applied?
70. Do the supplier skills set match standards referenced in the Skills Framework for the Information Age (SFIA framework)?<sup>6</sup>
71. Does the supplier highlight the importance of diversity in AI development and explain how this is considered in the composition of the delivery team and provide strategies to increase diversity in AI development if diversity requirements cannot be met by the immediate team?

D

# How to kick-start the implementation of the guidelines



## **World Economic Forum AI procurement workshop templates**

↓ Day One

↓ Day Two

↓ Day Three



# E

# Case studies



1

# Case study India

Controller General of Patents,  
Designs and Trade Marks



**“The aim is to enhance efficiency, uniformity and consistency.**

## Objective

The Indian Controller General of Patents, Designs and Trade Marks ([CGPDTM](#)) wanted to make use of artificial intelligence (AI), blockchain, internet of things (IoT) and other new technologies for its

patent processing system. The aim is to enhance efficiency, uniformity and consistency within issues ranging from inception of a possible IP to its enforcement.

## Why AI?

The patent processing system is a manually extensive and long process. As such, AI was considered a potential solution to modernize, automatize and strengthen the transparency of the process. It is also hoped that having a stable and efficient IP regime in the country encourages

innovation to achieve the country's industrial and economic development goals. The initiative was part of a larger government effort to explore the use of blockchain and AI in diverse areas such as education, healthcare, agriculture, electricity distribution and land records.

## Background

The CGPDTM is responsible for administration of all major IPR legislations in the country regarding patents, designs, trademarks, geographical

indications, copyrights and semiconductor integrated circuits layout-design. The office processes approximately 55,000 applications per year.

## Action

The procurement process was divided into two phases – the initial expression of interest (EOI) and request for proposal (RFP). The EOI was made available publicly on an existing e-tendering platform well-known to the business sector. The aim was to seek proposals as to how best to shortlist vendors for the purpose of hosting a limited tender. The participation of small and medium-sized enterprises was greatly encouraged through lower eligibility standards.

The agency suggested different areas for the proposals (electronic data processing, screening, prior art searching, pre-grant opposition etc.) and companies were invited to pitch various solutions and technologies. The selection criteria for the EOI was based on track-record for similar projects, general qualifications of key staff, financial strength and accreditation and certifications. Hence, the

agency ensured that the vendor had the right skills set to develop and deploy the AI solution by demanding proof of certifications, references and past experiences.

The RFP evaluation was much more focused on a specific type of solution and was based on technical bid evaluation, technical demonstration and financial bid. For the financial bid, the lowest bid was considered successful. Throughout the process, vendors were invited to submit queries for specific questions, which were answered at specific moments and made publicly available. It was agreed that the solution developed and furnished belongs exclusively to CGPDTM. The vendor had to grant a non-exclusive licence to access, replicate and use the application software, the custom software and any proposer owned software embedded in the systems.

## Ethical considerations

An important consideration for the deployment of the solution was the explicability of the search queries and the avoidance of biases. This was ensured by making the source code of the solution available to the public. The RFP also made clear

that any sensitive data provided would be hosted either on premises or through an API access<sup>8</sup> and would only be available to the successful vendor for testing/development phase. Furthermore, it was clarified that no data would be hosted outside India.

## Lessons learned: Which guidelines were harder to implement?

---

**“Support an iterative approach to product development.”**

**“Assess whether relevant data will be available for the project.”**

**“Develop an understanding of the skills that are needed to effectively acquire and maintain an AI-powered solution, before starting the procurement process.”**

The “Eligibility and Financial Criteria” methodology used to select a vendor was hard to understand for many RFP participants. One aspect that led to confusion was the required accuracy of 75% for developed models. The RFP did not give a clear definition of “accuracy” and did not provide historical data for training and testing of the models. As machine ML/AI models improve accuracy over time as they learn and get better, it was hard for the RFP participants to develop a 75% accuracy without access to relevant data. In addition, this evaluation criteria lacked transparency and didn’t support an iterative approach to product development. Following the concerns raised by the participants, the CGPDTM lifted that requirement.

---

## Success factors: Which guidelines were successfully implemented?

---

**“Aim to include your procurement within a strategy for AI adoption across government and learn from others.”**

**“Reach out in various ways to a wide variety of AI solution providers.”**

**“Create the conditions for a level and fair playing field among AI solution providers.”**

**“Focus on developing a clear problem statement, rather than on detailing specifications of a solution.”**

**“Define if and how you will share data with the vendor(s) for the procurement initiative and the subsequent project.”**

**“Require the successful bidder(s) to assemble a team with the right skills set.”**

Successfully designing and deploying AI in an organization as big and complex as the CGPDTM was a major technical and human challenge. Assembling a team with experience in change management and technical expertise on integration with existing software and datasets could have helped to better navigate the procurement and implementation process.

This project was part of a larger government of India-wide effort to adopt and enhance the use of latest technologies and as such, senior government functionaries were very active in making the procurement process a success. This strong leadership from the government ensured that the right resources were employed and the process moved forward.

---

While providing opportunities to various firms to compete, the public EOI also boosted innovation and the diversity of the proposed solutions. Newly established providers were also given the opportunity to compete for this public-sector contract through lower requirement standards.

An extensive and clear description of the IPO workflow and use-cases for AI made it easy for participants to identify opportunities. Documenting user needs and challenges for each stage of patent applications was crucial for AI system providers to understand the problem.

---

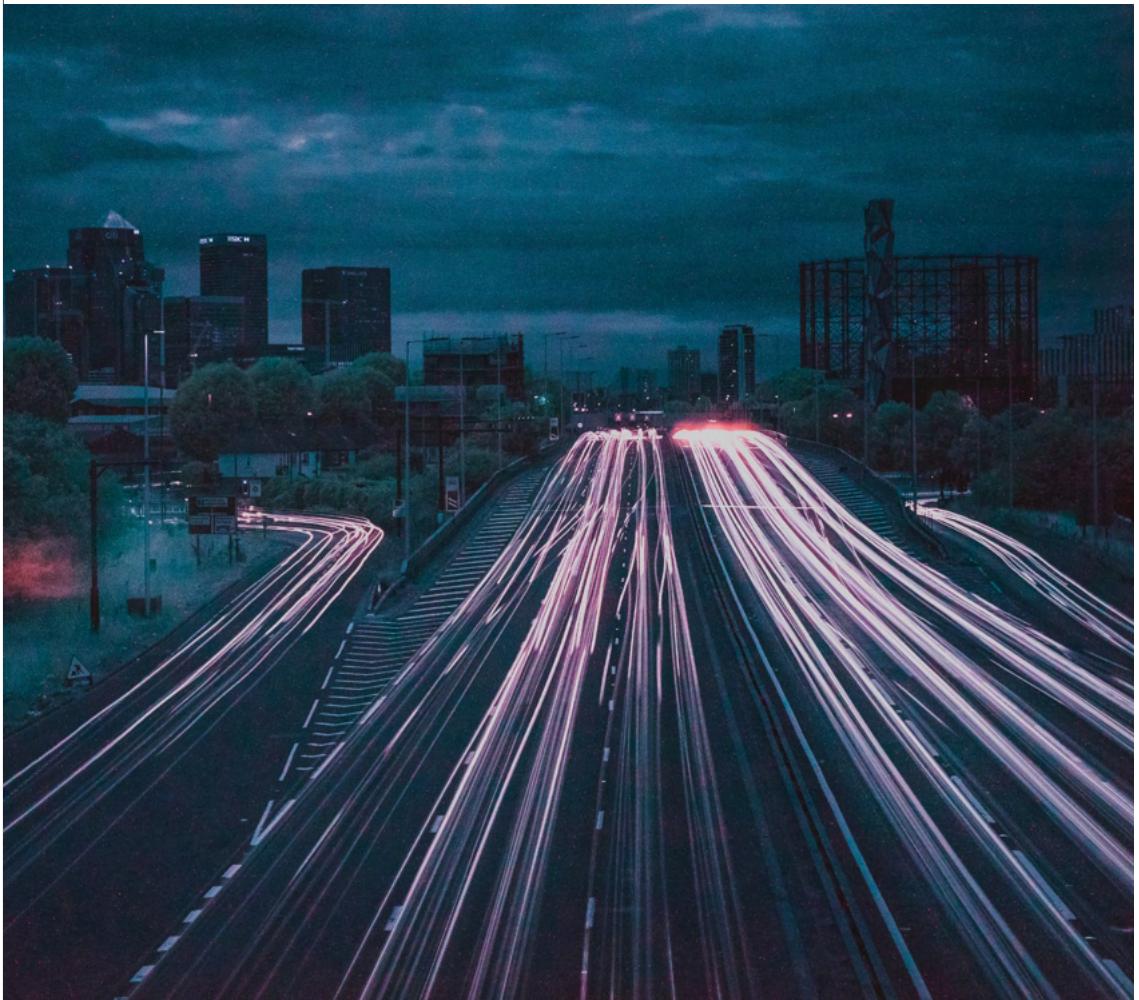
The RFP was clear on data governance during and after the procurement initiative. The governance approach specified who would be granted data access, the purposes for which a vendor would be authorized to use the data and the minimum requirements for hosting/reviewing the data.

Evidences of skills and qualifications of key team members were required in the initial EOI. Evidence of bidder’s resources for deploying the solution were also assessed and were part of the decision-making criteria.

2

# Case study United Kingdom

Driver & Vehicle Standards Agency



 **The department only became aware of the power and opportunities of applying AI when it received the responses to the invitation to tender – and, at a more detailed level – once it started working with the partners.**

## Objective

The Driver and Vehicle Standards Agency (DVSA) wanted to make use of digital technologies to ensure that vehicle standards are enforced while at the same time saving time and costs.

A data-driven approach should help the agency to conduct intelligent inspections of authorized garages conducting the vehicle standards test.

## Why AI?

The team held a lot of data that it couldn't use effectively. The testing was resource intensive and the previous process did not allow for targeted inspections. Clustering techniques offered insights

that were previously not available. This helped to make predictions that now support a more targeted approach to inspections.

## Background

The DVSA is an executive agency of the United Kingdom Department for Transport, which among other things supervises the MOT scheme, a vehicle standards examination, ensuring that authorized garages carry out tests to the correct standards. This examination, referred to as “the MOT”, assesses vehicle safety, roadworthiness and exhaust emissions and is required in the UK for most vehicles over three years old and used on anything that can be classified as a road. Each year, 66,000 testers conduct 40 million MOT tests in 23,000 garages. The inspection of the authorized garages was resource intensive and the knowledge was limited to effectively target inspections of these garages.

The DVSA made the decision to further invest in the MOT to improve the service in a number of

ways, including quality of the service to the end user (motorist), test quality, reduce fraud risks and improve efficiency. The DVSA had insufficient capacity to do this so chose to procure two digital partners. As well as delivering some of the improvements (in consort with DVSA as part of blended agile teams) the partners would also develop the department's in-house skills.

The DVSA released an invitation to tender (ITT). The AI aspects of the work were part of this larger contract for digital transformation and the department only became aware of the power and opportunities of applying AI when it received the responses to the invitation to tender – and, at a more detailed level – once it started working with the partners (as part of options for solving business challenges).

## Action

During the procurement process the DVSA ensured that the ITT set out clearly what challenges it wanted to solve and what outcomes it sought. The DVSA used the Digital Outcomes and Specialist Framework, which is a framework agreement that focusses on the digital transformation of public sector services.<sup>9</sup> The ITT did not ask for AI as a technology, but laid focus on the use of technologies that would deliver the most effective outcome. The aim of the procurement effort was to contract digital services and skills that would help the team to identify and deploy the right tools and systems to address the delivery challenges, in particular improving the DVSA inspection of authorized garages that conduct MOT tests. During the ITT stage, pricing arrangements were kept simple with partner effort paid on a time and

materials basis at agreed rates. It was required that all IP would be owned by the DVSA.

The project started with a set of mini discoveries, which enabled the agile nature of the work. These covered a number of areas and included the following:

- Improving MOT test quality through better supporting testers
- Better enabling the DVSA to know which garages presented the greatest risks of testing poorly
- Identifying those applying to be involved in MOT that may present risks to the integrity of the MOT service

**50%**  
the fall in examiners'  
preparation time for  
enforcement visits

In collaboration with the supplier, the DVSA applied a clustering model against garage test data from a three-month period.<sup>10</sup> The clustering model grouped MOT-authorized garages based on the behaviour they show when conducting MOT tests, such as the test duration, time of test and result of inspection (against expected). The DVSA created a risk (of testing incorrectly) score for each garage, which allowed the department to rank garages and their testers and helped it identify regional trends. The model was validated against those who had been identified as doing things incorrectly, ensuring that the model could learn what behaviours were good indicators of wrong-doing.

An important consideration was the ability to explain the model and the human in the loop. It is important to explain the outcome of the risk rating without losing the integrity of the test. Having a human

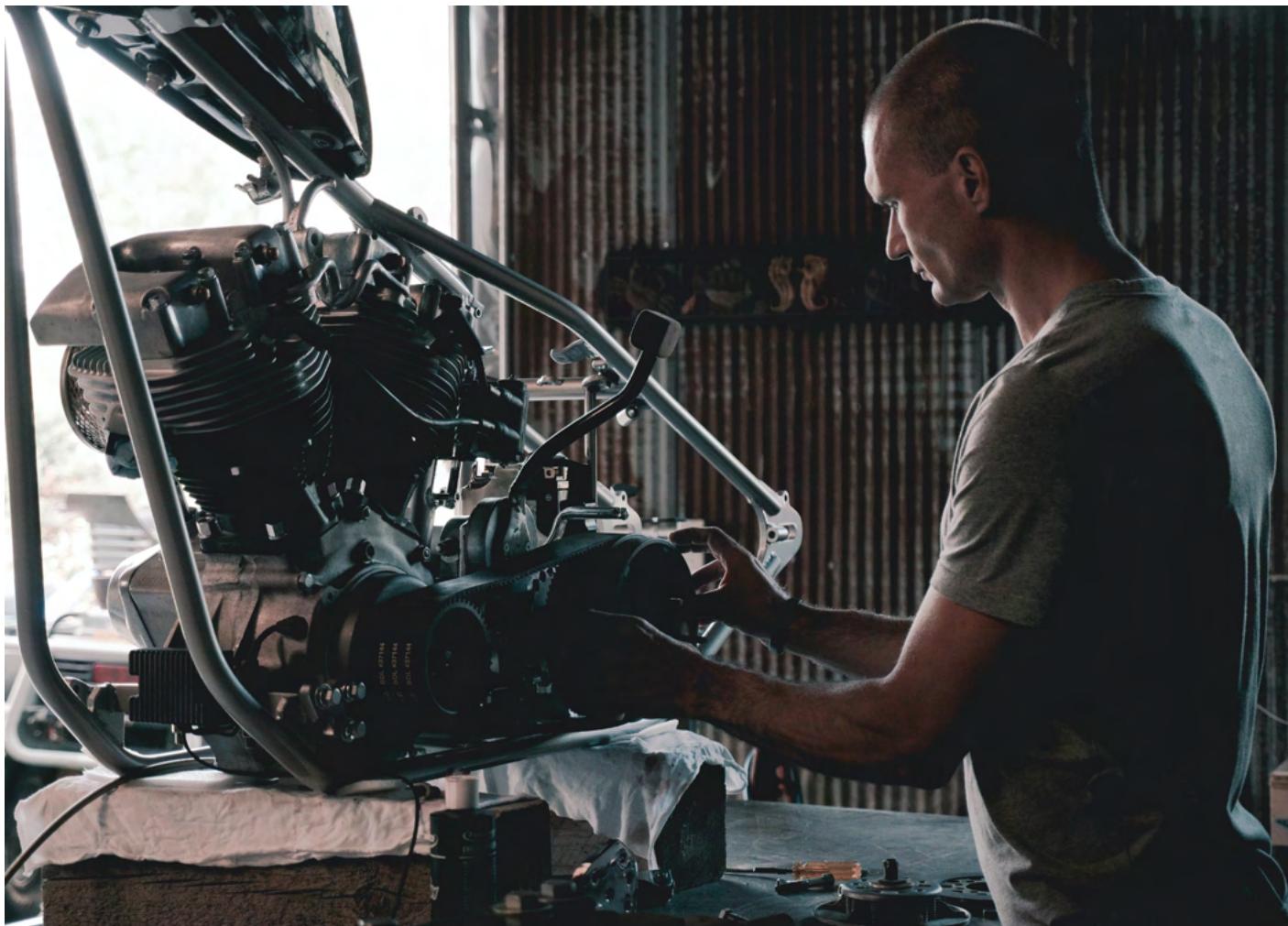
in the loop who interrogates and decides to take action on the risk score was crucial to make the use of AI successful. All the data used for the AI system was data that was already collected by the DVSA and it did not include a great amount of sensitive data. Suppliers had visibility of some data, but not off-site access.

The lifecycle management of the tool was not fully factored in upfront and became a challenge once the technology was developed. The DVSA team identified this as an issue and worked with suppliers to put together a plan to bolster the skills of the department's continuous improvement team. This ensures that the system continues to work effectively and meets users' needs, as well as technical support that addresses issues related to hosting and live service failures.

## Impact

The DVSA can now target its resources at the garages and testers with the highest risk score. By identifying areas of concern in advance, the examiners' preparation time for enforcement visits has fallen by 50%.

There has also been an increase in disciplinary action against garages, meaning standards are now being better enforced. As more garages are delivering better MOT standards, there are more cars on the road that comply with roadworthiness and environmental requirements.



## Lessons learned: Which guidelines were harder to implement?

---

**“Support an iterative approach to product development.”**

It was important to find the right balance between agile delivery and the focus on price in the evaluation of the proposals. Since prices and timelines might shift due to the agile nature of the work, you must ensure that you reflect this in the scoring of the invitation to tender and not only focus on the fixed lowest price of the delivery.

---

**“Consider during the procurement process that acquiring a tool that includes AI is not a one-time decision; testing the application over its lifespan is crucial.”**

Considering the life-cycle management and its impact on procurement revealed to be a challenge. The earlier the focus on the maintenance of the solution and the ongoing management of the AI system, the better it is for the project delivery.

## Success factors: Which guidelines were successfully implemented?

---

**“Make use of innovative procurement processes to acquire AI systems - encourage collaboration between different bidders.”**

It was important to rely on a team of suppliers for project delivery, rather than just one supplier. Partnering with three suppliers and asking them to deliver the project in collaboration ensures that all relevant skills were available and checks and balances were in place. Regarding AI delivery, one supplier developed the AI model and another supplier helped to test the model and ensured that it worked properly.

---

**“Focus on developing a clear problem statement, rather than on detailing specifications of a solution.”**

The requirements in the ITT focused on outcomes rather than the means of how to achieve those outcomes. This gave vendors the flexibility to select the technology that they found fit for purpose and ensure that the solution was innovative and effective.

---

**“Work with a diverse, multidisciplinary team.”**

The agency worked actively on upskilling internal teams and recruiting experts into the team where needed. This helped the agency to become a better customer for AI systems.

The delivery was supported through a close collaboration with the suppliers. During the project delivery the DVSA worked closely with delivery partners. Key to this was thinking as a single team and as partners, not contractors. At a practical level, this meant being open about the problems that needed to be solved, the challenges that different solutions may present and the costs of different options. This experience showed that openness brings real reward in getting value from the partnerships.

---

**“Engage vendors early and frequently throughout the process.”**

Extensive pre-market engagement helped to better target potential AI system providers. The DVSA hosted a supplier open day to explain the challenges that the agency faces to suppliers and gather initial ideas of how and with the help of which technologies to address these. After the initial tendering process, shortlisted suppliers were asked to present their approaches to the DVSA, which improved the ability to evaluate the different delivery approaches.

3

# Case study United Arab Emirates

Dubai Electricity and Water Authority



 **The ability of machine learning to leverage a range of enterprise information and improve its interactions combined with the chatbot's ease of interaction proved to be an ideal means to meet the data access needs.**

## Objective

To enable an efficient and comprehensive procurement process for digital and AI solutions, DEWA's top management had directed their team to demonstrate leadership on this topic. By identifying use-cases where the new procurement guidelines could be applied, DEWA's aim was to work on a pilot which could be then scaled across UAE and globally.

One of the use cases identified was the need for senior management at DEWA, to access reports and dashboards on a daily, weekly, monthly and quarterly

basis to review strategic performance indicators. These dashboards and reports are available on different platforms and some of them take a long time to generate and prepare before they can be presented to top management. As a result, DEWA was looking for a faster and easier way to access the required data to make correct and timely decisions. A technology was needed that was capable of understanding management's enquiries, providing the right data in a convenient and timely way and learning from the enquiries made.

## Why AI?

The use of AI to solve complex challenges was supported by the state's National AI Strategy, which seeks to position the UAE as an AI world leader by 2031. DEWA also has a vision to become a globally leading sustainable innovative cooperation, and its strategic objective is: "Enabling AI and digital technologies". To achieve these goals, DEWA defined three main pillars for its AI adoption. The first is Rammas for You, which covers customer-facing services. The second is Rammas at Work, which seeks to augment the work environment with AI tools, and the third is Powered by Rammas, which adds AI to DEWA's core business assets.

In January 2017, DEWA launched the Rammas Virtual Agent, a chatbot that answers customers' enquiries and is powered by AI, as part of the Rammas for You pillar. Following the virtual agent's success, DEWA began considering using the same concept to meet management's data access needs. The ability of machine learning to leverage a range of enterprise information and improve its interactions combined with the chatbot's ease of interaction proved to be an ideal means to meet the data access needs.

## Background

Dubai Electricity and Water Authority (DEWA) is a public utility founded on 1 January 1992, by a decree issued by the late Sheikh Maktoum bin Rashid Al Maktoum to merge Dubai Electricity Company and Dubai Water Department. DEWA's strategies and achievements are inspired and driven by the vision and directives of His Highness Sheikh Mohammed bin Rashid Al Maktoum, Vice President and Prime Minister of the UAE and Ruler of Dubai. Today, DEWA provides services to over 900,000 customers across Dubai.

DEWA was the 1st government organisation in the UAE to launch an online AI chatbot in 2017. The chatbot called Rammas communicates in both Arabic and English with customers and respond to their enquiries. AI helps DEWA's customers with

services, such as the Smart Response service on DEWA's smart app and website. This allows early self-diagnosis of technical interruptions at home, reducing the necessary steps to deal with complaints and follow-ups.

DEWA conceptualised the AI procurement guidelines with the World Economic Forum and Dubai Future Foundation to further drive cooperation between the public and private sectors, and to enable governments and companies to make their procurement processes as efficient and transparent as possible by employing a multi-stakeholder approach. DEWA implemented a framework that allowed for feedback and finding best practices and standards to govern AI technologies procurement process.

## Action

DEWA sent a request for proposal (RFP) to suppliers. Bidders had a month to respond, after which there was a window for bidders' questions and a bidder's conference to answer further questions.

The final evaluation of the solution proposals used seven criteria with different weights. Technical assessment and AI capability were the most important, and the proposed solutions were evaluated with a demonstration or evaluation of

a prototype from each bidder's solution. DEWA also evaluated project governance, deliverables, business value, solution dependency and vendor background, and awarded the contract to the highest scoring proposal evaluated by the procurement committee, which comprises important stakeholders and AI specialists.

After this, the source code for the solution was shared with DEWA. This is an open source system

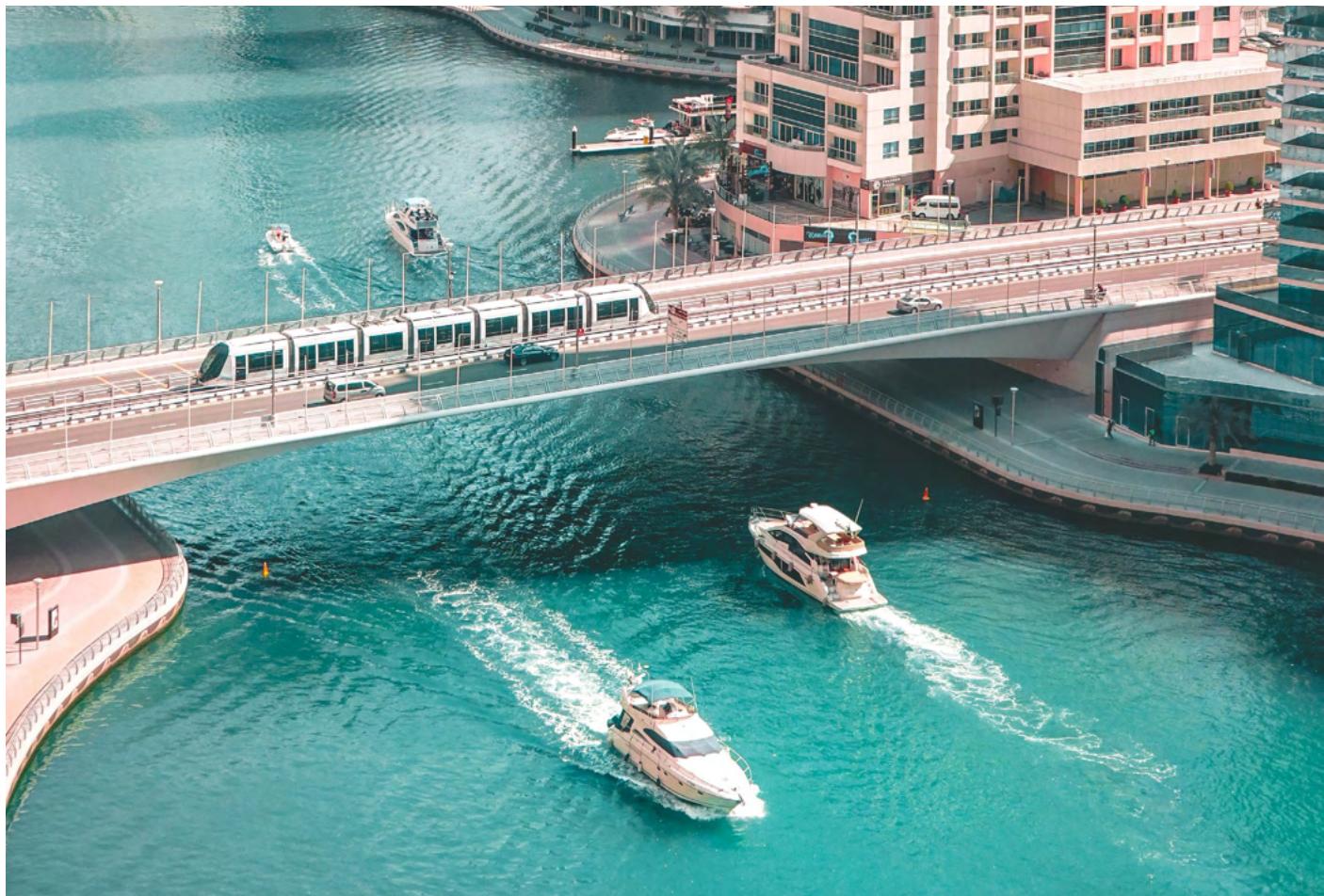
and will be developed from scratch and hosted by MORO, a digital platform launched in 2018 to support the Dubai 10X initiative. MORO provides hosting and data storage services and cloud-based digital services management. The supplier contract took into account additional requirements, such as training DEWA employees to maintain and improve it to ensure continuity and the proper communication of knowledge, to enable DEWA to further expand its capabilities.

## Ethical considerations

DEWA is committed to protecting customers' and stakeholder's data by adopting and complying with relevant UAE legislations and Dubai Government applicable regulations. This includes Federal Law No.1 for 2006 on Electronic Commerce & Transactions; Federal Legal Decree No. 5 for 2012 on combating cyber-crime, and the Regulatory Framework for stored values & Electronic Payment Systems (EPS Regulation), which regulates business offering electronic payment services.

DEWA also adheres to the Dubai Data Law, open data, shared data, data confidentiality and data sensitivity policies. DEWA also put in place internal measures to secure customer data. It drafted a contract that clearly stated the requirements to prevent sharing its information with any external parties; and that such data must always reside within DEWA's assets.

The solution works in tandem with multiple internal datasets related to strategic KPIs, employees' statistics, organisational data and sensitive information. The solution had to run on a private cloud within the UAE in adherence with the Dubai Data Law due to this sensitivity because it cannot be shared or processed externally. It was important that sensitive datasets remain protected at all times. To address this, the roles and responsibilities of each user were applied within the solution, and controlled by pre-defined access levels. There was considered to be no issue with data transparency or the ability to understand the AI model as the AI solution is only accessing data without any modification.



## Lessons learned: Which guidelines were harder to implement?

---

**“Make use of innovative procurement processes to acquire AI systems.”**

The procurement process took five months from the business case initiation until the announcement of the preferred bidder. The team considered this would take too long. As a result, DEWA developed a new procurement track specific to AI in cooperation with the World Economic Forum and Dubai Future Foundation. This track was benchmarked by Dubai Future Foundation to apply similar techniques to specifically expedite the adoption of AI tools within DEWA.

The new procurement track consists of a set of key milestones including:

- Establish a senior AI Committee which includes champions from multiple departments and specialities to guarantee a 360-degree approach when evaluating AI RFPs and aligning them with AI Procurement Guidelines to ensure the adoption of the Framework, define an AI pre-approved supplier list, thus, enhance the overall AI procurement process and accelerate the adoption of AI technologies in DEWA.
- Create the DEWA AI Definition to have a clear description for AI-use cases within DEWA, avoid confusion with other technologies, and facilitate the overall process.
- Create AI RFP templates. Early market engagement will also be a key component of this new track, as the procurement team will constantly be on the search for new AI vendors via conferences and info sessions.

## Success factors: Which guidelines were successfully implemented?

---

**“Focus on developing a clear problem statement, rather than detailing specifications of a solution.”**

DEWA implemented the first pilot for a virtual agent called Rammas, in 2016 and then launched the first version of the live solution in January 2017. Nine months later, the UAE AI Strategy was announced with a clear vision ‘to be an AI World Leader by 2031.’

The scope of the project was clear from the start as it was part of the AI roadmap initially. This made the process particularly efficient by leading to more relevant vendors’ responses and an increased probability of success.

**“Aim to include your procurement within a strategy for AI adoption across government and learn from others.”**

DEWA embedded AI in its strategy and developed a separate AI functional strategy that has been aligned and cascaded from the UAE National AI strategy. The functional AI strategy covers 6 main pillars, including AI in stakeholder happiness, AI in technology, AI in sustainability, AI in operations, AI investment, and enabling AI.

DEWA also responded immediately to the National AI Strategy by building a five-year roadmap to augment its work with AI tools. Moreover, DEWA is an active member of the Smart Dubai AI Advisory board and works closely with the Smart Dubai Office and other government entities for knowledge sharing and delivering new innovative services powered by AI to Dubai’s population.

For instance, the Rammas Virtual Agent content is integrated with Smart Dubai Office’s Virtual Agent, called Rashid, which is available on the Dubai Now smart application to ensure information availability and to maintain a seamless customer experience.

**“Work with a diverse, multidisciplinary team.”**

DEWA organised, in collaboration with Dubai Future Foundation, and World Economic Forum Fourth Industrial Revolution Centre, a four-day workshop in October 2019 about Artificial Intelligence (AI) Procurement guidelines.

This was part of DEWA's efforts to position the UAE as a global leader in AI by 2031 in line with the UAE Strategy for Artificial Intelligence.

One of the main outcomes of this workshop was to form a senior AI committee within DEWA, which includes champions from the Contract and Procurement department, an AI Team, an Intellectual Property Team, the BRM Team, and the PMO Team. This committee is responsible for evaluating the AI RFPs and to align them with AI Procurement Guidelines, to ensure the adoption of the Framework, by defining an AI pre-approved supplier list, improving the overall AI procurement process, and accelerating the adoption of AI technologies in DEWA.

This ensured a comprehensive evaluation of the proposed solutions and a good understanding of the issues at play.

**“Define if and how you will share data with the vendor(s) for the procurement initiative and the subsequent project.”**

DEWA adheres to the Dubai Data Law, open data, shared data, data confidentiality and data sensitivity policies. Moreover, DEWA has internal measures to control data privacy. Customers' data is not shared with any external parties and the data always resides within DEWA's Assets.

**“Ensure that you have proper data governance mechanisms in place from the start of the procurement process.”**

DEWA's security team is making sure that the data provided to the vendors is secured, encrypted and in compliance with Data Residency Law of UAE and DESC (Data Electronic Security Centre).

The Personal Identifier Information (PII) data was removed from the vendors' dataset and the rest was encrypted. This gave the vendors access to the structure of the data, which is all that was needed to build a prototype.



4

# Case study Kingdom of Bahrain

Information and eGovernment Authority



**“After discussion with different solution providers and evaluation of the first Proofs of Concept, it became clear that AI could add value to the proposed solution by using it for predictive analytics.**

## Objective

Decisions about advanced studies and career pathways in Bahrain have been traditionally based on strong cultural and social imperatives to pursue pure academic qualifications for traditional white-collar jobs, irrespective of whether there is labor market demand from those sectors. This social norm is compounded by the fact that there is no authoritative source of labor market intelligence on which prospective employees can base their study and career decisions. Together, these factors give rise to ill-informed decision-making, which has a detrimental impact on students, employers, and the government.

Therefore, the Labour Fund (Tamkeen) in collaboration with the Information and eGovernment Authority (iGA), and other government institutions, decided to develop an Employability Skills Portal (ESP) to serve as a repository of labor market information. This portal could be used by prospective employees to make informed career decisions and by educational institutions to tailor their programs to market demand. The portal needed a technology capable of cleaning and integrating data from multiple sources, finding correlation between the data and making prediction on the direction of various trends and indicators.

## Why AI?

The use of AI was not a requirement at the start of the project. However, after discussion with different solution providers and evaluation of the first Proofs of Concept, it became clear that AI could add value to the proposed solution by using it for predictive

analytics. In addition, the use of AI was in line with the vision of higher management and the Kingdom of Bahrain's leaders to support digital transformation and the use of modern technology.

## Background

The Information and eGovernment Authority (iGA) of Bahrain facilitates many public services related to the IT sector. It aims to achieve cyber security integration between the public sectors institutions,

as well as to work on implementing the knowledge in order to support decision making, creativity and encouraging innovation in the areas of public services and institutions.

## Action

As the portal would be based on the cloud, the project floated through an existing special procurement track for cloud technologies. This track accelerates the implementation of cloud projects by bypassing traditional tendering processes. In order to do that, this innovative procurement track offers access to dedicated funds for cloud technologies and a list of pre-approved vendors selected for their internal knowledge, links with global technology leaders and financial capabilities. The process started with a first, free of charge, Proof-of-Concept (POC), from different solution providers. These POCs were evaluated through an agile methodology until they reached an acceptable level of satisfaction by end users, the labor market, and internal users and iGA technical team. Each POC was then given a score based on both users' evaluation and a financial bid. Most weight was given to the ability to reach expected end results and user needs. The highest

scoring vendor solution was chosen to move to the next phase; the development of a complete POC with costs covered by iGA. If the required level of satisfaction from the final POC was not met, iGA would select the next highest scoring vendor solution to move to the second phase until the required level of satisfaction was reached and the contract was awarded. This iterating phase took about two months to complete.

The solution was agreed to be fully owned by iGA and its internal technical team was involved from the start in the implementation process to ensure a proper handover of the solution. iGA technical and management team also made sure to benefit from the bidders' knowledge through weekly meetings and close collaboration to better understand the implication and use of AI.

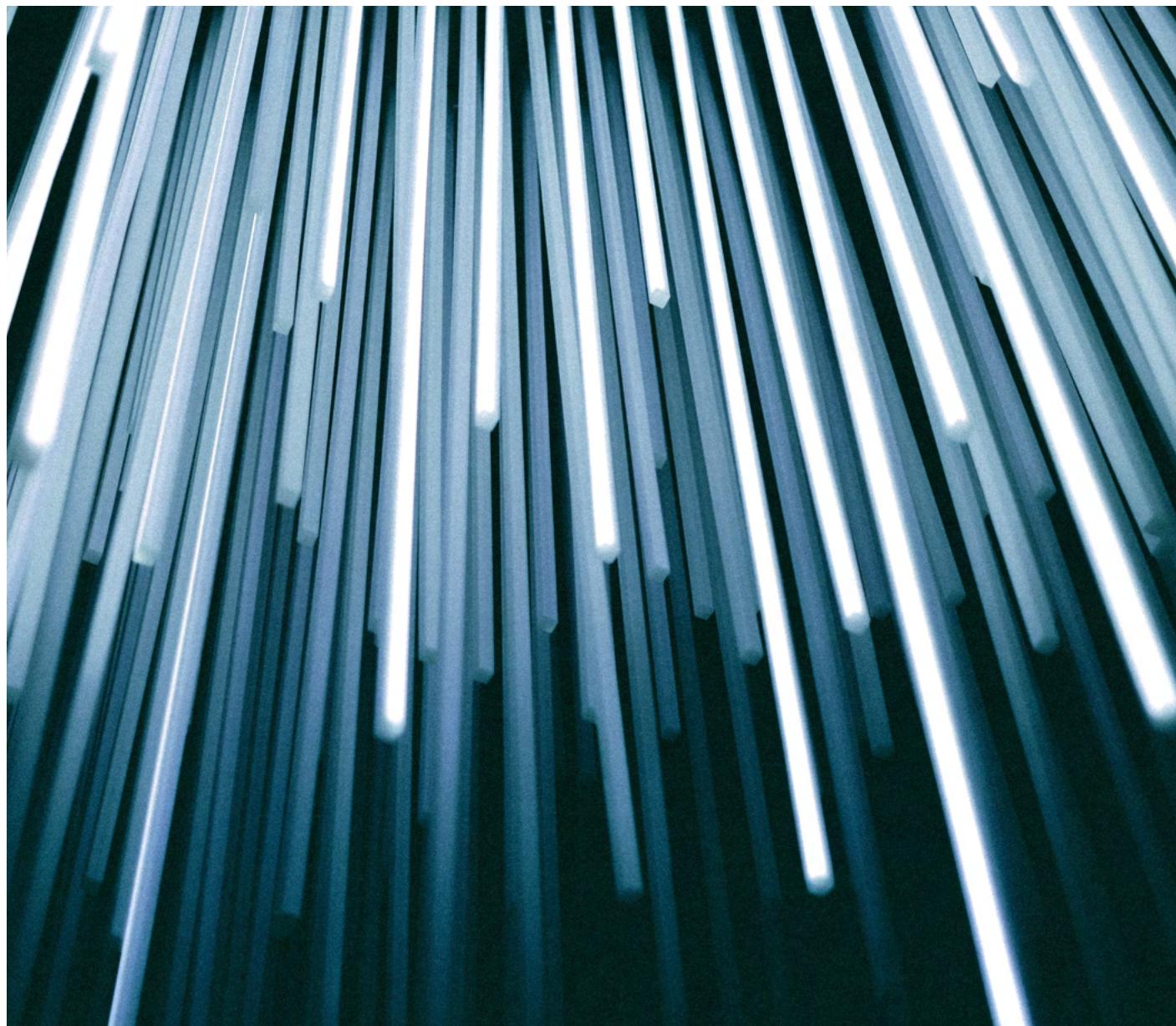
## Ethical considerations

The Data Protection Law of Bahrain, which regulates the use of personal data, was taken into consideration for the project and vendors had to comply with it. However, the project also involved other types of data not covered by the regulations. The use of various datasets from different government entities was an important issue because of the lack of regulations and governance for data sharing between organizations and the lack of governance for non-personal data. Hence, a task force leaded by iGA and top management from each involved organization was created. The role of this task force was, in part, to serve as a governance body for data sharing and also to gain an in depth understanding of each dataset and the biases that could emerge when using AI. Indeed, the best way to gain insights on the nature of each dataset and their potential bias was to partner with the providers of these datasets. iGA also appointed an external

legal consultant to conduct an impact assessment of the use of data before starting the project. The goal was to understand if the way each dataset would be used could create legal or ethical issues.

Concerning data sharing with the vendors for the POCs development, the vendors had access to the entire population to train their models, but synthetic data was used to mask personal information. The synthetic data was generated in such a way that the real aggregate results were preserved. In addition, the vendors could only access the data through temporary iGA internal accounts.

The AI model explainability was addressed by requiring the successful vendor to provide a non-technical description of the model that would be available to internal users.



## **Lessons learned: Which guidelines were harder to implement?**

---

**“Create the conditions for a level and fair playing field among AI solution providers.”**

The accelerated cloud technology procurement track being a new process, the list of pre-approved vendors was not fully developed at the time of the project. Work is being done to expand this list and give access to new innovative vendors.

**“Make use of innovative procurement processes to acquire AI systems.”**

The introduction of payment for the development of the second POC was a new concept that slowed the process as it was hard to get approvals. Moving forward, instead of requiring approval for each new payment, the accelerated cloud technology procurement track will include lump-sum funds that can be allocated as needed for each procurement project.

## **Success factors: Which guidelines were successfully implemented?**

---

**“Focus on developing a clear problem statement, rather than on detailing specifications of a solution.”**

The project didn't start with AI in mind. The need for a specific outcome was defined and the technical evaluation of the vendors' solution was focused on their capacity to meet the desired outcome. Hence, the project was open to a variety of technical solutions and was able to select the most appropriate technology.

**“Conduct an initial AI risk and impact assessment even before starting the procurement process, ensure that your interim findings inform the RFP, and revisit the assessment at decision points.”**

An external consultant was mandated to evaluate the potential impacts of the use of AI on the different datasets. Potential biases were identified as well as the mitigation strategies.

**“Conduct a review of relevant legislation, rights, administrative rules and other relevant norms that govern the types of data and kinds of applications in scope for the project.”**

Relevant regulations were identified and communicated to the vendors. In addition, blind spots within the current regulations were identified and strategies were put in place to address them. A government task force was formed to identify best practices and establish consensus on the use, processing and transfer of non-regulated data.

**“Ensure that you have proper data-governance mechanisms in place from the start of the procurement process.”**

A government task force comprised of top management from each organization where data would be collected was created. Hence, the vendors and iGA team were able to meet with the data providers and truly understand potential biases and limitation to the quality of each datasets in order to avoid misleading results. Vendors were then able to adapt their model accordingly and address these shortcomings.

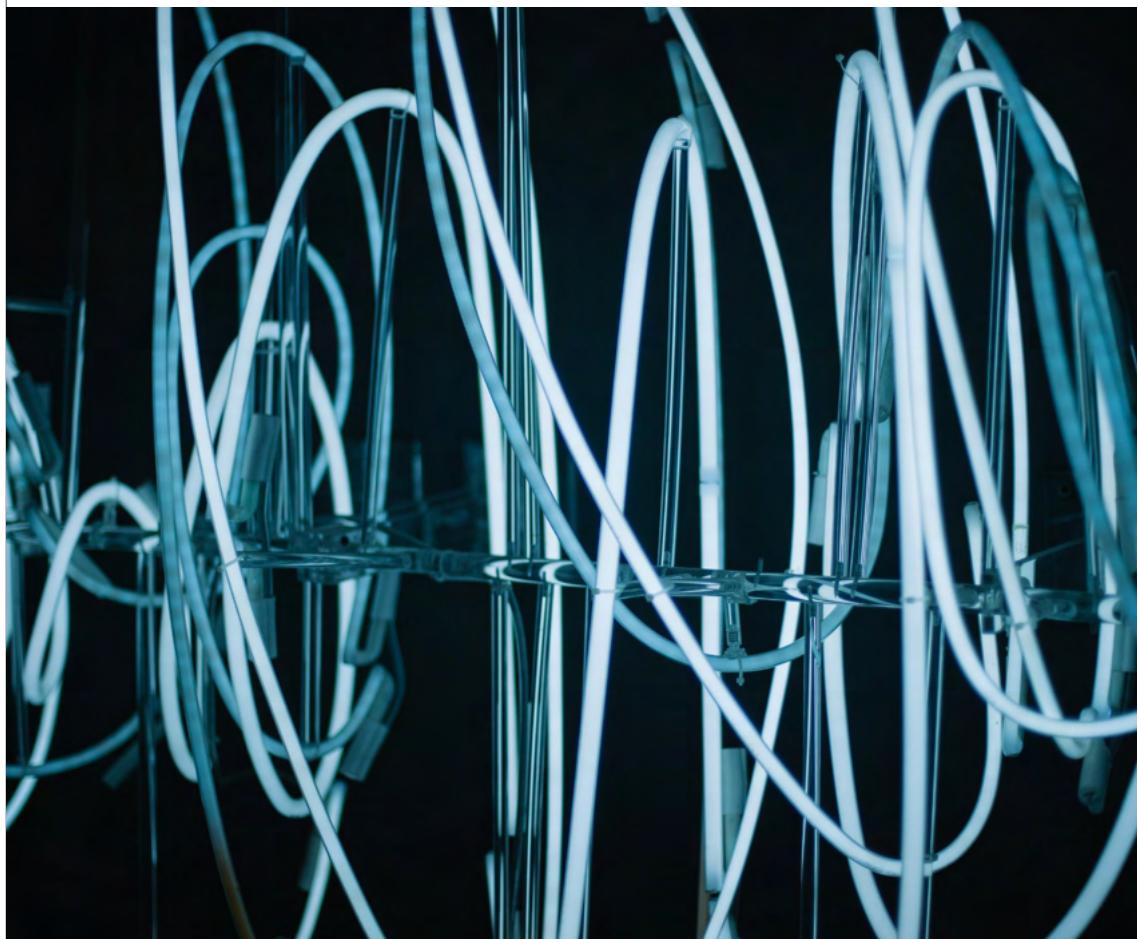
**“Ask the AI provider for knowledge transfer and training to be part of the engagement.”**

iGA internal technical team was involved from the start in the implementation process to ensure a proper handover of the solution. iGA technical and management team also made sure to benefit from the AI providers' knowledge through weekly meetings and close collaboration to better understand the implication and use of AI.

5

## Case study Splunk Inc.

Key considerations for successful adoption of AI as an added capability/functionality with an existing supplier and a system already in use



There are multiple ways of procuring and adopting AI technologies; they can be built from scratch, added as capabilities to commercial off-the-self (COTS) systems or acquired directly as a service (SaaS). Often solutions require a mix of these approaches to be successfully adopted. For most operational organizations AI capabilities are added iteratively to an existing solution or procured via an existing supplier as an added functionality to a product or service. When adopting AI as part of an

existing platform contract without going through an independent AI procurement process, some guidelines are more relevant than others.

Three important factors, highlighted in the guidelines, form the basis of success for public-sector agencies adding AI-capabilities to systems already in use. These have emerged from Splunk's experience supporting and working collaboratively with public-sector entities:

## Key guideline for AI as an added capability/functionality

**“Define the public benefit of using AI while assessing risks.”**

**“Articulate the technical and administrative feasibility of accessing relevant data.”**

**“Highlight the technical and ethical limitations of intended uses of data to avoid issues such as historical data bias.”**

## Key factor to consider to successfully implement the guidelines

### End users' background

When considering the benefits that can be realized with an AI system, understanding the end-user audience is of great importance. The end-user's understanding of pertinent mathematical principles (such as probability) and how they are likely to interpret and apply the output of the AI system should be considered. This will help inform the type and granularity of outputs (e.g. visual charts, key metrics etc.) that should be selected, how fast new techniques can be adopted and/or accepted and what cautions, if any, are desirable for the particular use case.

### Understanding data assets

Finding and understanding what data an organization holds and how it may be accessed, combined and processed in accordance with the law and organizational norms will help you determine project scope – what can be achieved with the data and with what controls. According to recent research, 97% of public-sector agencies agree that they must improve their ability to ingest, index and cross-correlate disparate data sets to optimize public policy outcomes.

### Data literacy

AI technologies can be complex and therefore, to be successful in the identification of technical and ethical limitations, it is critical that an organization's leadership and operations team be “data literate”. This does not mean each team member must become a data scientist, but they should understand the underlying mathematical principles (i.e. probability, accuracy, sampling etc.) and gain an appreciation of the different benefits and limitations of the main ML techniques. Innovation and education go hand in hand. Without a proper data and knowledge foundation, users will not be able to capitalize on the advances in automation and decision-making capability provided by AI.

# Acknowledgements

The World Economic Forum's Unlocking Public Sector Artificial Intelligence project, in collaboration with the Government of the United Kingdom, Deloitte Consulting and Splunk is a global, multistakeholder and cross-disciplinary initiative intended to help shape the public sector's adoption of AI, and emerging technologies in general, around the world. The project has engaged leaders from

private companies, governments, civil society organizations and academia to understand public-sector procurement of AI technology, identify challenges and define principles to guide responsible and ethical procurement. The opinions expressed herein may not correspond with the opinions of all members and organizations involved in the project.

## Lead authors:

### **Sabine Gerdon**

Artificial Intelligence and Machine Learning Fellow, World Economic Forum, Seconded from the Office for Artificial Intelligence, Government of the United Kingdom

### **Eddan Katz**

Project Lead, World Economic Forum

### **Emilie LeGrand**

McGill University Integrated Management Student Fellow

### **Gordon Morrison**

Director of EMEA Government Affairs, Splunk Inc.

### **Julián Torres Santeli**

Artificial Intelligence and Machine Learning Fellow, World Economic Forum, Seconded from Deloitte Canada's AI practice

We would like to thank our Unlocking Public-Sector AI project community as well as the following contributors for their insights:

### **Rashid Alahmedi**

Senior Specialist Technology and Solutions, Dubai Electricity and Water Authority

### **Greg Ainslie-Malik**

Machine Learning Architect, Splunk Inc.

### **Jesus Alvarez-Pinera**

Head of Data, Food Standards Agency

### **Shelby Austin**

Managing Partner, Growth and Investments and Omnia AI, Deloitte

### **Yousef Al-Barkawie**

Partner, Analytics and Cognitive Middle East Leader, Deloitte

### **Neil Barlow**

Head of Vehicle Policy and Engineering, Driver and Vehicle Standards Agency

### **Kathy Baxter**

Architect, Ethical AI Practice, Salesforce

### **Lorena Cano**

Digital Trade Fellow, World Economic Forum from Inter-American Development Bank

### **Ashley Casovan**

Executive Director, AI Global

### **Michael Costigan**

Artificial Intelligence and Machine Learning Fellow, World Economic Forum from Salesforce

### **Sue Daley**

Associate Director, techUK

### **Nihar Dalmia**

Government and Public Sector AI leader for Deloitte Canada, Deloitte

### **Gourav Dhiman**

Business Development Manager, XLPAT

### **Cosmina Dorobantu**

Deputy Director of Public Policy Programme, The Alan Turing Institute

### **Leslie Harper**

Senior Sector Specialist, Inter-American Development Bank

### **James Hodge**

Chief Technical Adviser, Splunk Inc.

### **Hamad Karam**

Senior Specialist Artificial Intelligence, Dubai Electricity and Water Authority

### **Andrew Kim**

Head of AI Policy, Google Cloud

<b>Steven Knight</b> AI Lead, Food Standards Agency	<b>Nada Al-Saeed</b> Data Policy Fellow, World Economic Forum from Bahrain Economic Development Board
<b>Benjamin Leich</b> Economic Adviser, Better Regulation Executive	<b>Komal Sharma Talwar</b> Director, XLPAT and TT Consultants
<b>Katherine Mayes</b> Programme Manager, techUK	<b>Leonard Stein</b> Senior Strategic Adviser, Splunk Inc.
<b>Maha Mofeed</b> Chief Corporate Officer, Bahrain Economic Development Board	<b>Jitin Talwar</b> Founder, XLPAT and TT Consultants
<b>Valesca Molinari</b> Automotive and Autonomous Mobility Fellow, World Economic Forum from Baker McKenzie	<b>Sandeep Singh Kohli</b> Co-founder, XLPAT
<b>Mariam Al Muhairi</b> Head, Centre for the Fourth Industrial Revolution United Arab Emirates	<b>Ahmad Al Tawallbeh</b> Specialist Artificial Intelligence, Dubai Electricity and Water Authority
<b>Khalid Al Mutawa</b> Director, Bahrain Information and eGovernment Authority	<b>Abbey Thornhill</b> Assistant Economist, Better Regulation Executive
<b>Brandie Nonnbecke</b> Founding Director, CITRIS Policy Lab	<b>Adrian Weller</b> Programme Director for AI, The Alan Turing Institute
<b>Arwa Al Qassim</b> AI Lead, Centre for the Fourth Industrial Revolution United Arab Emirates	<b>Mark Woods</b> Director, Technology and Innovation, Splunk Inc.
<b>Ana Rollan</b> Artificial Intelligence and Machine Learning Fellow, World Economic Forum from BBVA	<b>Tim Woodbury</b> Director of State and Local Government Affairs, Splunk Inc.

Thank you also to the teams in the UK from the Defence Science and Technology Laboratory, the Department for Transport, the Home Office Accelerated Capability Environment and local governments that supported the user testing and piloting. The steering and working group from the Department of Digital, Culture, Media and Sport, the Government Digital Service, the Cabinet Office, the Crown Commercial Service and the Centre for Data Ethics and Innovation has been instrumental to progressing this work, in particular:

<b>Sue Bateman</b> Deputy Director for Policy and Innovation, Government Digital Service	<b>Stephen Hennigan</b> Deputy Head of Office for Artificial Intelligence, United Kingdom Government
<b>Oliver Buckley</b> Executive Director, Centre for Data Ethics and Innovation	<b>Sana Khareghani</b> Head of Office for Artificial Intelligence, United Kingdom Government

Thank you to everyone who contributed through interviews, workshops and discussions in the last 18 months in Dalian, Dubai, London, Manama, San Francisco, Tianjin, Toronto and Washington DC.

# Endnotes

1. <https://www.gartner.com/en/information-technology/glossary/open-source>
2. <https://searchdatacenter.techtarget.com/definition/COTS-MOTS-GOTS-and-NOTS>
3. <https://www.gartner.com/en/information-technology/glossary/infrastructure-as-a-service-iaas>
4. <https://www.gartner.com/en/information-technology/glossary/platform-as-a-service-paas>
5. <https://www.gartner.com/en/information-technology/glossary/software-as-a-service-saas>
6. <https://www.sfia-online.org/en>
7. Factsheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie House, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, Kush R. Varshney.  
<https://arxiv.org/abs/1808.07261>
8. API stands for application programming interface. An API is a software intermediary that allows two applications to talk to each other.
9. <https://www.gov.uk/guidance/digital-outcomes-and-specialists-buyers-guide>
10. Unsupervised learning was used given the team did not have labelled data.



---

COMMITTED TO  
IMPROVING THE STATE  
OF THE WORLD

---

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

---

World Economic Forum  
91–93 route de la Capite  
CH-1223 Cologny/Geneva  
Switzerland

Tel.: +41 (0) 22 869 1212  
Fax: +41 (0) 22 786 2744  
[contact@weforum.org](mailto:contact@weforum.org)  
[www.weforum.org](http://www.weforum.org)



COMMITTED TO  
IMPROVING THE STATE  
OF THE WORLD

White Paper

# Guidelines for AI Procurement

September 2019



World Economic Forum  
91-93 route de la Capte  
CH-1223 Cologny/Geneva  
Switzerland  
Tel.: +41 (0)22 869 1212  
Fax: +41 (0)22 786 2744  
Email: [contact@weforum.org](mailto:contact@weforum.org)  
[www.weforum.org](http://www.weforum.org)

© 2019 World Economic Forum. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, including photocopying and recording, or by any information storage and retrieval system.

This white paper has been published by the World Economic Forum as a contribution to a project, insight area or interaction. The findings, interpretations and conclusions expressed herein are a result of a collaborative process facilitated and endorsed by the World Economic Forum, but whose results do not necessarily represent the views of the World Economic Forum, nor the entirety of its Members, Partners or other stakeholders.

# Contents

What is artificial intelligence (AI)?	4
Why do we need guidelines for public procurement of AI?	4
How were these guidelines developed?	5
How to use the guidelines	5
Guidelines overview	6
Detailed explanation of guidelines	7
Endnotes	17

# What is artificial intelligence (AI)?

AI has been formally defined as “technologies [that] aim to reproduce or surpass abilities (in computational systems) that would require ‘intelligence’ if humans were to perform them. These include: learning and adaptation; sensory understanding and interaction; reasoning and planning; optimization of procedures and parameters; autonomy; and creativity”.<sup>1</sup>

New AI approaches developed in the past decade, particularly the use of deep-learning neural networks, have dramatically advanced the capability of AI to recognize complex patterns, optimize for specific outcomes and make automated decisions. Doing this requires massive amounts of relevant data, a strong algorithm, a narrow domain and a concrete goal, and can result in dramatic improvements in reliability, efficiency and productivity.

## Why do we need guidelines for public procurement of AI?

Governments are increasingly seeking to capture the opportunities offered by AI to improve public-sector productivity and the provision of services to the public, and to stimulate the economy. AI holds the potential to vastly improve government operations and meet the needs of citizens in new ways, ranging from traffic management to healthcare delivery to processing tax forms. However, governments often lack experience in acquiring modern AI solutions and many public institutions are cautious about harnessing this powerful technology. Guidelines for public procurement can help in a number of ways.

First, government and the general public have justified concerns over bias, privacy, accountability, transparency and overall complexity. New incidents are emerging of negative consequences arising from the use of AI in areas such as criminal sentencing, law enforcement and even employment opportunities. As citizens increasingly demand the same level of service from their governments as they do from innovative private-sector companies, public officials will be required not only to identify the specific benefits AI can bring, but also to understand the negative outcomes that can be generated.

Governments do not have the latitude of using the inscrutable “black box” algorithms that increasingly characterize AI deployed by industry. Without clear guidance on how to ensure accountability, transparency and explainability, governments may fail in their responsibility to meet public expectations of both expert and democratic oversight of algorithmic decision-making and may inadvertently create new risks or harms.

Governments rely on the expertise, and previously developed models, of technology providers and may lack the necessary skills to fully understand or trace algorithmic causality. Technology providers understand these challenges and look to governments to create clarity and predictability about how to manage them, starting in the procurement process. While companies are generally wary of stricter guidelines for government procurement, common-sense frameworks can help governments overcome reluctance to procure

complex new technologies and actually open new markets for companies. Transparent guidelines will permit both established companies and new entrants to the AI space to compete on a level playing field for government contracts.

Second, AI procurement can build on a foundation of previous efforts to improve the effectiveness and efficiency of government technology procurement, which may include legislation or policy measures such as frameworks, model contracts, etc. Established principles of good government technology procurement may take on added significance in AI procurement. For example, many governments already ensure that procurement efforts are run by multidisciplinary teams. Experience has shown that a lack of diversity in AI teams and positions of leadership has correlated with inadvertent harms or discrimination to vulnerable minority groups and protected classes. Given government’s role in upholding inclusion, an added emphasis on a multidisciplinary approach and diversity may be necessary in AI procurement.

Third, as noted, AI has advanced rapidly in recent years, spurring further research and applications. New uses of AI that are of interest to governments will continue to emerge and will bring with them both benefits and risks. It is important that governments prepare for this future now by investing in building responsible practices for how they procure AI.

Finally, government procurement rules and purchasing practices often have a strong influence on markets, particularly in their early stages of development. As industry debates setting its own standards on these technologies, the government’s moral authority and credibility can help set a baseline for these discussions.

Overall, the guidelines aim to guide all parties involved in the procurement life cycle – policy officials, procurement officers, data scientists, technology providers and their leaders – towards the overarching goal of safeguarding public benefit and well-being.

# How were these guidelines developed?

The guidelines were developed by the World Economic Forum Centre for the Fourth Industrial Revolution, in consultation with a multistakeholder community. Project fellows from the UK Government's Office of AI, Deloitte and Salesforce worked with Forum staff, and in partnership with Splunk-convened workshops with appropriate representatives from government, academia, civil society and the private sector to explore key issues and co-design responses.

## How to use the guidelines

The guidelines provide fundamental considerations that a government should address before acquiring and deploying AI solutions and services. They apply once it has been determined that the solution needed for a problem *could* be AI-driven. The guidelines are not intended as a silver bullet for solving all public sector AI-adoption challenges, but by influencing how new AI solutions are procured, they can set government use and adoption of AI on a better path.

Specifically, the guidelines will help:

- Policy officials to accelerate attainment of their policy goals
- Procurement officials and commercial teams to develop AI-related requests for proposals and to manage procurement processes
- Data practitioners (e.g. statisticians, data scientists, digital and technology experts) to safeguard public benefit and identify and manage potential risks
- AI-solutions providers to better understand the core expectations for government AI projects and to align their proposals with emerging standards for public procurement

The guidelines consist of 10 high-level recommendations, ordered roughly sequentially in terms of their relevance to the cumulative process of procurement, each containing:

- Multiple principles relating to each guideline
- Explanatory text elaborating on the thinking and substance underlying each principle

As the technological sophistication, and the government use, of AI evolves, the guidelines should be updated to reflect new learning and leading practices. This is a living document that is intended to integrate feedback from practitioners over time. Much of that feedback will come from two sources: the project's community of subject matter experts, and the pilots to be held with the UK, the United Arab Emirates, Colombia and other partner governments. We also welcome feedback from other stakeholders and the general public. If you wish to provide feedback, please share via email: [AI@weforum.org](mailto:AI@weforum.org).

Ultimately, the goal is that these guidelines will enable governments and international bodies to set the right policies, protocols and perhaps even standards to facilitate effective, responsible and ethical public use of AI.

# Guidelines overview

What are the key considerations when starting a procurement process, writing a request for proposal (RFP), and evaluating RFP responses?

Guideline	Principles
<b>1. Use procurement processes that focus not on prescribing a specific solution but rather on outlining problems and opportunities, and allow room for iteration.</b>	<ul style="list-style-type: none"><li>a. Make use of innovative procurement processes to acquire AI systems.</li><li>b. Focus on developing a clear problem statement, rather than on detailing specifications of a solution.</li><li>c. Support an iterative approach to product development.</li></ul>
<b>2. Define the public benefit of using AI while assessing risks.</b>	<ul style="list-style-type: none"><li>a. Set out clearly in your RFP why you consider AI to be relevant to the problem and be open to alternative technical solutions.</li><li>b. Explain in your RFP that public benefit is a main driver of your decision-making process when assessing proposals.</li><li>c. Conduct an initial AI risk and impact assessment before starting the procurement process, ensure that your interim findings inform the RFP, and revisit the assessment at decision points.</li></ul>
<b>3. Align your procurement with relevant existing governmental strategies and contribute to their further improvement.</b>	<ul style="list-style-type: none"><li>a. Consult relevant governmental initiatives such as AI national strategies, innovation and/or industrial strategies, and guidance documents informing public policy about emerging technologies.</li><li>b. Collaborate with other relevant government bodies and institutions to share insights and learn from each other.</li></ul>
<b>4. Incorporate potentially relevant legislation and codes of practice in your RFP.</b>	<ul style="list-style-type: none"><li>a. Conduct a review of relevant legislation, rights, administrative rules and other relevant norms that govern the types of data and kinds of applications in scope for the project and reference them in the RFP.</li><li>b. Take into consideration the appropriate confidentiality, trade-secret protection, and data-privacy best practices that may be relevant to the deployment of the AI systems.</li></ul>
<b>5. Articulate the technical and administrative feasibility of accessing relevant data.</b>	<ul style="list-style-type: none"><li>a. Ensure that you have proper data governance mechanisms in place from the start of the procurement process.</li><li>b. Assess whether relevant data will be available for the project.</li><li>c. Define if and how you will share data with the vendor(s) for the procurement initiative and the subsequent project.</li></ul>
<b>6. Highlight the technical and ethical limitations of intended uses of data to avoid issues such as historical data bias.</b>	<ul style="list-style-type: none"><li>a. Consider the susceptibility of data that could be in scope and if usage of the data is fair.</li><li>b. Highlight known limitations (e.g. quality) of the data in the RFP and require tenderers to describe their strategies on how to address these shortcomings. Have a plan for addressing relevant limitations that you may have missed.</li></ul>
<b>7. Work with a diverse, multidisciplinary team.</b>	<ul style="list-style-type: none"><li>a. Develop ideas and make decisions throughout the procurement process in a multidisciplinary team.</li><li>b. Require the successful bidder(s) to assemble a team with the right skill set.</li></ul>
<b>8. Focus throughout the procurement process on mechanisms of algorithmic accountability and of transparency norms.</b>	<ul style="list-style-type: none"><li>a. Promote a culture of accountability across AI-powered solutions.</li><li>b. Ensure that AI decision-making is as transparent as possible.</li><li>c. Explore mechanisms to enable interpretability of the algorithms internally and externally as a means of establishing accountability and contestability.</li></ul>
<b>9. Implement a process for the continued engagement of the AI provider with the acquiring entity for knowledge transfer and long-term risk assessment.</b>	<ul style="list-style-type: none"><li>a. Consider during the procurement process that acquiring a tool that includes AI is not a one-time decision; testing the application over its lifespan is crucial.</li><li>b. Ask the AI provider to ensure that knowledge transfer and training are part of the engagement.</li><li>c. Ask the AI provider for insights on how to manage the appropriate use of the application by non-specialists.</li></ul>
<b>10. Create the conditions for a level and fair playing field among AI solution providers.</b>	<ul style="list-style-type: none"><li>a. Reach out in various ways to a wide variety of AI solution providers.</li><li>b. Engage vendors early and frequently throughout the process.</li><li>c. Ensure interoperability of AI solutions and require open licensing terms to avoid vendor lock-in.</li></ul>

# Detailed explanation of guidelines

## 1. Use procurement processes that focus not on prescribing a specific solution, but rather on outlining problems and opportunities and allow room for iteration.

### Why is it important?

To acquire the AI systems that best address the challenge you want to address and encourage responsible innovation.

- a. *Make use of innovative procurement processes to acquire AI systems.*
  - Innovation-oriented procurement procedures provide opportunities to accelerate the adoption of new technologies such as AI systems, to promote innovation and to support secondary policy criteria such as support for small and medium-sized enterprises and the ethical development of AI.
  - For example, these processes support early market engagement, enable you to go to market in different stages and can include the use of proofs of concept. These provide the opportunity to test the technologies
- on your problem area before making a final buying decision. Innovative public procurement processes that include practices such as detailing challenging problems, organizing technology contests, providing opportunities for demonstrators, and giving newly established providers the opportunity to compete for public-sector contracts, have the potential to boost innovation and help new companies become established. This market-making role also encourages small enterprises with new ideas and reduces the risks for new technology start-ups.
- By strategically choosing the procurement approach depending on the nature of the challenge that you mean to address, these processes could include, for example:
    - Agile procurement processes that allow you to go to market in different stages and can include proofs of concept to test the technologies before the final purchase.
    - Challenge-based procurement processes that have vendors compete against each other based on their AI skills and include an evaluation of the technologies applied to the challenges they mean to address.

Visual to depict the challenge-based procurement process used by the UK GovTech Catalyst challenge



- Innovation partnerships that enable the procurement of technologies that cannot be delivered by the current options available to the market.
- Dynamic purchasing systems – procedures currently used mainly for products commonly available on the market – can accelerate uptake of technologies that are rapidly developing. As a procurement tool, it is similar in some ways to an electronic framework agreement

but, as new suppliers can join at any time, this allows newly established firms to participate in the framework agreements when they meet the set criteria.

- AI procurement frameworks that prescribe the terms and conditions applying to any subsequent contract and allow the pre-vetting of providers against a set of predefined criteria that can include ethical requirements.

- When making use of novel approaches to procuring emerging technologies you should also focus on best practices that have been shown to increase the supplier base of smaller and innovative suppliers, which is important for fast-developing markets such as AI. These practices include but are not limited to:
  - Setting out and following a detailed procurement timeline at the start of the campaign.
  - Breaking down large proposals into smaller work components.
  - Encouraging collaboration between different bidders.
- b. *Focus on developing a clear problem statement, rather than on detailing the specifications of a solution.*
  - AI technologies are developing rapidly, with new technologies and products constantly being introduced to the market. By focusing on describing the challenges and/or opportunities that you want to address and drawing on the expertise of technology partners, you can better decipher what technology is most appropriate for the issue at hand. By focusing on the challenge and/or opportunity, you might also discover a higher-priority issue, or realize you were focusing on a symptom rather than the root cause.
  - Beyond playing to each stakeholder's strength, this approach has two added benefits. First, it demands and promotes early market engagement, which we explain in further detail in Guideline 10. Second, it makes it easier for newer AI service providers (such as start-ups) to participate, as the government will not be focused on a specific product. Nurturing an emerging AI ecosystem is a key economic investment in the future.
- c. *Support an iterative approach to product development.*
  - AI-powered solutions differ significantly from other technology tools in their unique ability to learn and adapt through ongoing, periodic training with new data. Therefore, the procurement process should allow room for iteration, while ensuring a robust, fair and transparent evaluation and decision process.
  - For example, a phased challenge-based procurement could serve to evaluate different competitors' minimum viable products (MVPs) during phase one of procurement, with only the winner going on to develop the full solution. This building and testing in phases within the procurement cycle facilitates informed decision-making, innovation and transparency. It also provides you with relevant information to conduct meaningful impact assessments and evaluate risks.

## 2. Define the public benefit of using AI while assessing risks.

### Why is this important?

Defining the public benefit goal provides an anchor for the overall project and procurement process that the AI is intended to achieve. AI also brings new and specific risks that must be identified and managed as early as the procurement phase of the project.

- a. *Set out clearly in your RFP why you consider AI to be relevant to the problem and be open to alternative technical solutions.*
  - In most circumstances, you should refer to the need for an AI solution in your invitation to tender only if there is strong indication that the technology will address the problem that you are trying to solve. A need for the acquisition of an AI system should arise through analysis of policy challenges and alternatives, and be compared to other potential courses of action when the AI project does not have a clear research and innovation focus. If, during the evaluation of the tender responses, it becomes evident that another solution that doesn't incorporate AI is better able to address the problem, you should make the decision to follow this alternative delivery path.
  - Assess whether AI could be part of a solution to your problem, before starting the procurement process. If you lack the capabilities in your team to carry out this assessment, you should seek these from elsewhere in your organization or relevant professional network (e.g. academia, trusted vendors) and make the consultation and collaboration with appropriate stakeholders a priority. For this assessment, it is crucial to engage a multistakeholder community to define and test a clear policy problem statement and reflect the findings in the RFP.
  - Pre-market engagement is also often essential in helping you to describe your problem and narrow down the tasks that AI may be able to assist with. This will help you better communicate to potential suppliers what you are asking for and why, as well as highlighting where the gaps are. Documenting user need and challenges to the best of your ability is crucial, since the success of the project also depends on how well AI system providers understand the problem.
- b. *Explain in your RFP that public benefit is a main driver of your decision-making process when assessing proposals.*
  - When setting out the requirements in the RFP, you should consider explicitly referring to public benefit as well as user needs. When determining user needs, public servants should be confident that they are acting in the public benefit. With regard to AI systems, the public benefit extends beyond value for money and also includes considerations about transparency of the decision-making process and other factors that are included in these guidelines.
  - In practice this requires you, for example, to specify success and failure criteria in the context of public benefit: What do you expect such a system to achieve and be capable of, and what are the types of failure and harm that must be avoided? Conducting an impact assessment will help you to set these issues out. Refer to Guideline 7 for additional information on adding ethical requirements to the RFP.
- c. *Conduct an initial AI risk and impact assessment even before starting the procurement process, ensure that your interim findings inform the RFP, and revisit the assessment at decision points.*

- To better understand the potential impacts of the use of AI and to mitigate the risks, it is important to start an assessment in a systematic manner before the acquisition of an AI system and to make sure that the findings also inform your commercial strategy. There will be different considerations depending on which policy challenges you are trying to solve and which potential application of AI could help to address this challenge. Without knowing which AI system you will acquire, it is not possible to conduct a whole assessment.
  - An initial assessment should outline user needs and affected communities, as well as potential risks such as inaccuracy and bias of the AI system. It should also include some consideration of scenarios involving unintended consequences. The initial assessment should make you think about strategies to address these potential impacts, including but not limited to citizen panels, transparency reports and testing on differentially private or synthetic datasets. Associated risks and their respective mitigation strategies must be recognized by a suitable risk owner with decision-making power and should include a go/no-go decision.
  - In your invitation to tender, you should consider asking potential suppliers to identify risks and explain how they would mitigate them. This can give you valuable information regarding how careful each tenderer is and how aware they
- are of potential risks. Where you identified significant risks in your initial assessments, you should specifically require tenderers to set out how they would address those.
- Data protection impact assessments and equality impact assessments can provide a useful starting point for assessing potential unintended consequences. In assessing these, you should consider how the use of these systems, such as semi-automated or solely automated decisions, interact with mechanisms of oversight, review and other safeguards. For examples of risk assessment questionnaires for automated decision-making, refer to the government of Canada's [Directive on Automated Decision Making](#), and the framework on [Algorithmic Impact Assessments](#) from AI Now.
  - In addition to the above, there should be systematic and continuous risk monitoring during every stage of the AI solution's life cycle, from design to post-implementation maintenance. AI solution providers can do this by identifying, drafting mitigations for and reporting risks through a project management function, which is central to the implementation (refer to Guideline 9 for more information on how to consider life-cycle management during the procurement process). The impact assessment should be revisited where necessary (e.g. in the event of significant changes to the opportunity statement).

Visual of the SDLC stages, with sample AI risk assessment question for each stage.

SDLC stage	Sample AI risk mitigation considerations
1. Requirements gathering and analysis	<ul style="list-style-type: none"> <li>- Is the use of AI/ML necessary for the desired outcome?</li> <li>- Should AI/ML even be discussed at this stage?</li> </ul>
2. Design	<ul style="list-style-type: none"> <li>- Do we have consent to use the data sources required by the solution?</li> <li>- Do we fully understand the implications of using external data, models or solutions?</li> </ul>
3. Implementation and coding	<ul style="list-style-type: none"> <li>- Do we have the right skills or domain expertise to develop the solution?</li> <li>- Does the development process protect data confidentiality and integrity?</li> </ul>
4. Testing	<ul style="list-style-type: none"> <li>- What level and type of bias is acceptable in the solution?</li> <li>- Do the acceptance criteria set appropriate levels of accuracy to ensure the model performance is satisfactory?</li> </ul>
5. Deployment	<ul style="list-style-type: none"> <li>- Have users received adequate training to ensure they understand the output of the system?</li> <li>- Is it transparent to users how the solution is deriving an output?</li> </ul>
6. Maintenance	<ul style="list-style-type: none"> <li>- Do the system administrators know what metrics to examine to validate that models are operating as expected?</li> <li>- Is there a clear process for updating or refining models using new data?</li> </ul>

### **3. Aim to include your procurement within a strategy for AI adoption across government and learn from others.**

#### **Why is it important?**

To ensure that you use procurement strategically to support efforts on AI development and deployment, and to spread the knowledge of the public application of an emerging technology.

- a. *Consult relevant AI national strategy initiatives and guidance documents from ministries and departments informing public policy of emerging technologies.*
- Many countries are currently in the process of drafting and releasing national AI strategies, and some have already published theirs. Prior to commencing an AI rollout, evaluate how your pursuit of acquiring an AI system aligns to your country's overall strategy.
- This allows you to include secondary policy aims in your strategic procurement and potentially make use of economies of scale by pooling the demand for AI systems. An added benefit of aligning to a national AI strategy is that there may be special support for initiatives that align to the strategy, such as access to additional experts.
- b. *Consult with government agencies that have looked into procuring AI solutions, irrespective of the outcome of the procurement efforts.*
- To improve your practices and share your experiences, you could actively seek out collaboration across departments and fields of expertise. You could also share knowledge and feedback via expert communities, such as the digital-buying community, professional networks or meet-ups.
- Within your department it can be helpful to set up platforms and networks that allow for the exchange of information, experiences and best practices about the purchasing of AI-powered solutions.

### **4. Ensure that legislation and codes of practice are incorporated in the RFP.**

#### **Why is this important?**

Conforming with existing laws and regulations ensures compliance; incorporating codes of practices supports the standardization of norms; and surveying the relevant rules enables better policy-making in a dynamic innovation technology ecosystem.

- a. *Conduct a review of relevant legislation, rights, administrative rules and other relevant norms that govern the types of data and kinds of applications in scope for the project.*
- Conduct a review of relevant legislation, human and civil rights, administrative rules, and other relevant norms that govern the types of data and kinds of applications connected to the problem being addressed and solutions being proposed. Clarify the appropriate adjudicative and administrative jurisdictions within the domestic government in relation

to conflicts of laws concerning the data. Depending on the problem being addressed in the invitation to tender, existing laws and regulations relevant to that government function may already have some rules on the use, processing, transfer etc. of data. Incorporate those rules and norms into the RFP by referring to the originating laws and regulations.

- When identifying the relevant rules, sources should include not only formal law, but also industry best practices, trade organization consensus guidelines and other forms of norm-setting mechanisms of soft law. For example, freedom of information laws<sup>2</sup> establish rules about what needs to be made available to the public, and data ethics frameworks guide the design of appropriate data use in government and the wider public sector.
- b. *Take into consideration the appropriate confidentiality, trade secret protection and data privacy best practices that may be relevant to the deployment of the AI solutions.*
- To meaningfully evaluate proposed AI solutions, government officials must strike the right balance between preserving accountability through transparency and reassuring vendors that the trade secrets associated with their products and services, as well as their business confidentiality, will not be compromised. Information about government processes should be open by default, with the limits of disclosure justified in exceptional circumstances such as export controls, national security or ongoing criminal investigations.
- In those circumstances where confidentiality and trade-secrecy protection can be justified in light of public-interest considerations, investigate the possibilities of facilitating transparency through partial disclosure, limited review options and other means of enhancing public trust.

### **5. Articulate the technical feasibility and governance considerations of obtaining relevant data.**

#### **Why is this important?**

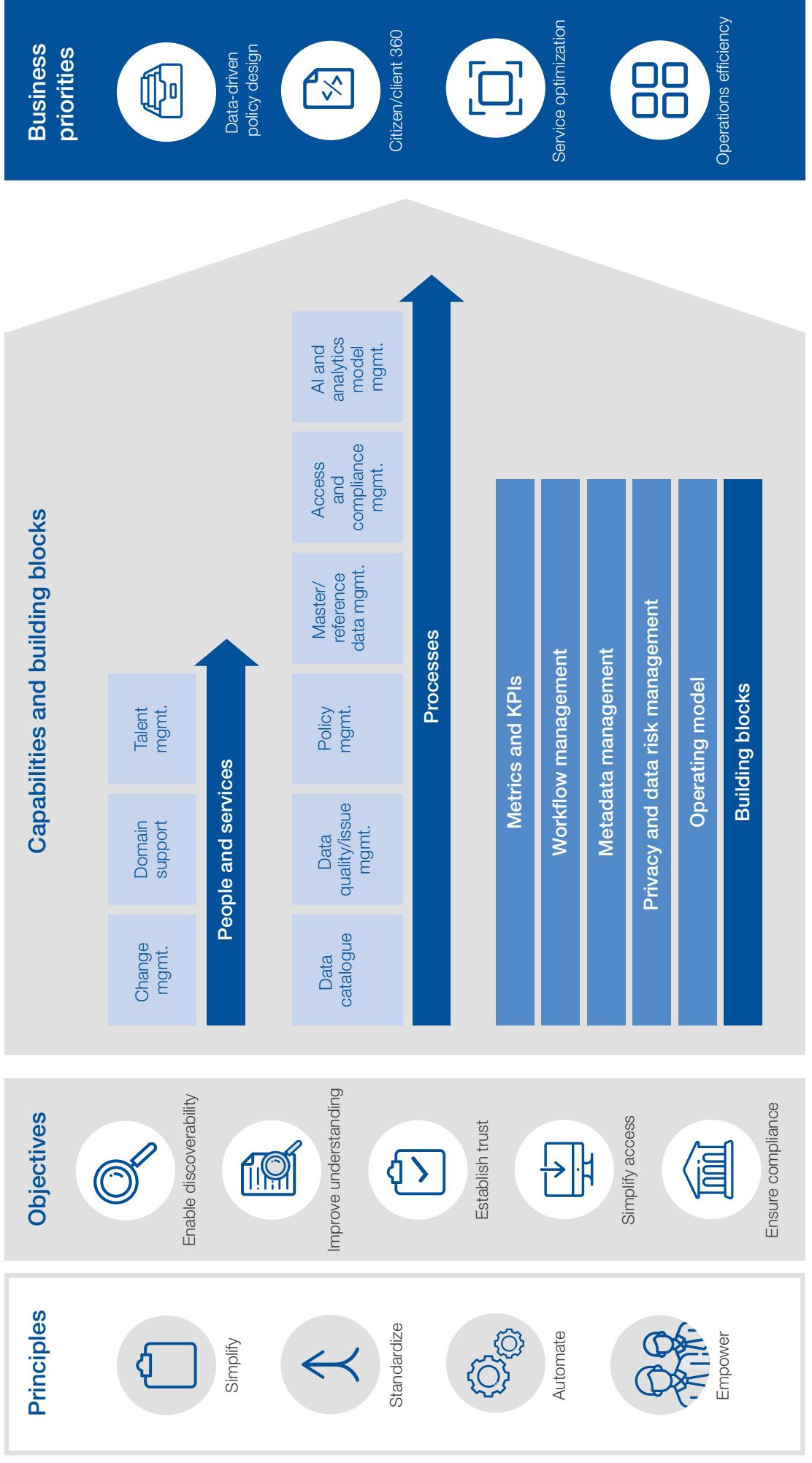
Availability of relevant data is a prerequisite for any AI solution, so time should not be spent discussing AI procurement if no data will be available. In addition, access to data should be granted only after careful consideration by the data-governing party(ies).

- a. *Ensure that you have proper data-governance mechanisms in place from the start of the procurement process.*
- Set out a data-governance approach from the start of the procurement process. Given the importance and complexity of data governance, it is almost mandatory to implement sound data-governance processes before engaging in transformative AI projects. Governance needs to cover all data activities related to the proposed project, such as granting data access to project members, moving/storing data in other locations for analysis, and reviewing data consent (the purposes for which we are authorized to use the data).

## Sample data governance framework

Deloitte's data governance framework enables organizations to be specific in terms of what goals will be prioritized, what capabilities will be deployed and what results are expected

### Knowledge and insights (AI) strategic objectives



- Data governance, and all other aspects of an AI initiative, require a multidisciplinary team. Refer to Guideline 7 for more information on multidisciplinary teams.
  - In the absence of a data-governance framework, ensure that it is clear who is accountable (who is responsible for data management during the procurement process and the subsequent project).
- b. Assess whether relevant data will be available for the project.
- Data is crucial for modern-day AI tools. You should determine, at a high level, data availability before starting your procurement process. This entails developing an understanding of what data might be required for the project. The idea is not to assess all possible data sources, but to build general awareness of data sources of potential interest. Data documentation, using data dictionaries, for example, is helpful when trying to build a high-level understanding of data assets.
  - In cases where data is not available for the use case in mind, you may be able to find data through third parties, for example, vendors, partners or data brokers. If data is not available through any channel, engage skilled data scientists (for example, through vendors) to assess whether the use case can be addressed at all in a data-driven manner.
- c. Define if and how you will share data with the vendor(s) for the procurement initiative and the subsequent project.
- Depending on the sensitivity of your project and data, it is worth considering the release of data to providers during procurement so that bidders can craft a response to the RFP that is tailored to your needs, with assumptions, timelines and fees that match your situation as closely as possible. This improves the quality of RFP responses you receive.
  - If you are releasing data that is sensitive and not meant for public consumption, consider releasing only a sample, so that vendors have a clear idea of what the data enables them to do without having access to all of it. When doing this, make sure that you provide a sample that is representative of the overall dataset. Otherwise, vendors might make erroneous assumptions that can impact the quality of bids and consequently the integrity of the project.
  - Create and document the appropriate data-sharing conditions. For example:
    - Minimum requirements for the environment where the vendor will host the data (e.g. enterprise laptop that meets the vendor's standards for their sensitive data).
    - Data consent form signed by the vendor's lead for the project, stating that the data will be used exclusively for the pursuit and for no other purpose. It should be clear to vendors that while in possession of the data they are not allowed to use the data for any purpose other than that specified in the RFP.
    - Date for data deletion (e.g. immediately upon submission of the vendor's RFP response). In no circumstances should governments allow vendors to keep data after the procurement process, or after the conclusion of the project for successful bidders.
- Confirmation of deletion of all data (e.g. written confirmation of deletion signed and submitted by the vendor's lead for the project).
  - There are many anonymization techniques to help safeguard data privacy, including data aggregation, masking and synthetic data.<sup>3</sup> Keep in mind, however, that you must manage anonymized data as carefully as the original data, since it may inadvertently expose important insights. RFPs should encourage innovative technological approaches, such as those mentioned above, that make less intrusive use of data or that achieve the same or similar outcomes with less sensitive datasets.
  - Certain vendors may have data that is complementary to the initiative, and it is in your best interest to consider using this data. It is important to have a framework that gives guidance regarding the circumstances under which it is reasonable to accept data from a vendor. Decision criteria could include:
    - Vendor: some vendors could be pre-qualified as accepted data providers, be considered more trustworthy as a result of their previous track record as existing suppliers or have a strong reputation related to their data assets.
    - Domain: some domains – such as health, justice and immigration – are very sensitive. Use of third-party data in these domains requires careful scrutiny before it is accepted.
    - Data precedence and integrity: before using any third-party data, the government should have a clear understanding of how the data was collected, the governance processes employed to ensure its integrity, and whether the third party offering the data is legally allowed to commercialize it for the RFP.

## 6. Highlight the technical and ethical limitations of using the data to avoid issues such as bias.

### Why is this important?

Though available, legal to use and proportionate to need, there may be limitations to data (e.g. data bias) that make an AI approach inappropriate, unreliable or misleading.

- a. Consider the susceptibility of data that could be in scope and whether usage of the data is fair.
  - As important as data protection is, not all data is sensitive (e.g. open-government data is freely accessible online). All data, sensitive or not, must have its integrity safeguarded, but it is not necessary to keep non-sensitive data behind closed doors. It is important to assess the privacy needs of different datasets to determine the right level of protection. Normally, personally identifiable information (PII), such as financial and health data, is considered extremely sensitive. The RFP needs to reflect data governance requirements for both the procurement process and the project that are in accordance with the nature of the data.
  - Select data that fits criteria of fairness. For example, the data should be representative of the population that the AI solution will address, as well as being reasonably recent.<sup>4</sup>

- b. *Highlight known limitations of the data (e.g. quality) in your RFP and require tenderers to describe their strategies on how to address these shortcomings. Have a plan for addressing relevant limitations that you may have missed.*
- Considerations when deciding if a source of data is suitable include:<sup>5</sup>
  - Representativeness (whether the data accurately represents the segment of the population in scope for the AI solution)
  - Provenance (including how and why the data was collected)
  - Gaps in data quality (e.g. many values missing from a particular data element)
  - Bias present in the data (if it is not representative of the population to which the algorithm will be applied)
  - Lack of clarity in metadata (for example, confusing or vague data element names)
  - Check data completeness, representativeness and accuracy of potential sources before starting the procurement process. Articulate data quality observations and the apparent limitations and, if possible, share those insights through the RFP. Bidders must be aware of these data considerations during the procurement process or, in cases where data is sensitive, the selected provider(s) must be made aware after the contract has been awarded.
  - If you do not have the right skills or means to comprehensively check for possible limitations of your data, provide vendors with guiding insights into the high-level state of the data and its origin,<sup>6</sup> so that they can draft adequate proposals. Also, ensure the RFP's data requirements include performing a comprehensive data quality assessment and, if required, development of mitigation strategies for low-quality data.

## 7. Work with a diverse, multidisciplinary team.

### Why is this important?

Developing and fulfilling a proper AI RFP will require a diverse team that understands the interdependent disciplines that AI covers, including: domain expertise (e.g. healthcare, transportation), systems and data engineering, model development (e.g. deep learning) and visualization/information design, among others.

- a. *Develop ideas and make decisions throughout the procurement process in a diverse and multidisciplinary team.*
- Develop an understanding of the skills that are needed to effectively acquire and maintain an AI-powered solution, before starting the procurement process.
  - Assemble multidisciplinary teams that specialize in designing, procuring and operationalizing AI systems. These multidisciplinary teams should include expertise

in: policy from the domain (e.g. justice) in which the AI solution will be applied, machine learning/data science, data engineering, technology (software and hardware), procurement, ethics and human rights.<sup>7</sup>

- Ensure that you have a diverse team. This should include people from different genders, ethnicities, socioeconomic backgrounds, disabilities and sexualities. You should also make sure that there is a mix of perspectives and points of view. This ensures that problems and solutions are tackled from different angles and helps to mitigate bias.
- If expertise is lacking within your team, you can reach out to pools or professional networks within your organization or across the civil service.

Note that many value-laden decisions will likely be made during development (i.e. post-procurement), and it is essential that your team maintains the skills, or at the very least access to expertise, to ensure that all important decisions and trade-offs are made or overseen internally, rather than exclusively by a contractor or vendor.

- b. *Require the successful bidder(s) to assemble a team with the right skill set.*
- As part of your requirements, ensure bidders provide evidence of the skills and qualifications of the proposed project resources who will develop and deploy the AI solution.<sup>8</sup> This should be part of the RFP response and it should be one of your decision criteria when evaluating the proposals.

## 8. Focus throughout the procurement process on mechanisms of accountability and transparency norms.

### Why is this important?

To build public trust in the legitimacy of the AI system, the procurement process should enable accountability in understanding how the AI solution works, so that it can be evaluated independently and thus promote a culture of responsibility over the AI solution life cycle.

- a. *Promote a culture of accountability across AI-powered solutions.*
- Public institutions cannot rely on black-box algorithms to justify decisions that affect individual and collective citizens' rights, especially with the increased understanding about algorithmic bias and its discriminatory effects on access to public resources. There will be different considerations depending on the use case and application of AI that you are aiming to acquire, and you should plan to work with the supplier to explain the application for external scrutiny, ensuring your approach can be held to account. These considerations should link to the risk and impact assessment described in Guideline 2. Under certain scenarios, you could consider making it a requirement for providers to allow independent audit(s) of their solutions. This can help prevent or mitigate unintended outcomes.

## Diagram to explain what is meant by a ‘black box’ algorithm and why they’re an issue

In a traditional model where a service provider interfaces with the service recipient, the recipient can communicate back and forth with the service provider regarding an outcome and/or determination. The recipient can ask questions regarding a decision and challenge an outcome.



With a fully automated system that uses a technique such as neural networks, the service recipient cannot expect to understand the outcome. This is because certain algorithms, such as neural networks, are very accurate but do not explain their path to a decision.



- Providers and public officials should incorporate risk analysis for the unexpected and unintended effects of AI-powered solutions, within the limits prescribed by the law, and specify their respective responsibilities in the contract. Note that the laws and standards for assigning accountability may differ according to jurisdiction. For example, the Canadian federal government's Directive on Automated Decision-Making requires the associate deputy minister of the respective federal entity to sign off on an algorithmic impact assessment (AIA) as part of an AI project.
  - Consider how applicable accountability requirements in law, such as freedom of information legislation and data-protection logging requirements, will be implemented throughout the project life cycle.
- b. Ensure that AI decision-making is as transparent as possible.
- Encourage transparency of AI decision-making (i.e. the decisions and/or insights generated by AI). One way to do this is to encourage the use of explainable AI. You can also make it a requirement for the bidder to provide the required training and knowledge transfer to your team, even making your team part of the AI-implementation journey. Finally, you can ask for documentation that provides information about the algorithm (e.g. data used for training, whether the model is based on supervised, unsupervised or reinforcement learning, or any known biases).
- Documentation is especially important when the algorithm is a pre-packaged solution that the bidder will bring to the project, as opposed to an algorithm that will be built and/or customized as part of the upcoming project. Finally, you can also ask bidders to provide information on their model-building methodology, including how they select variables, build samples (where applicable) and validate the model. Be aware, however, that algorithm-building is an iterative process and that it depends on creativity as much as it does on science.
  - Documentation provided by a bidder will give you directional awareness of their practices and methods; it will not give you a step-by-step guide that details exactly what would be done during the project, as the exact process will invariably shift from project to project to meet the needs of each scenario.
- c. Explore mechanisms to enable interpretability of the algorithms internally and externally as a means of establishing accountability and contestability.
- With AI solutions that make decisions affecting people's rights and benefits, it is less important to know exactly how a machine-learning model has arrived at a result if we can show logical steps to achieving the outcome. In other words, the ability to know how and why a model performed in the way it did is a more appropriate means of evaluating transparency in the context of AI. For example, this might include what training data was used,

which variables have contributed most to a result, and the types of audit and assurance the model went through in relation to systemic issues such as discrimination and fairness. This should be set out as documentation needed by your supplier.

- It is also important to consider the potential tension between explainability and accuracy of AI when acquiring AI solutions. Classic statistical techniques such as decision-tree models are easier to explain but might have less predictive power, whereas more complex models, such as neural networks, have high predictive power but are considered to be black boxes. Given these challenges you should think carefully about achieving the right balance between accuracy and transparency in the AI-powered solution procured, especially for topics of great social concern (e.g. healthcare, education) where citizens might expect full transparency. Address this concern in the RFP.
- If an algorithm will be making decisions that affect people's rights and public benefits, describe how the administrative process would preserve due process by enabling the contestability of automated decision-making in those circumstances.

## 9. Implement a process for the continued engagement of the AI provider with the acquiring entity for knowledge transfer and long-term risk assessment.

### Why is this important?

The functionality and consequences of AI systems may not be apparent in the procurement process and often become evident only over the duration of its application, requiring extended communication and information-sharing between the procuring entity and the system developer.

For AI systems in the public sector, sustainable and ongoing evaluation methods and means of providing feedback on the data model are crucial to ensure that the tool's use remains ethical. You should make clear in your RFP that this should be considered by the provider and discussed as part of the procurement process.

- a. Consider during the procurement process that acquiring a tool that includes AI is not a one-time decision; testing the application over its lifespan is crucial.

- The tool will need support during its life cycle. Knowing where to go for that support and how much support is available will be vital for getting the most out of any tool. Accepting the potential impact of any support gaps or employing outside expertise both come at a cost. This should be factored in when purchasing an intelligent tool.
- Consider the implementation of a process-based governance framework that provides a template for the integration of the norms, values and principles that inform the procedures and protocols organizing the project workflow.
- Testing the model on an ongoing basis is necessary to maintain its accuracy. An inaccurate model can result in erroneous decisions and affect users of public services.

Therefore, you should establish with the provider how the efficacy of the model will be monitored once deployed.

- b. Ask the AI provider for knowledge transfer and training to be part of the engagement.
  - Make knowledge transfer a requirement under the RFP. Evaluate the thoroughness and logic of the knowledge-transfer plan to ensure that government resources will be able to use the tool appropriately on their own once the project is finalized.
  - Set out clearly your expectations for project documentation. Ensure that maintenance and auditing of the AI solution would be possible by entities independent of the vendor.
- c. Ask the AI provider for insights on how to manage the appropriate use of the application by non-specialists.
  - Operational or service staff must have enough knowledge or training on the solution to understand how to use it and successfully exploit its outputs. You should address the need for enough training and support to avoid the misuse of AI applications with the AI provider. The application must make it easy to report any suspected unauthorized behaviour to the relevant authority(ies) within and/or outside the organization. Enable end-to-end auditability with a process log that gathers the data across the modelling, training, testing, verifying and implementation phases of the project life cycle. Such a log will allow for the variable accessibility and presentation of information with different users in mind to achieve interpretable and justifiable AI.
- d. Make ethical considerations part of your evaluation criteria for proposals.
  - There are robust ethical practices that you should require suppliers to demonstrate when providing AI solutions. Leading AI-solution providers have begun to create internal frameworks for the ethical design, development and deployment of AI, which cover processes to ensure accountability over algorithms, avoiding outputs of analysis that could result in unfair and/or biased decision-making, designing for reproducibility, testing the model under a range of conditions and defining acceptable model performance. Bidders should be able not only to describe their approach to the above, but also to provide examples of projects, complete with client references, where these considerations have been followed.<sup>9</sup>
  - Make comprehensive, transparent algorithm assessment one of the requirements in the proposal and, if applicable, state minimum performance metrics that the model must meet. If possible, work with bidders to determine what the thresholds should be. As part of testing the model, you should work with the provider to establish how often you need to update the model with new data. Testing over the lifespan of the model for suitability and accuracy is highly important, especially when the AI is supporting critical services.

## 10. Create the conditions for a level and fair playing field among AI solution providers.

### Why is this important?

Government spending can be used to create a fair, competitive market, which leads to better AI systems. In addition, early engagement with AI vendors can result in more relevant responses, increasing the probability of success for the procurement and the subsequent project.

- a. *Reach out in various ways to a wide variety of AI solution providers.*
  - Given the rapidly developing landscape of AI service providers, largely comprising smaller enterprises such as start-ups, consider non-traditional methods of market engagement to attract AI solution providers. For example, explain the needs that lead to the proposal through in-person presentations, webinars, information sessions at co-working spaces and/or online platforms such as LinkedIn or Twitter.
  - Consider reaching out to non-traditional stakeholders, such as research institutes and academia. In some cases, these may have the right skills to be part of an AI implementation, and in all cases, they can act as advisers.<sup>10</sup>
  - Keep in mind that successfully designing and deploying AI in organizations as big and complex as public agencies requires much more than technical expertise. It requires experience in change management, familiarity with public organizations, and the ability to manage complex projects.
- b. *Engage vendors early and frequently throughout the process.*
  - Market engagement is a process; it takes place prior to procurement and aims to identify potential bidders and/or solutions, build capacity in the market to address challenges and opportunities, and inform the design of the procurement and contract.
  - Early engagement between government and industry is vital to a successful AI purchasing campaign. Early supplier engagement can help to determine the scope and feasibility of the RFP and, in turn, the most appropriate way to design and structure the requirements, increasing the likelihood that the winning bidder will meet your needs at a competitive cost. Ways to engage vendors early include having vendors provide inputs on possible evaluation criteria for the RFP, and hosting vendors to walk them through the RFP. Approaches like this are already being deployed in Canada, for example, and greatly help government and the private sector increase the effectiveness of procurement.
  - To mitigate any risks that could be associated with market engagement (e.g. commercial confidentiality, protection of intellectual property [IP], fettering discretion of tender process), be sure to broadly advertise the engagement opportunity, allow all interested parties to participate, ensure that there is adequate time for responses and reasonable time for bidder selection and, where appropriate, that RFP responses can be marked as confidential.
- c. *Ensure interoperability of AI solutions and require open licensing terms to avoid vendor lock-in.*
  - Consider strategies to avoid vendor lock-in, particularly in relation to black-box algorithms. These practices could involve the use of open standards, royalty-free licensing and public domain publication terms.
  - During the design and deployment of the AI solution, it is likely that either a new algorithm will be designed, or an existing one will be tailored (e.g. retrained through your data). It is therefore useful to consider whether your department should own that IP and how it would control it. The arrangements should be mutually beneficial and fair, and require royalty-free licensing when adopting a system that includes IP controlled by a vendor.
  - In order to preserve access to systems that become obsolete, ensure the ability to reverse-engineer the system to allow for maintenance of the AI solution independent of the vendor.

# Endnotes

1. Definition from the Engineering and Physical Science Research Council, a UK government research funding body.
2. For an up-to-date list of freedom of information laws around the world, see [https://en.wikipedia.org/wiki/Freedom\\_of\\_information\\_laws\\_by\\_country](https://en.wikipedia.org/wiki/Freedom_of_information_laws_by_country) (link as of 6 September 2019).
3. For more information on data anonymization, refer to: “Guide to basic data anonymisation techniques”, Personal Data Protection Commission, Singapore. 25 January 2018.
4. For more information on fairness during data selection, refer to: “Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector”, section “Data fairness”, David Leslie, the Alan Turing Institute.
5. For more information on data selection criteria, refer to: “Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector”, section “Data fairness”, David Leslie, the Alan Turing Institute.
6. For example, summary statistics such as number of rows present, number of missing values for each data field, description of how the data is collected and processed.
7. For more information on the domain and technical skills required to deliver an AI engagement, refer to: “Searching for superstars isn’t the answer. How organizations can build world-class analytics teams that deliver results”, Deloitte.
8. *ibid.*
9. AI ethics is a deep and evolving field, and there are various publications on the matter, including those listed below. Refer to these sources for a full background on the topic.
  - “OECD principles on artificial intelligence”, Organization for Economic Co-operation and Development
  - “Ethics guidelines for trustworthy AI”, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission
  - “Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector”, section “Data fairness”, David Leslie, the Alan Turing Institute.
  - “For a meaningful artificial intelligence. Towards a French and European strategy”, Cédric Villani
10. Examples of organizations include the Alan Turing Institute in the UK and the Vector Institute, MILA, and the Alberta Machine Intelligence Institute in Canada.



---

COMMITTED TO  
IMPROVING THE STATE  
OF THE WORLD

---

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

---

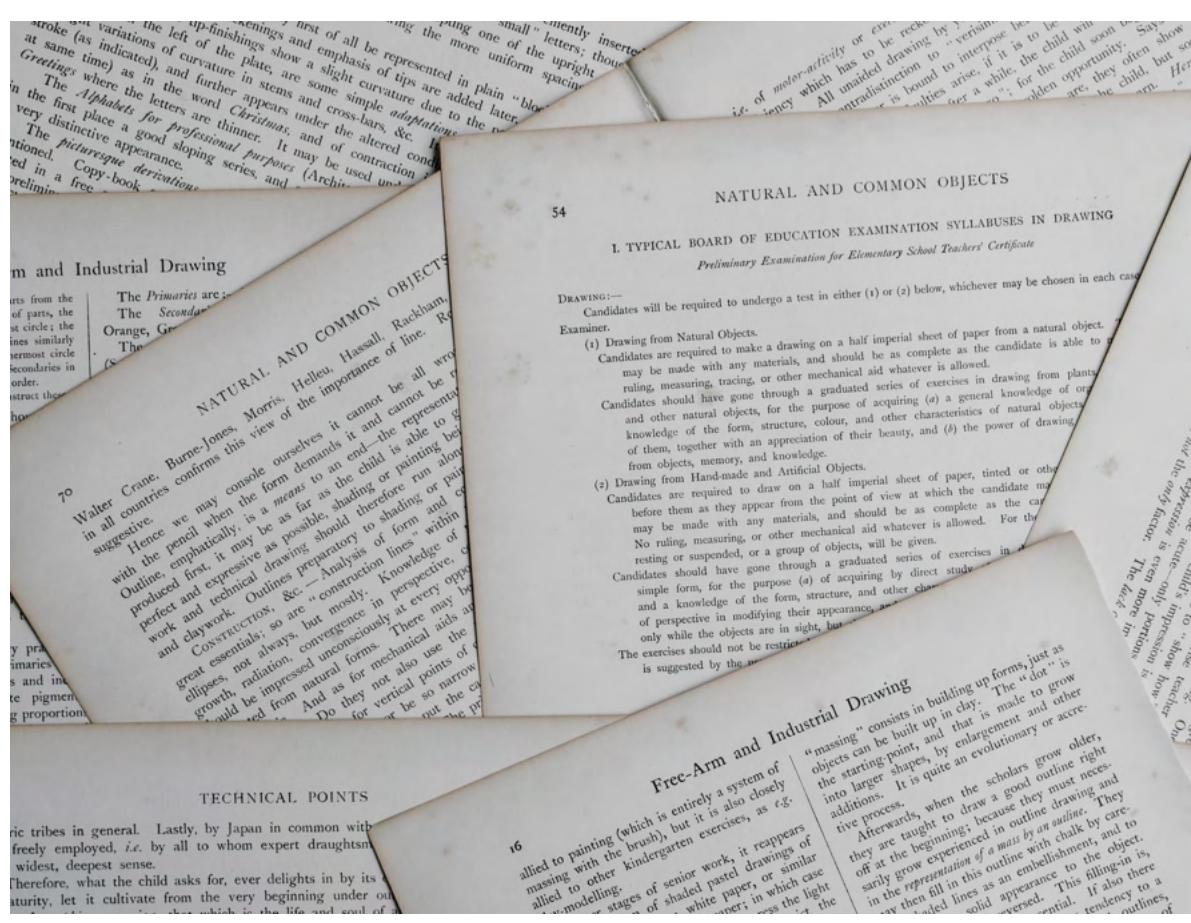
World Economic Forum  
91–93 route de la Capite  
CH-1223 Cologny/Geneva  
Switzerland

Tel.: +41 (0) 22 869 1212  
Fax: +41 (0) 22 786 2744

[contact@weforum.org](mailto:contact@weforum.org)  
[www.weforum.org](http://www.weforum.org)

# Ethics in AI research papers and articles

By: Kathy Baxter



*Last updated: July 15, 2021*

This is my obsessively curated list of research papers and articles on ethics in AI that I have been collecting over the years. Let me know if I am missing your favorites.

## 2021

### Peer-reviewed

[Racial/Ethnic Disparities in the Performance of Prediction Models for Death by Suicide After Mental Health Visits \(7/1/21\)](#)

[The Limits of Differential Privacy \(and Its Misuse in Data Release and Machine Learning\) \(July 2021\)](#)

[The Values Encoded in Machine Learning Research \(6/29/21\)](#)

[A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability \(6/23/21\)](#)

[External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients \(6/21/21\)](#)

[Are we done with ImageNet? \(6/12/21\)](#)

[Intrinsic Bias Metrics Do Not Correlate with Application Bias \(6/8/21\)](#)

[Measuring Model Biases in the Absence of Ground Truth \(6/6/21\)](#)

[Reliability Testing for Natural Language Processing Systems \(6/1/21\)](#)

## Explaining the Principles to Practices Gap in AI (June 2021)

Racial Segregation and the Data-Driven Society: How Our Failure to Reckon with Root Causes Perpetuates Separate and Unequal Realities  
(5/24/21)

Online Selection of Diverse Committees (5/19/21)

Towards accountability in the use of Artificial Intelligence for Public Administrations (5/18/21)

Including Signed Languages in Natural Language Processing (5/11/21)

A Step Toward More Inclusive People Annotations for Fairness  
(5/5/21)

To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes (May 2021)

[Denmark] "We Would Never Write That Down": Classifications of Unemployed and Data Challenges for AI (April 2021)

Precarity: Modeling the Long Term Effects of Compounded Decisions on Individual Instability (4/24/21)

Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation (4/20/21)

Engineering Bias Out of AI (4/20/21)

Moral zombies: why algorithms are not moral agents (4/16/21)

Autonomous Driving Ethics: from Trolley Problem to Ethics of Risk

(4/12/21)

How medicine discriminates against non-white people and women

(4/10/21)

Moving beyond “algorithmic bias is a data problem” (4/9/21)Evaluating eligibility criteria of oncology trials using real-world data and AI (4/7/21)How artificial intelligence could make clinical trials smarter (4/7/21)The Myth of Complete AI-Fairness (4/6/21)How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals (4/5/21)SENSEI: SENSITIVE SET INVARIANCE FOR ENFORCING INDIVIDUAL FAIRNESS (4/1/21)The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making (Apr 2021)INDIVIDUALLY FAIR GRADIENT BOOSTING (3/31/21)STATISTICAL INFERENCE FOR INDIVIDUAL FAIRNESS (3/30/21)Can an AI Algorithm Mitigate Racial Economic Inequality? An Analysis in the Context of Airbnb (3/22/21)Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets

(3/22/21)

Lawyers are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources (3/21/21)

INDIVIDUALLY FAIR RANKING (3/19/21)

Does “AI” stand for augmenting inequality in the era of covid-19 healthcare? (3/16/21)

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans (3/15/21)

People Analytics must benefit the people. An ethical analysis of data-driven algorithmic systems in human resources management (3/12/21)

The race to teach sign language to computers (3/6/21)

Measuring Model Biases in the Absence of Ground Truth (3/5/21)

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law (3/3/21)

Machines, Artificial Intelligence and rising global transphobia (3/1/21)

Narratives and Counternarratives on Data Sharing in Africa (3/1/21)

Artificial intelligence has a problem with grammar (2/27/21)

Sustainable AI: AI for sustainability and the sustainability of AI (2/26/21)

## Benchmarking and Survey of Explanation Methods for Black Box Models (2/25/21)

Ethics-Based Auditing to Develop Trustworthy AI (2/19/21)

The Privatization of AI Research(-ers): Causes and Potential Consequences (2/15/21)

Algorithmic injustice: a relational ethics approach (2/12/21)

The Limits of Global Inclusion in AI Development (2/7/21)

Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models (2/4/21)

Insiders and Outsiders in Research on Machine Learning and Society (2/3/21)

About Face: A Survey of Facial Recognition Evaluation (2/1/21)

Adversarial Learning with Cost-Sensitive Classes (1/29/21)

Hiding Behind Machines: When Blame Is Shifted to Artificial Agents (1/27/21)

Re-imagining Algorithmic Fairness in India and Beyond (1/25/21)

Black Feminist Musings on Algorithmic Oppression (1/25/21)

It's Too Easy to Hide Bias in Deep-Learning Systems (1/19/21)

Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases (1/19/21)

Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis  
(1/18/21)

Persistent Anti-Muslim Bias in Large Language Models (1/14/21)

AI Ethics: A Long History and a Recent Burst of Attention (1/14/21)

Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19 (Jan 2021)

The Role of Arts in Shaping AI Ethics (2021)

## **Government, NGO, and expert publications**

America's global leadership in human-centered AI can't come from industry alone (7/6/21)

Demystifying the Draft EU Artificial Intelligence Act (July 2021)

Research Summary: Experts Doubt Ethical AI Design Will Be Broadly Adopted as the Norm Within the Next Decade (6/29/21)

WHO issues first global report on Artificial Intelligence (AI) in health and six guiding principles for its design and use (6/28/21)

Mobilising the intellectual resources of the arts and humanities  
(6/25/21)

[China] Lofty principles, conflicting incentives: AI ethics and governance in China (6/24/21)

[Experts Doubt Ethical AI Design Will Be Broadly Adopted as the Norm](#)

[Within the Next Decade \(6/16/21\)](#)

[A Step Toward More Inclusive People Annotations in the Open Images](#)

[Extended Dataset \(6/15/21\)](#)

[\[videos\] NIST AI Measurement and Evaluation Workshop \(6/15/21\)](#)

[3 GUIDING PRINCIPLES TO KICK-START YOUR AI INITIATIVE](#)

[\(6/15/21\)](#)

[A Step Toward More Inclusive People Annotations in the Open Images](#)

[Extended Dataset \(6/15/21\)](#)

[The role of the arts and humanities in thinking about artificial](#)

[intelligence \(AI\) \(6/14/21\)](#)

[10 steps to educate your company on AI fairness \(6/9/21\)](#)

[Karlsruhe win against biometric mass surveillance in Germany](#)

[\(6/8/21\)](#)

[The current state of affairs and a roadmap for effective carbon-](#)

[accounting tooling in AI \(6/7/21\)](#)

[Managing the risk in AI: Spotting the “unknown unknowns” \(6/5/21\)](#)

[The Contestation of Tech Ethics: A Sociotechnical Approach to Ethics](#)

[and Technology in Action \(6/3/21\)](#)

[Facial Recognition Technology: Federal Law Enforcement Agencies](#)

[Should Better Assess Privacy and Other Risks \(6/3/21\)](#)

[EU] Guidance on Strengthening the Code of Practice on Disinformation (6/2/21)

Automated Decision-Making Systems in the Public Sector An Impact Assessment Tool for Public Authorities (June 2021)

AI GOVERNANCE IN 2020: A YEAR IN REVIEW: OBSERVATIONS FROM 52 GLOBAL EXPERTS (June 2021)

NIST: A Proposal for Identifying and Managing Bias in Artificial Intelligence (June 2021)

AI for disability inclusion: Enabling change with advanced technology (May 2021)

AI Ethics Has a Surplus Problem (5/26/21)

[Spain] In Catalonia, the RisCanvi algorithm helps decide whether inmates are paroled (5/25/21)

Understanding Contextual Facial Expressions Across the Globe (5/24/21)

People analytics in the workplace – how to effectively enforce labor rights (5/19/21)

Tech for disabled people is booming around the world. So where's the funding? (5/19/21)

Data Capitalism (2021)

The Chief AI Ethics Officer: A Champion or a PR Stunt? (5/16/21)

Image classification algorithms at Apple, Google still push racist tropes

(5/14/21)

US House AI Task Force Is the latest authority to address algorithms

and racism (5/12/21)

Detecting and mitigating bias in natural language processing

(5/10/21)

[Italy] In Italy, general practitioners and some regions adopt COVID-19

vaccine prioritization algorithms (5/19/21)

Can a COVID-19 face mask protect you from facial recognition

technology too? (5/10/21)

New research helps make AI fairer in decision-making (5/7/21)

[France] Evaluation Methods for Content Recommendation Algorithms

(5/6/21)

Hard choices: AI in health care (2021)

The hidden work created by artificial intelligence programs (5/4/21)

Reclaim Your Face – A European Citizens Initiative to ban biometric

mass surveillance (5/4/21)

[Greece] Greek camps for asylum seekers to introduce partly

automated surveillance systems (4/27/21)

Will Artificial Intelligence Foster or Hamper the Green New Deal?

(4/22/21)

[UK] [Home Office algorithm to detect sham marriages may contain built-in discrimination \(4/19/21\)](#)

[Aiming for truth, fairness, and equity in your company's use of AI \(4/19/21\)](#)

[Face Recognition Is Far from the Sci-Fi Super-Tool Its Sellers Claim \(4/16/21\)](#)

[61 MEPs urge the EU to ban biometric mass surveillance \(4/16/21\)](#)

[France] [How French welfare services are creating 'robo-debt' \(4/15/21\)](#)

[Singapore] [Are we stuck with Covid-19 surveillance? – Join the democracy classroom \(4/12/21\)](#)

[The Abuse and Misogynoir Playbook, explained \(workshop transcript\) \(4/8/21\)](#)

[Emails between NYPD and Clearview AI obtained by FOIL request prompt questions \(4/6/21\)](#)

[Research Summary: Mapping the Ethicality of Algorithmic Pricing \(4/5/21\)](#)

[Identity systems and social protection in Venezuela and Bolivia: GENDER IMPACTS AND OTHER INEQUALITIES \(4/5/21\)](#)

[Research Summary: Algorithmic Impact Assessments – What Impact Do They Have? \(4/4/21\)](#)

[If Your Company Uses AI, It Needs an Internal Review Board \(4/1/21\)](#)

Error-riddled data sets are warping our sense of how good AI really is

(4/1/21)

[UK] Technology Managing People – the legal implications (Mar 2021)

Six Unexamined Premises Regarding Artificial Intelligence and National Security (3/31/21)

Are You Overestimating Your Responsible AI Maturity? (3/30/21)

Responsible AI: From principles to practice (3/30/21)

[China] Research Summary: The Chinese Approach to AI: An Analysis

of Policy, Ethics, and Regulation (3/29/21)

The IEEE Trusted Data and Artificial Intelligence Systems (AIS)

Playbook for Financial Services (3/29/21)

Using Clinical Text to Combat Selection Bias in Medical Research

(3/29/21)

From Principles to Practice: An interdisciplinary framework to

operationalize AI ethics (2020)

How AI Can Help Companies Set Prices More Ethically (3/26/21)

[UK] Technology Managing People – the legal implications (3/25/21)

Alation State of Data Culture Report Reveals Barriers in Adopting

Artificial Intelligence (3/24/21)

EUGENICS POWERS IQ AND AI (3/24/21)

[A Better Measuring Stick: Algorithmic Approach to Pain Diagnosis](#)

[Could Eliminate Racial Bias \(3/24/21\)](#)

[Research Summary: Artificial Intelligence and the Privacy Paradox of Opportunity, Big Data and The Digital Universe \(3/21/21\)](#)

[Mapping India's AI Potential \(3/21/21\)](#)

[Hiring by Algorithm: Legal Issues Presented by the Use of Artificial Intelligence in Sourcing and Selection \(3/17/21\)](#)

[The Algorithmic Auditing Trap \(3/17/21\)](#)

[Should AI Models Be Explainable? That depends. \(3/16/21\)](#)

[How AI is Changing Our Understanding of Language \(3/16/21\)](#)

[Who Is Governing AI Matters Just as Much as How It's Designed \(3/15/21\)](#)

[10 Takeaways from the State of AI Ethics in Canada & Spain \(3/15/21\)](#)

[How Can We Better Regulate AI Companies? \(3/15/21\)](#)

[Myanmar: Facial Recognition System Threatens Rights \(3/12/21\)](#)

[Auditing employment algorithms for discrimination \(3/12/21\)](#)

[Fairness On The Ground: Applying Algorithmic Fairness Approaches To Production Systems \(3/11/21\)](#)

[Assessing Regulatory Fairness Through Machine Learning \(3/10/21\)](#)

[Rome] [The Ethics of AI to Ensure Food Security and Development](#)

(3/8/21)

[Britain] [Malgorithm: How Instagram's algorithm publishes](#)

[misinformation and hate to millions during a pandemic.](#) (3/8/21)

[If It's Free, You're the Product: The New Normal in a Surveillance](#)

[Economy](#) (3/7/21)

[Corsight AI addresses face biometrics ethics in Asia Pacific expansion](#)

(3/5/21)

[Where are the women? Mapping the gender job gap in AI Policy](#)

[Briefing – Full Report \(Mar 2021\)](#)

[Stanford HAI: AI Index Report 2021 \(Mar 2021\)](#)

[KPMG: Trust in Artificial Intelligence: A five country study \(March 2021\)](#)

[WEF: A 5-step guide to scale responsible AI \(3/1/21\)](#)

[RESPONSIBLE AI #AIFORALL: Approach Document for India Part 1 –](#)

[Principles for Responsible AI \(Feb 2021\)](#)

[WEF: Responsible Use of Technology: The Microsoft Case Study](#)

(2/25/21)

[WEF: 4 lessons on designing responsible, ethical tech: Microsoft case](#)

[study \(2/25/21\)](#)

[A Better Measuring Stick: Algorithmic Approach to Pain Diagnosis](#)

[Could Eliminate Racial Bias \(2/24/21\)](#)

We need to talk about Artificial Intelligence (2/22/21)

The Artificiality of AI – Why are We Letting Machines Manage Employees? (2/22/21)

Research Summary: The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms (2/21/21)

Why Organizations Should Care About Responsible AI & Digital Ethics (2/19/21)

Black voices bring much needed context to our data-driven society (2/18/21)

Privacy enhancing technologies for trustworthy use of data (2/9/21)

The Physician's Conundrum: Assigning Moral Responsibility for Medical Artificial Intelligence and Machine Learning (2/8/21)

Legislation Related to Artificial Intelligence

TRAINING AI SYSTEMS FOR USE IN DANGEROUS SITUATIONS (2/4/21)

As the FDA clears a flood of AI tools, missing data raise troubling questions on safety and fairness (2/3/21)

Governance by Algorithms (Research Summary) (2/1/21)

[Germany] 10 DIFFICULT TERMS TO TRANSLATE REGARDING RACE (Feb 2021)

MAEI The State of AI Ethics Report (Jan 2021)

Addressing Ethical Dilemmas in AI: Listening to Engineers (Jan 2021)

The CREO AI Creativity Report 2021 (Jan 2021)

NYC Summary of Agency Compliance Reporting (Jan 2021)

What algorithm auditing startups need to succeed (1/30/21)

EU Guidelines on Facial Recognition (1/28/21)

FDA Announces Action Plan for Oversight of AI/ML in Medical Devices (1/28/21)

Mapping AI in the Global South (1/26/21)

Beyond Engagement: Aligning Algorithmic Recommendations With Prosocial Goals (1/21/21)

The building blocks of Microsoft's responsible AI program (1/19/21)

Seven Legal Questions for Data Scientists (1/19/21)

To achieve Responsible AI, close the "believability gap" (1/18/21)

Civil society calls for AI red lines in the European Union's Artificial Intelligence proposal (1/12/21)

China's social credit system was due by 2020 but is far from ready (1/12/21)

California Company Settles FTC Allegations It Deceived Consumers about use of Facial Recognition in Photo Storage App (1/11/21)

## How to Build a Likable Chatbot (1/11/21)

In Poland, a law made loan algorithms transparent. Implementation is nonexistent. (1/6/21)

## Getting Specific About AI Risks (1/6/21)

In Focus: Facial Recognition Tech Stories and Rights Harms Around the World (Jan 2021)

AI Governance in Japan Ver. 1.0 (Jan 2021)

## **Popular Press**

Promoting Trustworthy AI in Government (7/8/21)

OpenAI warns AI behind GitHub's Copilot may be susceptible to bias (7/8/21)

We tested AI interview tools. Here's what we found. (7/7/21)

How TikTok's hate speech detection tool set off a debate about racial bias on the app (7/7/21)

THE MOST "WOKE" COMPANY COULD CONTRIBUTE MOST TO ONLINE BIAS (7/5/21)

AI legislation must address bias in algorithmic decision-making systems (7/5/21)

COULD AI KEEP PEOPLE 'ALIVE' AFTER DEATH? (7/3/21)

## Pandemic Wave of Automation May Be Bad News for Workers

(7/3/21)

What happened when a ‘wildly irrational’ algorithm made crucial healthcare decisions (7/2/21)

Study finds that few major AI research papers consider negative impacts (7/1/21)

The importance of having accountability in AI ethics (7/1/21)

Facial recognition tech has been widely used across the US government for years, a new report shows (6/30/21)

Using A.I. to Find Bias in A.I. (6/30/21)

A Hippocratic Oath for your AI doctor (6/29/21)

Fired by Bot at Amazon: ‘It’s You Against the Machine’ (6/28/21)

DeepMind AGI paper adds urgency to ethical AI (6/26/21)

NIST proposes ways to identify and address AI bias (6/25/21)

The World Needs Deepfake Experts to Stem This Chaos (6/24/21)

IBM explores AI tools to spot, cut bias in online ad targeting (6/24/21)

Combating racial bias in AI (6/23/21)

LinkedIn’s job-matching AI was biased. The company’s solution? More AI. (6/23/21)

[Google & DeepMind Researchers Revamp ImageNet \(6/23/21\)](#)

[Navigating a surprising pandemic side effect: AI whiplash \(6/21/21\)](#)

[Europe's privacy regulators call for a ban on facial recognition in publicly accessible spaces \(6/21/21\)](#)

['Nobody is catching it': Algorithms used in health care nationwide are rife with bias \(6/21/21\)](#)

[A popular algorithm to predict sepsis misses most cases and sends frequent false alarms, study finds \(6/21/21\)](#)

[Google searches for new measure of skin tones to curb bias in products \(6/18/21\)](#)

[Facial Recognition Failures Are Locking People Out of Unemployment Systems \(6/18/21\)](#)

[Hey computer, am I okay? \(6/18/21\)](#)

[The Efforts to Make Text-Based AI Less Racist and Terrible \(6/17/21\)](#)

[Ethical AI will not see broad adoption by 2030, study suggests \(6/17/21\)](#)

[Amazon Alexa head scientist on developing trustworthy AI systems \(6/16/21\)](#)

[Facebook's latest AI doesn't just detect deep fakes, it knows where they came from \(6/16/21\)](#)

[Want a job? Employers say: Talk to the computer \(6/14/21\)](#)

How and Why Enterprises Must Tackle Ethical AI (6/14/21)

Fact-Checks, Info Hubs, and Shadow-Bans: A Landscape Review of Misinformation Interventions (6/14/21)

Google's Privacy Backpedal Shows Why It's So Hard Not to Be Evil (6/14/21)

The secrets behind the plastic spoon: a 'perfect' design with terrible consequences (6/14/21)

How To Make Sure That Diversity In AI Works (6/14/21)

Human Cognitive Bias And Its Role In AI (6/14/21)

How to mitigate bias in AI (6/13/21)

Rules around facial recognition and policing remain blurry (6/12/21)

These creepy fake humans herald a new age in AI (6/11/21)

Biden administration forms new AI task force (6/10/21)

[China] Across China, AI 'city brains' are changing how the government runs (6/10/21)

TikTok changed the shape of some people's faces without asking (6/10/21)

EleutherAI claims new NLP model approaches GPT-3-level performance (6/9/21)

10 steps to educate your company on AI fairness (6/9/21)

Facial Verification Won't Fight Fraud (6/9/21)

First U.S. Artificial Intelligence Czar Seeks 'Responsible Use' of AI Tools (6/8/21)

Bias and discrimination in AI: whose responsibility is it to tackle them? (6/8/21)

Instagram pushes back against 'The Algorithm' – because it tracks every action you take on the app using 'a variety of algorithms' (6/8/21)

Investors call for ethical approach to facial recognition technology (6/8/21)

Skin in the game: Video chat apps tout 'inclusive' AI features (6/7/21)

That AI scanning your X-ray for signs of COVID-19 may just be looking at your age (6/7/21)

As AI develops, so does the debate over profits and ethics (6/4/21)

Federal facial recognition ban will be reintroduced 'soon,' Sen. Jeff Merkley says (6/4/21)

China's response to COVID showed the world how to make the most of A.I. (6/4/21)

TikTok just gave itself permission to collect biometric data on US users, including 'faceprints and voiceprints' (6/3/21)

'Care bots' are on the rise and replacing human caregivers (6/3/21)

Don't End Up on This Artificial Intelligence Hall of Shame (6/3/21)

Dark Patterns that Mislead Consumers Are All Over the Internet  
(6/3/21)

How Amazon became an engine for anti-vaccine conspiracy theories  
(6/3/21)

Amazon's Cost Saving Routing Algorithm Makes Drivers Walk Into Traffic (6/2/21)

McDonald's is testing automated drive-thru ordering at 10 Chicago restaurants (6/2/21)

Machine learning is booming in medicine. It's also facing a credibility crisis (6/2/21)

Senate Democrats Urge Google To Investigate Racial Bias In Its Tools And The Company (6/2/21)

FACIAL RECOGNITION TECHNOLOGY: Federal Law Enforcement Agencies Should Better Assess Privacy and Other Risks (June 2021)

A rogue killer drone 'hunted down' a human target without being instructed to, UN report says (5/30/21)

Skin in the frame: black photographers welcome Google initiative  
(5/28/21)

How a largely untested AI algorithm crept into hundreds of hospitals  
(5/28/21)

Anthropic is the new AI research outfit from OpenAI's Dario Amodei, and it has \$124M to burn (5/28/21)

[Spain] Workers vs Algorithms (5/27/21)

The ethical use of data in artificial intelligence (5/27/21)

This \$5 billion insurance company likes to talk up its AI. Now it's in a mess over it (5/27/21)

Scared, human? Emotion-detection AI meets eye-tracking technology (5/27/21)

[EU] Clearview AI hit with sweeping legal complaints over controversial face scraping in Europe (5/27/21)

[EU] The Grand Normalization of Mass Surveillance: ECtHR Grand Chamber Judgments in Big Brother Watch and Centrum för rättvisa (5/26/21)

Google Strikes Deal With Hospital Chain to Develop Healthcare Algorithms (5/26/21)

AI emotion-detection software tested on Uyghurs (5/26/21)

Report finds startling disinterest in ethical, responsible use of AI among business leaders (5/25/21)

65% of execs can't explain how their AI models make decisions, survey finds (5/25/21)

How 'blind learning' could solve A.I.'s 'Little Shop of Horrors' dilemma (5/25/21)

'AI' is being used to profile people from their head vibrations – but is there enough evidence to support it? (5/24/21)

Facial Recognition Is Racist. Why Aren't More Cities Banning It?  
(5/24/21)

AI Can Write Disinformation Now—and Dupe Human Readers  
(5/24/21)

AI didn't invent privacy abuse - how the history of data privacy informs our future (5/24/21)

Citizen's dystopian new feature is mass surveillance disguised as public safety (5/22/21)

Facebook AI cuts by more than half the error rate of unsupervised speech recognition (5/21/21)

AI's Future Doesn't Have to Be Dystopian (5/20/21)

Google's New Dermatology App Wasn't Designed for People With Darker Skin (5/20/21)

Publicis Media updates ad tech evaluation process to avoid cultural stereotypes, but getting information about algorithms 'is hard'  
(5/20/21)

Statistical bias in context - AI didn't invent quantitative methods of bias (5/20/21)

The race to understand the exhilarating, dangerous world of language AI (5/20/21)

[Spotify, don't spy: global coalition of 180+ musicians and human rights groups take a stand against speech-recognition technology \(5/19/21\)](#)

[Twitter finds its AI tends to crop out Black people, men from photos \(5/19/21\)](#)

[Innovative free course empowers citizens to advocate for ethical AI \(5/19/21\)](#)

[The disinformation threat from text-generating AI \(5/19/21\)](#)

[\[France\] Environmentalists do not cut video surveillance but slow down its development \(5/18/21\)](#)

[Why companies should carefully read government A.I. plans \(5/18/21\)](#)

[Is it time for an algorithm bill of rights? These analysts think so \(5/18/21\)](#)

[Amazon extends moratorium on police use of facial recognition software \(5/18/21\)](#)

[Amazon extends ban on police use of its facial recognition technology indefinitely \(5/18/21\)](#)

[Evolving to a more equitable AI \(5/18/21\)](#)

[Ethical AI Is Our Responsibility \(5/17/21\)](#)

[The Impact of Dark Patterns on Communities of Color \(5/17/21\)](#)

[AIs are getting smarter, fast. That's creating tricky questions that we can't answer \(5/17/21\)](#)

The Chief AI Ethics Officer: A Champion or a PR Stunt? (5/16/21)

Police departments adopting facial recognition tech amid allegations of wrongful arrests (5/16/21)

Critics raise alarm over Big Tech's most powerful tools (5/16/21)

In charts: facial recognition technology – and how much do we trust it? (5/16/21)

[New Zealand] Jacinda Ardern calls for 'ethical algorithms' to help stop online radicalisation (5/14/21)

We need to design distrust into AI systems to make them safer (5/13/21)

[UK] AI and robots to assess patients in NHS plan to tackle record waiting lists (5/13/21)

[China] Chinese Users Do Care About Online Privacy (5/12/21)

Instagram Censored Posts About One Of Islam's Holiest Mosques, Drawing Employee Ire (5/12/21)

Trustworthy AI versus ethical AI - what's the difference, and why does it matter? (5/11/21)

Expert.ai adds emotion, style detection tools to natural language API (5/10/21)

Blind people, advocates slam company claiming to make websites ADA compliant (5/9/21)

[Israel] Sheikh Jarrah: Activists raise concerns over deleted social media content (5/7/21)

Deepfake detectors and datasets exhibit racial and gender bias, USC study shows (5/6/21)

California's "Equity" Algorithm Could Leave 2 Million Struggling Californians Without Additional Vaccine Supply (5/6/21)

Proper data hygiene critical as enterprises focus on AI governance (5/6/21)

[Denmark] Ready for face recognition at the restaurant? (5/5/21)

AI Bias Problem Needs More Academic Rigor, Less Hype (5/4/21)

Anyone can use this powerful facial-recognition tool – and that's a problem (5/4/21)

AI bias is an ongoing problem, but there's hope for a minimally biased future (5/4/21)

Fighting algorithmic bias in artificial intelligence (5/4/21)

YOUR CAR IS SPYING ON YOU, AND A CBP CONTRACT SHOWS THE RISKS (5/3/21)

Shhhh, they're listening: Inside the coming voice-profiling revolution (5/3/21)

Appen combats skewed AI data to ensure end-users have the same experience (5/3/21)

## Using AI to root out unconscious bias (5/1/21)

Ethics of AI: Benefits and risks of artificial intelligence (4/30/21)

[Russia] 'Deepfake' that supposedly fooled European politicians was just a look-alike, say pranksters (4/30/21)

[Portugal] "It is worrying for democracy." TAP uses "heartless" algorithm to fire (4/30/21)

King County government must turn its back on facial recognition technology (4/30/21)

Credit Card Ads Were Targeted by Age, Violating Facebook's Anti-Discrimination Policy (4/29/21)

Black Man Accuses AI of Being Racist After Passport Photo Is Rejected in Viral Video (4/28/21)

Kind Environments for technology organisations? [part two] (4/28/21)

N.Y.P.D. Robot Dog's Run Is Cut Short After Fierce Backlash (4/28/21)

Landmark AI legislation could tackle algorithmic bias (4/28/21)

IBM and Microsoft Have Integrated AI Ethical Standards into Their Operations, So Can You (4/28/21)

Congress drags algorithms out of the shadows (4/27/21)

Antiracism in AI: How to Build Bias Checkpoints Into Your Development and Delivery Process (4/27/21)

To Fight Social Media Disinformation, Look to the Algorithms(4/27/21)Ex-Google employee: Big Tech's biz model is 'a society that is addicted, outraged, polarized' (4/27/21)Artificial Intelligence Is Misreading Human Emotion (4/27/21)[UAE] Sharjah Police drones use face-recognition technology to identify wanted criminals (4/26/21)Intersectionality, Inequity And Imminence Of AI: A Human Perspective(4/26/21)AI at work isn't always intelligent (4/26/21)Despite acknowledged promise: Fear, uncertainty and doubt surround AI adoption (4/26/21)'Make Algorithmic Audits As Ubiquitous As Seatbelts'—Why Tech Needs Outside Help To Serve Humanity (4/26/21)AI that stops you getting mad at your customers? Are you crazy?(4/24/21)Big tech's algorithms are not color blind (4/23/21)Stop talking about AI ethics. It's time to talk about power. (4/23/21)Algorithmic Nudges Don't Have to Be Unethical (4/22/21)Stanford study questions how medical AI devices are evaluated (4/22/21)

[Shadow Bans, Dopamine Hits, and Viral Videos, All in the Life of TikTok](#)

[Creators \(4/22/21\)](#)

[The global race to regulate AI \(4/22/21\)](#)

[To ensure inclusivity, the Biden administration must double down on AI development initiatives \(4/22/21\)](#)

[Microsoft and Big Tech Can't Distance Themselves From the Police Violence They Fuel \(4/21/21\)](#)

[Europe throws down gauntlet on AI with new rulebook \(4/21/21\)](#)

[In scramble to respond to Covid-19, hospitals turned to models with high risk of bias \(4/21/21\)](#)

[6 key battles ahead for Europe's AI law \(4/21/21\)](#)

[Europe's Proposed Limits on AI Would Have Global Consequences \(4/21/21\)](#)

[Google Ethical AI Group's Turmoil Began Long Before Public Unraveling \(4/21/21\)](#)

[FTC issues stern warning: Biased AI may break the law \(4/20/21\)](#)

['Detoxified' language models might marginalize minorities, says study \(4/20/21\)](#)

[AI vs. Maya Angelou: Experimental Evidence That People Cannot Differentiate AI-Generated From Human-Written Poetry \(4/19/21\)](#)

Nextdoor will alert users if it thinks they're about to post something racist (4/19/21)

She's taking Jeff Bezos to task (4/19/21)

US banks deploy AI to monitor customers, workers amid tech backlash (4/19/21)

Hackers Used to Be Humans. Soon, AIs Will Hack Humanity (4/19/21)

Google translation AI botches legal terms 'enjoin,' 'garnish' -research (4/19/21)

Here's why we should never trust AI to identify our emotions (4/18/21)

Data and computer scientists, ecologists, pathologists and legal scholars study AI's biases (4/16/21)

[Sweden] World Moving Towards a “Devastating Marriage” of Artificial Intelligence & Weapons of War (4/16/21)

[Italy] Facial recognition: Sari Real Time does not comply with the privacy policy (4/16/21)

[Netherlands] Rotterdam Court of Audit: risk of biased outcomes due to the use of algorithms (4/15/21)

Amazon and Microsoft team up to defend against facial recognition lawsuits (4/15/21)

MEPs call for European AI rules to ban biometric surveillance in public (4/15/21)

[AI's social justice problem: It's amplifying human bias \(4/15/21\)](#)

[Coded Bias: An Insightful Look At AI, Algorithms And Their Risks To Society \(4/15/21\)](#)

[AI is increasingly being used to identify emotions – here's what's at stake \(4/15/21\)](#)

[The Slow Violence of Emerging Technologies \[part one\] \(4/14/21\)](#)

[The new lawsuit that shows facial recognition is officially a civil rights issue \(4/14/21\)](#)

[The Computer Got it Wrong: Why We're Taking the Detroit Police to Court Over a Faulty Face Recognition 'Match' \(4/13/21\)](#)

[Why AI Is Failing for Enterprises: Predetermination Bias \(4/12/21\)](#)

[Revealed: the Facebook loophole that lets world leaders deceive and harass their citizens \(4/12/21\)](#)

[Society dictated your face, according to biometrics research and a patent \(4/12/21\)](#)

[Hitting the Books: How biased AI can hurt users or boost a business's bottom line \(4/10/21\)](#)

[Black women, AI, and overcoming historical patterns of abuse \(4/10/21\)](#)

[\[Sweden\] CKP's test use of a program intended for face recognition has been reported to the Data Protection Ombudsman \(4/9/21\)](#)

[South Asia] From Lahore to Lucknow, crimes against women spur more surveillance (4/9/21)

Artificial intelligence isn't helping you hire the best person for the job (4/9/21)

Shedding light on fairness in AI with a new data set (4/8/21)

A \$2 Billion Government Surveillance Lab Created Tech That Guesses Your Name By Simply Looking At Your Face (4/8/21)

Students of color are getting flagged to their teachers because testing software can't see them (4/8/21)

In all police stations in France, facial recognition is used (4/7/21)

Study suggests that AI model selection might introduce bias (4/7/21)

[EU] Europeans can't talk about racist AI systems. They lack the words. (4/6/21)

Twitter nukes AI-generated twits who backed Amazon and pushed anti-union rhetoric (4/6/21)

Discover the stupidity of AI emotion recognition with this little browser game (4/6/21)

Surveillance Nation (4/6/21)

Your Local Police Department Might Have Used This Facial Recognition Tool To Surveil You. Find Out Here. (4/6/21)

Can A.I. help Facebook cure its disinformation problem? (4/6/21)

[India] Modi govt now plans a ‘touchless’ vaccination process, with Aadhaar-based facial recognition (4/6/21)

[France] Suresnes wants to detect suspicious behavior (4/6/21)

Your 'smart home' is watching – and possibly sharing your data with the police (4/5/21)

Government audit of AI with ties to white supremacy finds no AI (4/5/21)

Can AI read your emotions? Try it for yourself (4/5/21)

Scientists create online games to show risks of AI emotion recognition (4/4/21)

Why Silicon Valley's most astute critics are all women (4/3/21)

Hitting the Books: The bias behind AI assistants' failure to understand accents (4/3/21)

Here's how enterprises say they're deploying AI responsibly (4/2/21)

[EU] Seeing stones: pandemic reveals Palantir's troubling reach in Europe (4/2/21)

A South Korean Chatbot Shows Just How Sloppy Tech Companies Can Be With User Data (4/2/21)

These Companies Track Millions Of Cars–Immigration And Border Police Have Been Grabbing Their Data (4/1/21)

Study finds that even the best speech recognition systems exhibit bias

(4/1/21)

[UK] Technology Managing People – the legal implications (March 2021)

AI experts warn Facebook's anti-bias tool is 'completely insufficient'  
(3/31/21)

The Foundations of AI Are Riddled With Errors (3/31/21)

[UK] We've won our lawsuit over Matt Hancock's £23m NHS data deal with Palantir (3/30/21)

Utah Gave \$20 Million Contract to AI Surveillance Firm That Didn't Have AI (3/30/21)

MIT study finds labelling errors in datasets used to test AI (3/29/21)

Automated translation is hopelessly sexist, but don't blame the algorithm or the training data (3/29/21)

Americans think AI has the most potential to cause harm over next decade (3/29/21)

FBI warns of the rise of 'deepfakes' in coming months and explains how to spot them easily (3/29/21)

[Australia] 'Dangerous future' in NDIS tech experimentation (3/28/21)

[Brazil] Anti-racism, technology and artificial intelligence in Brazil: a chat at Mozfest, by Mozilla (3/28/21)

AI: Ghost workers demand to be seen and heard (3/28/21)

Why Representation Matters When Building AI (3/28/21)

[Spain] Aragón will implement an algorithm for the detection of suicide risk in his clinical history (3/26/21)

If Mark Zuckerberg won't fix Facebook's algorithms problem, who will? (3/26/21)

Scotland launches AI strategy with a focus on ethics and inclusion (3/26/21)

[Germany] Allocation of vaccination appointments: Are older people disadvantaged? (3/25/21)

AI at work: Staff 'hired and fired by algorithm' (3/25/21)

[Netherlands] Roeland (22) was kicked from TikTok because the app thinks he is under 13 (3/23/21)

Amazon Delivery Drivers Forced to Sign 'Biometric Consent' Form or Lose Job (3/23/21)

After Clearview, more bad actors in A.I. facial recognition might show up (3/23/21)

Major flaws found in machine learning for COVID-19 diagnosis (3/23/21)

How synthetic data could save AI (3/20/21)

Uber under pressure over facial recognition checks for drivers

(3/19/21)

[UK] Uber drivers claim they were fired after company's identification software failed to recognise their faces (3/19/21)

What happens when your massive text-generating neural net starts spitting out people's phone numbers? If you're OpenAI, you create a filter (3/18/21)

The Pandemic Could Accelerate Job Automation—and Inequality

(3/18/21)

A New York Lawmaker Wants to Ban Police Use of Armed Robots

(3/18/21)

AI can be unintentionally biased: Data cleaning and awareness can help prevent the problem (3/18/21)

'Anonymized' X-ray datasets can reveal patient identities (3/17/21)

California bans 'dark patterns' that trick users into giving away their personal data (3/16/21)

Europe's artificial intelligence blindspot: Race (3/16/21)

[Netherlands] Tens of thousands of people may falsely enter police face database (3/16/21)

This is the EU's chance to stop racism in artificial intelligence

(3/16/21)

Diversity And Inclusion In AI (3/16/21)

Researchers blurred faces that launched a thousand algorithms

(3/15/21)

Who Is Making Sure the A.I. Machines Aren't Racist? (3/15/21)

Preventing bias in AI is hard. Bug bounties could point the way forward

(3/15/21)

Tenant screening software faces national reckoning (3/14/21)

Getting to trustworthy AI (3/14/21)

Google might ask questions about AI ethics, but it doesn't want answers (3/13/21)

Dutch court rulings break new ground on gig worker data rights

(3/12/21)

Auditing employment algorithms for discrimination (3/12/21)

[Spain] The 'riders' law will oblige companies to inform unions about algorithms that affect working conditions (3/11/21)

[France] In the Official Journal, "smart" cameras to measure the rate of mask wearing in transport (3/11/21)

GAO's emerging tech shop developing framework to test reliability of AI algorithms (3/11/21)

How Facebook got addicted to spreading misinformation (3/11/21)

[UK] Hundreds of sewage leaks detected thanks to AI (3/11/21)

[China] [Honest passengers first! Beijing subway to pilot credit-based fast entry system \(3/10/21\)](#)

[Biased AI can be bad for your health – here's how to promote algorithmic fairness \(3/9/21\)](#)

[Artificial intelligence tool helps detect unconscious racial, cultural bias in the workplace \(3/8/21\)](#)

[Algorithms are increasingly treating workers like robots. Canada needs policy to protect them \(3/8/21\)](#)

[Underpaid Workers Are Being Forced to Train Biased AI on Mechanical Turk \(3/8/21\)](#)

[Inside the fall of Watson Health: How IBM's audacious plan to 'change the face of health care' with AI fell apart \(3/8/21\)](#)

[Building AI for the Global South \(3/7/21\)](#)

[I asked an AI to tell me how beautiful I am \(3/5/21\)](#)

['Facebook has a blind spot': why Spanish-language misinformation is flourishing \(3/3/21\)](#)

[Looking For An AI Ethicist? Good Luck \(3/2/21\)](#)

[Russia] [In Moscow, Big Brother Is Watching and Recognizing Protesters \(3/1/21\)](#)

[Does your company need a Chief AI Ethics Officer, an AI Ethicist, AI Ethics Council, or all three? \(Mar. 2021\)](#)

Taking on the tech giants: the lawyer fighting the power of algorithmic systems (2/28/21)

5 steps to creating a responsible AI Center of Excellence (2/28/21)

Virginia Approves Limits on Police Use of Facial Recognition (2/26/21)

An AI is training counselors to deal with teens in crisis (2/26/21)

TikTok agrees legal payout over facial recognition (2/26/21)

'This is bigger than just Timnit': How Google tried to silence a critic and ignited a movement (2/26/21)

Why Google's AI ethics blunders are a PR nightmare (2/25/21)

Can Auditing Eliminate Bias from Algorithms? (2/23/21)

Whistleblowers: Software Bug Keeping Hundreds Of Inmates In Arizona Prisons Beyond Release Dates (2/22/21)

HUMANS ARE TRYING TO TAKE BIAS OUT OF FACIAL RECOGNITION PROGRAMS. IT'S NOT WORKING-YET. (2/22/21)

Can We Engineer Ethical AI? (2/21/21)

The Shoddy Science Behind Emotional Recognition Tech (2/19/21)

How NSF and Amazon Are Collectively Tackling Artificial Intelligence-Based Bias (2/18/21)

DeepMind researchers say AI poses a threat to people who identify as queer (2/18/21)

Studies find bias in AI models that recommend treatments and diagnose diseases (2/18/21)

Why Tech Companies Are Limiting Police Use of Facial Recognition (2/18/21)

Biden Must Halt Face Recognition Technology to Advance Racial Equity (2/17/21)

This AI reads children's emotions as they learn (2/17/21)

Can Computer Algorithms Learn to Fight Wars Ethically? (2/17/21)

Spain: Ethnic origin, gender or clothing: Renfe's controversial system to monitor its travelers (2/17/21)

Germany: Objective or biased (2/16/21)

Study shows that federated learning can lead to reduced carbon emissions (2/16/21)

Development of an AI Technology able to 'Read Emotions' across entire cities to Stop Crime Before it Happen (2/16/21)

How computers see us: Doctoral student working to curb discrimination by artificial intelligence (2/15/21)

France: The CNIL dampens the appetite of Big Brother Bercy (2/15/21)

Washington state lawmakers seek to ban government from using discriminatory AI tech (2/13/21)

Amazon's driver monitoring app is an invasive nightmare (2/13/21)

Minneapolis poised to ban facial recognition for police use (2/12/21)

Facebook's next big AI project is training its machines on users' public videos (3/12/21)

Techno-utopianism in the workplace and the threat of excessive automation (2/12/21)

Amazon uses an app called Mentor to track and discipline delivery drivers (2/12/21)

Are AI investors shorting Black lives? (2/11/21)

Russia: "Face control": Russian police go digital against protesters (2/11/21)

Why Is Facebook Rejecting These Fashion Ads? (2/11/21)

Fighting AI bias needs to be a key part of Biden's civil rights agenda (2/11/21)

COUNCIL OF EUROPE PROPOSES BAN ON CERTAIN FACIAL RECOGNITION APPLICATIONS

Despite Scanning Millions of Faces, Feds Caught Zero Imposters at Airports Last Year (2/10/21)

EU's top privacy regulator urges ban on surveillance-based ad targeting (2/10/21)

[UK] [BAILII grants Oxford University unprecedented access to case data for AI analysis in historic agreement](#) (2/10/21)

[The computers rejecting your job application](#) (2/8/21)

[How Data Can Drive Inequality](#) (2/8/21)

[Diversity in AI is awful](#) (2/8/21)

[Where Do Vaccine Doses Go, and Who Gets Them? The Algorithms Decide](#) (2/7/21)

[Rethinking Gaming: The Ethical Work of Optimization in Web Search Engines](#) (2/7/21)

[‘Orwellian’ AI lie detector project challenged in EU court](#) (2/5/21)

[UPDATE 1-High-tech lie detector used at Europe borders face scrutiny](#) (2/5/21)

[New AI Can Detect Emotion With Radio Waves](#) (2/4/21)

[Your iPhone's Adult Content Filter Blocks Anything 'Asian'](#) (2/4/21)

[There Are Spying Eyes Everywhere—and Now They Share a Brain](#) (2/4/21)

[AI brain drain to Google and pals threatens public sector's ability to moderate machine-learning bias](#) (2/4/21)

[The Role Of Bias In Artificial Intelligence](#) (2/4/21)

[The AI industry is built on geographic and social inequality, research shows](#) (2/4/21)

## 21 States Are Now Vetting Unemployment Claims With a ‘Risky’ Facial Recognition System (2/3/21)

U.S. technology company Clearview AI violated Canadian privacy law: report (2/3/21)

“Even if you can do it, should you?” Researchers talk combating bias in artificial intelligence (2/3/21)

[Netherlands] Can algorithms look into a child's future? In Hjørring and Silkeborg, experiments were made on vulnerable children (2/2/21)

[Singapore] Application installed on students' devices does not track personal information: MOE (2/1/21)

Confidence, uncertainty, and trust in AI affect how humans make decisions (2/1/21)

Embracing Diversity Using AI [Infographic] (2/1/21)

Inside a Hot-Button Research Paper: Dr. Emily M. Bender Talks Large Language Models and the Future of AI Ethics (2/1/21)

Here’s a Way to Learn if Facial Recognition Systems Used Your Photos (1/31/21)

Forget user experience. AI must focus on ‘citizen experience’ (1/31/21)

Why some companies are thinking twice about using artificial intelligence (1/31/21)

The Coup We Are Not Talking About (1/29/21)

New Spotify Patent Involves Monitoring Users' Speech to Recommend Music (1/28/21)

Spotify Secures Horrifying Patent to Monitor Users' Speech (1/28/21)

Police Say They Can Use Facial Recognition, Despite Bans (1/28/21)

[Greece] Flush with EU funds, Greek police to introduce live face recognition before the summer (1/28/21)

University will stop using controversial remote-testing software following student outcry (1/28/21)

Bumble's algorithm will now report you for body-shaming (1/28/21)

It's time to use all of Twitter's archives to teach AI about bias (1/27/21)

Independent auditors are struggling to hold AI companies accountable (1/26/21)

SCIENTISTS TRYING TO BUILD MEDICAL AI THAT'S LESS RACIST THAN DOCTORS (1/26/21)

NJ Transit will test AI-powered face-mask detection (1/25/21)

New Algorithms Could Reduce Racial Disparities in Health Care (1/25/21)

Governance: Companies mature in their use of AI know that it needs guardrails (1/25/21)

[France] Suspended feminist accounts: for once, Twitter admits a real "error" (1/25/21)

[Netherlands] The Dutch Government's Benefits Scandal Is Rooted in Stigma Against Welfare Recipients (1/23/21)

AI could make healthcare fairer—by helping us believe what patients say (1/22/21)

Center for Applied Data Ethics suggests treating AI like a bureaucracy (1/22/21)

[France] In Bercy, a cell of computer scientists to help the State regulate GAFA (1/22/21)

[India] Why Lucknow Police wanting to use AI to ‘read’ distressed expressions of women won’t work (1/21/21)

Democrats urge tech giants to change algorithms that facilitate spread of extremist content (1/21/21)

‘For Some Reason I’m Covered in Blood’: GPT-3 Contains Disturbing Bias Against Muslims (1/21/21)

India’s biometric ID system is eroding the rights of pregnant women (1/21/21)

This App Claims It Can Detect ‘Trustworthiness.’ It Can’t (1/19/21)

FDA publishes Action Plan to regulate AI and ML based products (1/18/21)

The FTC Forced a Misbehaving A.I. Company to Delete Its Algorithm (1/18/21)

[Social-Media Algorithms Rule How We See the World. Good Luck](#)

[Trying to Stop Them. \(1/17/21\)](#)

[France] [Twitter and Facebook send "dykes" and "queers" to the closet](#)  
[\(1/16/21\)](#)

[Chatbot Gone Awry Starts Conversations About AI Ethics in South Korea \(1/16/21\)](#)

[Using AI to Make Hiring Decisions? Prepare for EEOC Scrutiny](#)  
[\(1/15/21\)](#)

[Facebook uses AI to predict if COVID-19 patients will need more care](#)  
[\(1/15/21\)](#)

[An Algorithm Is Helping a Community Detect Lead Pipes \(1/14/21\)](#)

[Democrats push for AI bias testing \(1/14/21\)](#)

[Why Ethics Matter For Social Media, Silicon Valley And Every Tech Industry Leader \(1/14/21\)](#)

[How explainable artificial intelligence can help humans innovate](#)  
[\(1/13/21\)](#)

[Salesforce researchers release framework to test NLP model robustness \(1/13/21\)](#)

[Flo gets FTC slap for sharing user data when it promised privacy](#)  
[\(1/13/21\)](#)

[Italy] [The Viminal-Guarantor of privacy clash on facial recognition in real time \(1/13/21\)](#)

Huawei patent mentions use of Uighur-spotting tech (1/12/21)

The Robot Made Me Do It: Human–Robot Interaction and Risk-Taking Behavior (Research Summary) (1/12/21)

The Ethics of Emotion in AI Systems (Research Summary) (1/12/21)

Job Screening Service Halts Facial Analysis of Applicants (1/12/21)

“I Don’t Want Someone to Watch Me While I’m Working”: Gendered Views of Facial Recognition Technology in Workplace Surveillance (Research Summary) (1/12/21)

Facial Recognition Technology Isn’t Good Just Because It’s Used to Arrest Neo-Nazis (1/12/21)

Facial Recognition Technology Has A Bias Problem (1/11/21)

Listening to Black Women: The Innovation Tech Can't Figure Out (1/11/21)

[Netherlands] SyRI coalition to Senate: 'Super SyRI' blueprint for more benefits affairs (1/11/21)

Outlandish Stanford facial recognition study claims there are links between facial features and political orientation (1/11/21)

[S. Korea] CEO says controversial AI chatbot ‘Luda’ will socialize in time (1/11/21)

Algorithms and the coronavirus pandemic (1/9/21)

New York City Proposes Regulating Algorithms Used in Hiring (1/8/21)

[ExamSoft's proctoring software has a face-detection problem \(1/5/21\)](#)

[Italy] [Court Rules Deliveroo Used 'Discriminatory' Algorithm \(1/5/21\)](#)

[Medicine's machine learning problem \(1/4/21\)](#)

[Singapore Police May Use Contact Tracing Data for Investigations \(1/4/21\)](#)

[2020 in Review: 10 AI Failures \(1/1/21\)](#)

## 2020

### Peer-reviewed

[Racial Bias in Pulse Oximetry Measurement \(12/17/20\)](#)

[Switzerland] [Smart Criminal Justice \(12/10/20\)](#)

[Invited Talk: You Can't Escape Hyperparameters and Latent Variables: Machine Learning as a Software Engineering Enterprise \(12/7/20\)](#)

[FAIROD: Fairness-aware Outlier Detection \(5/20/21\)](#)

[Civil Unrest on Twitter \(CUT\): A Dataset of Tweets to Support Research on Civil Unrest \(Nov 2020\)](#)

[Protecting consumers from collusive prices due to AI \(11/27/20\)](#)

[The ethical questions that haunt facial-recognition research \(11/18/20\)](#)

[How to Measure Gender Bias in Machine Translation: Optimal Translators, Multiple Reference Points \(11/12/20\)](#)

We need to talk about deception in social robotics! (11/11/20)

Redistribution and Rekognition: A Feminist Critique of Algorithmic Fairness (11/7/20)

"What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness (10/30/20)

Recipes for Safety in Open-domain Chatbots (10/22/20)

The "black box" at work (10/19/20)

Exchanging Lessons Between Algorithmic Fairness and Domain Generalization (10/14/20)

The Ethics of Emotion in AI Systems (10/8/20)

Brave: what it means to be an AI Ethicist (10/6/20)

CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models (9/30/21)

REAL TOXICITY PROMPTS: Evaluating Neural Toxic Degeneration in Language Models (9/25/20)

Counterfactual Explanation and Causal Inference in Service of Robustness in Robot Control (9/18/20)

Compounding Injustice: The Cascading Effect of Algorithmic Bias in Risk Assessments (8/11/20)

Discovering and Categorising Language Biases in Reddit\* (8/6/20)

## What are you optimizing for? Aligning Recommender Systems with Human Values (2020)

A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government (Aug 2020)

Artificial Intelligence and Inequality in the Middle East: The Political Economy of Inclusion (7/20/20)

Emerging challenges in AI and the need for AI ethics education (7/15/20)

Green Algorithms: Quantifying the carbon footprint of computation (7/15/20)

Machine Learning Explainability for External Stakeholders (7/10/20)

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList (July 2020)

Social Bias Frames: Reasoning about Social and Power Implications of Language (July 2020)

AI and the Global South: Designing for Other Worlds (July 2020)

Don't ask if artificial intelligence is good or fair, ask how it shifts power (7/7/20)

Speech Recognition Tech Is Yet Another Example of Bias (7/5/20)

Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis? (7/5/20)

The Ethics of Algorithms: Key Problems and Solutions (July 2020)

Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics (6/24/20)

Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices (6/22/20)

Algorithmic Bias: On the Implicit Biases of Social Technology (6/20/21)

Hidden in Plain Sight – Reconsidering the Use of Race Correction in Clinical Algorithms (6/17/20)

The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation (6/17/20)

Mind the app -- Considerations on the ethical risks of COVID-19 apps (6/13/20)

In AI We Trust: Ethics, Artificial Intelligence, and Reliability (6/10/20)

Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis (6/9/20)

Understanding the Essence of Artificial Intelligence: Towards Ecological Safety of AI in Human Society (Jun 2020)

Fairness, Equality, and Power in Algorithmic Decision-Making (preprint)

Language (Technology) is Power: A Critical Survey of "Bias" in NLP (5/28/20)

IEEE 7010: A New Standard for Assessing the Well-being Implications of Artificial Intelligence (5/7/20)

Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation (5/3/20)

Social Biases in NLP Models as Barriers for Persons with Disabilities (5/2/20)

Multi-Dimensional Gender Bias Classification (5/1/20)

Politeness Transfer: A Tag and Generate Approach (5/1/20)

Ethics and governance for digital disease surveillance (May 2020)

Mind Your Inflections! Improving NLP for Non-Standard English with Base-Inflection Encoding (4/30/20)

The Politics of Adversarial Machine Learning (4/26/20)

Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI (4/25/20)

Learning to Diversify from Human Judgments: Research Directions and Open Challenges (4/25/20)

A Human in the Loop is Not Enough: The Need for Human-Subject Experiments in Facial Recognition (4/25/20)

Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning (4/25/20)

StereоСet measures racism, sexism, and other forms of bias in AI language models (4/22/20)

Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims (4/15/20)

Research Collection: Research Supporting Responsible AI (4/13/20)

Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem (4/9/20)

Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal (4/7/20)

Measuring the predictability of life outcomes with a scientific mass collaboration (3/30/20)

AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media (3/29/20)

Racial disparities in automated speech recognition (3/23/20)

What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis (3/11/20)

Recommender Systems and their Ethical Challenges (2/27/20)

No computation without representation: Avoiding data and algorithm biases through diversity (2/26/20)

Informed Consent and Medical Artificial Intelligence: What to Tell the Patient? (2/26/20)

Learning the difference that makes a difference with counterfactually-augment data (2/14/20)

Diversity and Inclusion Metrics in Subset Selection (2/9/20)

Hazard Contribution Modes of Machine Learning Components (2/7/20)

The Ethics of AI Ethics: An Evaluation of Guidelines (2/1/20)

Deontological Ethics By Monotonicity Shape Constraints (1/31/20)

From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy (1/29/20)

The Explanation Game: A Formal Framework for Interpretable Machine Learning (1/17/20)

Robot Rights? Let's Talk about Human Welfare Instead (1/14/20)

Algorithmic Fairness from a Non-ideal Perspective (1/9/20)

Toward Situated Interventions for Algorithmic Equity: Lessons from the Field (Jan 2020)

Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning (Jan 2020)

“The Human Body is a Black Box”: Supporting Clinical Decision-Making with Deep Learning (Jan 2020)

Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure (Jan 2020)

## Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing (Jan 2020)

Fairness and utilization in allocating resources with uncertain demand (Jan 2020)

## **Government, NGO, and expert publications**

Crucial Yet Overlooked: Why We Must Reconcile Legal and Technical Approaches to Algorithmic Bias (12/17/20)

Landmark artificial intelligence legislation advances toward becoming law (12/16/20)

How Many Jobs Will AI Destroy? As Many As We Tell It To. (12/16/20)

Creating Trustworthy AI (12/15/20)

WEF: Responsible Limits on Facial Recognition Use Case: Flow Management (12/14/20)

In AI ethics, “bad” isn’t good enough (12/14/20)

The ethics of artificial intelligence (12/10/20)

Why people may not trust your AI, and how to fix it (12/8/20)

What the AI Community Can Learn From Sneezing Ferrets and a Mutant Virus Debate (12/8/20)

The tech industry needs regulation for its systemically important companies (12/8/20)

Hong Qu: Shining a Headlight on AI Blindspots (12/3/20)

Report – Algorithm-driven Hiring Tools: Innovative Recruitment or Expedited Disability Discrimination? (12/3/20)

When Algorithmic Fairness Fixes Fail: The Case for Keeping Humans in the Loop (12/2/20)

How News Organizations Use Algorithms to Decide What to Show You (12/1/20)

Can AI Fairly Decide Who Gets an Organ Transplant? (12/1/20)

How AI bots and voice assistants reinforce gender bias (11/23/20)

Corsight AI launches real-time facial recognition that identifies masked faces (11/7/20)

Google AI Blog: Mitigating Unfair Bias in ML Models with the MinDiff Framework (11/16/20)

Retooling AI: Algorithm Bias and the Struggle to Do No Harm (11/10/20)

A Legal Approach to “Affirmative Algorithms” (11/9/20)

AUTOMATING SOCIETY REPORT 2020 (EU) (Nov 2020)

Stanford HAI Policy Brief: Domain Shift and Emerging Questions in Facial Recognition Technology (Nov 2020)

Stanford HAI Policy Brief: Preparing for the Age of Deepfakes and Disinformation (Nov 2020)

Root Out Bias at Every Stage of Your AI-Development Process

(10/30/20)

Spanish police plan to extend use of its lie-detector while efficacy is unclear (10/27/20)HBR: When Do We Trust AI's Recommendations More Than People's?

(10/16/20)

How to use AI hiring tools to reduce bias in recruiting (10/13/20)Men drive trucks, women raise children - discriminatory gendertargeting through Facebook (10/11/20)The Short Anthropological Guide to the Study of Ethical AI (10/10/20)Net worthy: How to get budget and buy-in for tech ethics (and other forms of responsible business) (Oct 2020)Stanford HAI Policy Brief: Toward Fairness in Health Care Training Data

(Oct 2020)

HBR: Automated Hardship: How the Tech-Driven Overhaul of the UK'sSocial Security System Worsens Poverty (9/29/20)NEC and The Face Recognition Company each upgrade biometricsystems accuracy for masked faces (9/24/21)POVERTY LAWGORITHMS: A Poverty Lawyer's Guide to FightingAutomated Decision-Making Harms on Low-Income Communities

(9/15/20)

AI ethics groups are repeating one of society's classic mistakes

(9/14/20)

AI's Promise and Peril for the U.S. Government [Stanford HAI] (Sept 2020)

Framework for Promoting Workforce Well-being in the AI-Integrated Workplace (8/27/20)

GAO Facial Recognition Technology: Privacy and Accuracy Issues Related to Commercial Uses (8/11/20)

Risks of Discrimination through the Use of Algorithms [German Federal Anti-Discrimination Agency] (8/7/20)

NIST: Four Principles of Explainable Artificial Intelligence (Aug 2020)

BMM Privacy Bot: Tool to anonymize BLM protesters

Report: Bridging AI's trust gaps: Aligning policymakers and companies (The Future Society) (7/22/20)

Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment (7/17/20)

Facebook's Civil Rights Audit – Final Report (7/8/20)

AI Watch: Artificial Intelligence in public services (July 2020)

AUTOMATED SUSPICION: THE EU'S NEW TRAVEL SURVEILLANCE INITIATIVES (July 2020)

The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law: Ad Hoc Committee on Artificial Intelligence (6/24/20)

A call for a critical look at the metrics for success in the evaluation of AI  
(6/17/20)

ARTIFICIAL INTELLIGENCE ETHICS FRAMEWORK FOR THE INTELLIGENCE COMMUNITY (June 2020)

FACIAL RECOGNITION TECHNOLOGIES: May 29, 2020 A PRIMER  
(5/29/20)

When AI Writes Your Email (5/6/20)

UC Berkeley Decision Points in AI Governance THREE CASE STUDIES EXPLORE EFFORTS TO OPERATIONALIZE AI PRINCIPLES (May 2020)

Artificial Intelligence for Social Good in Latin America and the Caribbean: The Regional Landscape and 12 Country Snapshots (May 2020)

Is AI trustworthy enough to help us fight COVID-19? (5/1/20)

A Scalable Approach to Reducing Gender Bias in Google Translate  
(4/22/20)

How to create a trustworthy COVID-19 tracking technology (4/20/20)

Aligning AI to Human Values means Picking the Right Metrics  
(4/15/20)

Measuring the predictability of life outcomes with a scientific mass collaboration (4/14/20)

Statement Regarding the Ethical Implementation of Artificial Intelligence Systems (AIS) for Addressing the COVID-19 Pandemic (April 2020)

Interpretability [Cloudera Fast Forward Report] (Apr 2020)

FTC: Using Artificial Intelligence and Algorithms (4/8/20)

From Principles to Practice An interdisciplinary framework to operationalise AI ethics (4/2/20)

The DoD AI Ethical Principles: Shifting From Principles to Practice (4/1/20)

AI is transforming society. Here's what we can do to make sure it prioritizes human needs. (3/27/20)

To Prevent Algorithmic Bias, Legal and Technical Definitions around Algorithmic Fairness Must Align (3/23/20)

Digital Commerce, AI, and Constraining Consumer Choice (3/19/20)

Mapping coronavirus, responsibly (2/25/20)

Algorithm Groups People More Fairly to Reduce AI Bias (2/12/20)

Protecting privacy in an AI-driven world (2/10/20)

ML-fairness-gym: A Tool for Exploring Long-Term Impacts of Machine Learning Systems (2/5/20)

The measure and mismeasure of fairness: a critical review of fair machine learning (2/3/20)

## Bringing Facial Recognition Systems To Light (Feb 2020)

Artificial Intelligence in the Asia-Pacific Region: Examining policies and strategies to maximise AI readiness and adoption (Feb 2020)

IBM: Precision Regulation for Artificial Intelligence (1/21/20)

Black-Boxed Politics: Opacity is a Choice in AI Systems (1/17/20)

What do we teach when we teach tech & AI ethics? (1/17/20)

How businesses can create an ethical culture in the age of tech (1/7/20)

Artificial Intelligence Principles For Vulnerable Populations in Humanitarian Contexts (Jan 2020)

Excavating AI: The Politics of Images in Machine Learning Training Sets (Jan 2020)

Learning from the past to create Responsible AI (evergreen)

## **Popular Press**

Unregulated facial recognition must stop before more Black men are wrongfully arrested (12/31/20)

In 2020, COVID-19 derailed the privacy debate in the U.S (12/29/20)

Why 2020 was a pivotal, contradictory year for facial recognition (12/29/20)

[Researchers find race, gender, and style biases in art-generating AI systems \(12/28/20\)](#)

[AI jobs in 2021: here are some key trends \(12/28/20\)](#)

[AI research survey finds machine learning needs a culture change \(12/26/20\)](#)

[He spent 10 days in jail after facial recognition software led to the arrest of the wrong man, lawsuit says \(12/28/20\)](#)

[Algorithms are deciding who gets the first vaccines. Should we trust them? \(12/23/20\)](#)

[Dozens sue Amazon's Ring after camera hack leads to threats and racial slurs \(12/23/20\)](#)

[Pulse Oximeter Devices Have Higher Error Rate in Black Patients \(12/22/20\)](#)

[This is the Stanford vaccine algorithm that left out frontline doctors \(12/21/20\)](#)

[How to create space for ethics in AI \(12/19/20\)](#)

[Stanford Apologizes After Vaccine Allocation Leaves Out Nearly All Medical Residents \(12/18/20\)](#)

[Pôle emploi makes it much too easy for anyone to obtain French CVs \(12/17/20\)](#)

[From whistleblower laws to unions: How Google's AI ethics meltdown could shape policy \(12/16/20\)](#)

[Canada] Use of surveillance software to crack down on exam cheating has unintended consequences (12/16/20)

Researchers describe profound complexity of biometrics ethics puzzle (12/15/20)

Surveillance companies are using mobile ads to obtain scarily accurate location data (12/11/20)

AI needs to face up to its invisible-worker problem (12/11/20)

Sci-fi surveillance: Europe's secretive push into biometric technology (12/10/20)

How our data encodes systematic racism (12/10/20)

Study finds crime-predicting judicial tool exhibits gender bias (12/10/20)

Study finds diversity in data science teams is key in reducing algorithmic bias (12/9/20)

IBM announces new AI language, explainability, and automation services (12/9/20)

Uni revealed it killed off its PhD-applicant screening AI – just as its inventors gave a lecture about the tech (12/8/20)

Huawei tested AI software that could recognize Uighur minorities and alert police, report says (12/8/20)

You can buy a robot to keep your lonely grandparents company. Should you? (12/8/20)

## Artificial Intelligence Is An Amazing Disruptor, But Has A Major Impact On Unskilled Workers (12/8/20)

Surprise, surprise: AI cameras sold to schools in New York struggle with people of color and are full of false positives (12/7/20)

How banks use AI to catch criminals and detect bias (12/7/20)

We can reduce gender bias in natural-language AI, but it will take a lot more work (12/6/20)

The coming war on the hidden algorithms that trap people in poverty (12/4/20)

Health algorithms discriminate against Black patients, also in Switzerland (12/4/20)

[France] The Interior strengthens the possibilities of political identification (12/4/20)

Study shows how AI exacerbates recruitment bias against women (12/2/20)

Microsoft apologises for feature criticised as workplace surveillance (12/2/20)

How The Department Of Defense Approaches Ethical AI (11/29/20)

Ethical AI isn't the same as trustworthy AI, and that matters (11/28/20)

[UK] The Covid data spies paid to know ALL your secrets: Town halls harvest millions of highly personal details including if you're being

unfaithful or having unsafe sex (11/27/20)

Uganda is using Huawei's facial recognition tech to crack down on dissent after anti-government protests (11/27/20)

Who will your algorithm harm next? Why businesses need to start thinking about evil AI now (11/26/20)

Can an Algorithm Prevent Suicide? (11/23/20)

[France] French tax authority pushes for automated controls despite mixed results (11/23/20)

The UK Government Isn't Being Transparent About Its Palantir Contracts (11/23/20)

China's Surveillance State Sucks Up Data. U.S. Tech Is Key to Sorting It. (11/22/20)

Watch: Facial recognition at Dubai Metro stations to identify wanted criminals (11/22/20)

Can We Make Our Robots Less Biased Than We Are? (11/22/20)

Hundreds of Facebook moderators complain: AI content moderation isn't working and we're paying for it (11/21/20)

New York City wants to restrict artificial intelligence in hiring (11/20/20)

These Algorithms Could Bring an End to the World's Deadliest Killer (11/20/20)

[Facebook's latest efforts to combat hate speech aren't enough, ADL says \(11/20/20\)](#)

[When AI Sees a Man, It Thinks 'Official.' A Woman? 'Smile' \(11/19/20\)](#)

[Facebook's A.I. is getting better at finding malicious content—but it won't solve the company's problems \(11/19/20\)](#)

[Banks turn to AI as regulators press for Libor exit \(11/19/20\)](#)

[Zest AI Joins Forces With Freddie Mac to Help Make Homeownership Possible for More Americans \(11/19/20\)](#)

[When AI Systems Fail: Introducing the AI Incident Database \(11/18/20\)](#)

[The way we train AI is fundamentally flawed \(11/18/20\)](#)

[How regulators can get facial recognition technology right \(11/17/20\)](#)

[Health Care AI Systems Are Biased \(11/17/20\)](#)

[Online exams raise concerns of racial bias in facial recognition \(11/17/20\)](#)

[\[Australia\] Robodebt class action: Coalition agrees to pay \\$1.2bn to settle lawsuit \(11/16/20\)](#)

[Activists urge EU to ban live facial recognition in public spaces \(11/12/20\)](#)

['Coded Bias' Review: When the Bots Are Racist \(11/11/20\)](#)

Why AI can't move forward without diversity, equity, and inclusion

(11/12/20)

Equitable tech: AI-enabled platform to reduce bias in datasets released

(11/11/20)

AI research finds a 'compute divide' concentrates power and

accelerates inequality in the era of deep learning (11/11/20)

"Data Trusts" Could Be the Key to Better AI (11/10/20)

The US Government Will Pay Doctors to Use These AI Algorithms

(11/10/20)

Face for sale: Leaks and lawsuits blight Russia facial recognition

(11/9/20)

How artificial intelligence may be making you buy things (11/9/20)

Pope urges Catholics to pray that AI does not widen inequality

(11/7/20)

How IBM Is Working Toward a Fairer AI (11/5/20)

Proctorio used DMCA to take down a student's critical tweets

(11/5/20)

Artificial intelligence is making the beauty industry work for everyone

(11/4/20)

When AI Says, 'You're Fired' (11/4/20)

Portland, Maine has voted to ban facial recognition (11/4/20)

Police Will Pilot a Program to Live-Stream Amazon Ring Cameras

(11/3/20)

How the Police Use AI to Track and Identify You (11/3/20)

Schools Adopt Face Recognition in the Name of Fighting Covid

(11/3/20)

Artificial Intelligence in healthcare is racist (11/2/20)

When Algorithms Infer Pregnancy or Other Sensitive Information

About People (11/2/20)

Fairness Definitions Explained (Research Summary) (11/2/20)

Big tech's 'blackbox' algorithms face regulatory oversight under EU

plan (10/30/20)

The AI Company Helping the Pentagon Assess Disinfo Campaigns

(10/28/20)

Nearly half of councils in Great Britain use algorithms to help make

claims decisions (10/28/20)

Uber Is Getting Sued Over Its Allegedly Racist Ratings System

(10/27/20)

Why Getting Paid for Your Data Is a Bad Deal (10/26/20)

Uber Drivers Launch Legal Action Over 'Robo-Firing' By Algorithm

(10/26/20)

How an Algorithm Blocked Kidney Transplants to Black Patient

(10/26/20)

AI bias: blame the workman, not his tools (10/26/20)

Surveillance Startup Used Own Cameras to Harass Coworkers

(10/26/20)

How the Racism Baked Into Technology Hurts Teens (10/24/20)

How to make a chatbot that isn't racist or sexist (10/23/20)

AI Weekly: Constructive ways to take power back from Big Tech

(10/23/20)

The De-democratization of AI: Deep Learning and the Compute Divide

in Artificial Intelligence Research (10/22/20)

AI researchers urge tech to go beyond scale to address systemic social

issues (10/22/20)

Putting Responsible AI Into Practice (10/22/20)

Algorithms Are Making Economic Inequality Worse (10/22/20)

The true dangers of AI are closer than we think (10/21/20)

Activists Turn Facial Recognition Tools Against the Police (10/21/20)

GOOGLE AI TECH WILL BE USED FOR VIRTUAL BORDER WALL, CBP

CONTRACT SHOWS (10/21/20)

Facial recognition datasets are being widely used despite being taken down due to ethical concerns. Here's how. (10/21/20)

White House Nears New Rules on Artificial Intelligence (10/21/20)

Salesforce's Simulation Cards spell out uses, risks, and bias to make AI models more transparent (10/20/20)

Photoshop's AI neural filters can tweak age and expression with a few clicks (10/20/20)

AI Fairness Isn't Just an Ethical Issue (10/20/20)

Singapore releases AI ethics, governance reference guide (10/16/20)

UK passport photo checker shows bias against dark-skinned women (10/16/20)

The real promise of synthetic data (10/16/20)

BMW develops AI ethics code for its cars (10/16/20)

A practical guide to building ethical AI (10/15/20)

AI for good: A better, more inclusive future of work (10/15/20)

Six ways machine learning threatens social justice (10/15/20)

'Machines set loose to slaughter': the dangerous rise of military AI (10/15/20)

Six ways machine learning threatens social justice (10/15/20)

AI Reads Human Emotions. Should it? (10/14/20)

Politicians have made an algorithm to fix the housing crisis. It's bad  
(10/14/20)

Which Company Uses the Most of Your Data? (10/14/20)

From a small town in North Carolina to big-city hospitals, how software infuses racism into U.S. health care (10/13/20)

When AI hurts people, who is held responsible? (10/12/20)

Shrinking the 'data desert': Inside efforts to make AI systems more inclusive of people with disabilities (10/12/20)

AI Can Help Diagnose Some Illnesses–If Your Country Is Rich  
(10/11/20)

Toddlers Are Being Scooped Up in Buenos Aires' Live Facial Recognition Dragnet (10/9/20)

UK passport photo checker shows bias against dark-skinned women  
(10/7/20)

The biggest barrier to humane, ethical AI: Capitalism itself (10/5/20)

Facial Recognition, Racial Recognition and the Clear and Present Issues with AI Bias (10/5/20)

Twitter vows to fix bias image cropping issue (10/2/20)

AI and ethics: One-third of executives are not aware of potential AI bias (10/2/20)

## 6 big ethical questions about the future of AI (Oct 2020)

A Facial Action Coding system to explain what women are thinking?

You could just ask (9/30/20)

How not to build a biased algorithm (9/30/20)

Alexa, do I have COVID-19? (9/30/20)

The True Impact of Sexist and Racist Bias in Algorithms (9/29/20)

ExamSoft's remote bar exam sparks privacy and facial recognition concerns (9/29/20)

Netherlands: End dangerous mass surveillance policing experiments (9/29/20)

Our goal shouldn't be to build merely 'trustworthy' AI (9/28/20)

Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI (9/28/20)

AI Democratization in the Era of GPT-3 (9/25/20)

'I'm extremely controversial': the psychologist rethinking human emotion (9/25/20)

Students Are Pushing Back Against Proctoring Surveillance Apps (9/25/20)

How close is AI to decoding our emotions? (9/24/20)

Amazon is embracing surveillance-as-a-service (9/24/20)

GPT-3's bigotry is exactly why devs shouldn't use the internet to train AI

(9/24/20)

A controversial photo editing app slammed for AI-enabled 'blackface'

feature (9/24/20)

Locked Out: Can Algorithms Violate Fair Housing Laws? (9/24/20)

Regulating biometrics: Taking stock of a rapidly changing landscape

(9/23/20)

The geographic bias in medical AI tools (9/23/20)

Can ethics classes actually influence students' moral behavior?

(9/23/20)

To Clean Up Comments, Let AI Tell Users Their Words Are Trash

(9/22/20)

To Make Fairer AI, Physicists Peer Inside Its Black Box (9/22/20)

Another reminder that bias, testing, diversity is needed in machine

learning: Twitter's image-crop AI may favor white men, women's chests

(9/21/20)

Twitter and Zoom's algorithmic bias issues (9/21/20)

Twitter apologises for 'racist' image-cropping algorithm (9/21/20)

Diversity in AI: The Invisible Men and Women (9/21/20)

Twitter is looking into why its photo preview appears to favor white

faces over Black faces (9/20/20)

Algorithms may never really figure us out – thank goodness (9/20/20)

Is an Algorithm Less Racist Than a Loan Officer? (9/18/20)

Facebook Testing Implications of Privacy-Invasive Tech By Invading People's Privacy (9/18/20)

Cutting-edge language models can produce convincing misinformation if we don't stop them (9/18/20)

Fake Data Could Help Solve Machine Learning's Bias Problem—if We Let It (9/17/20)

Unethical AI unfairly impacts protected classes - and everybody else as well (9/16/20)

Ethical Tech Starts With Addressing Ethical Debt (9/16/20)

Where is the accountability for AI ethics gatekeepers? (9/15/20)

AI researchers devise failure detection method for safety-critical machine learning (9/15/20)

AI ethics groups are repeating one of society's classic mistakes (9/14/20)

How Algorithms Can Fight Bias Instead of Entrench It (9/14/20)

Why the AI industry urgently needs more uncomfortable conversations about BAME representation (9/14/20)

New standards for AI clinical trials will help spot snake oil and hype (9/11/20)

Could you live here? Decades-old redlining still shapes San Diego's neighborhoods today, advocates say (9/11/20)

Portland passes broadest facial recognition ban in the US (9/9/20)

Portland Passes Groundbreaking Ban on Facial Recognition in Stores, Banks, Restaurants and More (9/9/20)

Mutant Algorithms Are Coming for Your Education (9/8/20)

When Algorithms Give Real Students Imaginary Grades (9/8/20)

Here are the biased algorithms the UK government uses to make high-level decisions (9/7/20)

From viral conspiracies to exam fiascos, algorithms come with serious side effects (9/6/20)

Facebook AI open-sources Opacus, a new high-speed library for training PyTorch models with differential privacy (DP) (9/6/20)

Pasco's sheriff created a futuristic program to stop crime before it happens: It monitors and harasses families across the county. (9/3/20)

What Does Building a Fair AI Really Entail? (9/3/20)

Racial biases infect artificial intelligence (9/2/20)

Biometrics Have Crept Into Humanitarian Aid, But the Systems May Disadvantage Women Who Need Help Most (9/2/20)

Microsoft launches a deepfake detector ahead of US election (9/2/20)  
Bodies into Bits (8/31/20)

[Google Offers to Help Others With the Tricky Ethics of AI \(8/28/20\)](#)

[Rooting out racism in AI systems -- there's no time to lose \(8/28/20\)](#)

[Facebook's discriminatory ad targeting illustrates the dangers of biased algorithms \(8/28/20\)](#)

[PopID's face-based payments pose privacy and security risks \(8/27/20\)](#)

[The utopian promise and dystopian potential of real-time detection of police, fire, and medical emergencies \(8/27/20\)](#)

[Participation-washing could be the next dangerous fad in machine learning \(8/25/20\)](#)

[Does Facebook Still Sell Discriminatory Ads? \(8/25/20\)](#)

[LinkedIn open-sources toolkit to measure AI model fairness \(8/25/20\)](#)

[Councils scrapping use of algorithms in benefit and welfare decisions \(8/24/20\)](#)

[The term 'ethical AI' is finally starting to mean something \(8/23/20\)](#)

[Algorithms can drive inequality. Just look at Britain's school exam chaos \(8/23/20\)](#)

[Best Practices for Indigenous Keywording for Stock Images \(8/20/20\)](#)

[Why 'Ditch the algorithm' is the future of political protest \(8/19/20\)](#)

[INSIGHT: Facial Recognition Is a Threat to People of Color \(8/18/20\)](#)

Too many AI researchers think real-world problems are not relevant

(8/18/20)

Does technology increase the problem of racism and discrimination?

(8/17/20)

An A.I. Training Tool Has Been Passing Its Bias to Algorithms for Almost Two Decades (8/17/20)

Governments have collected large amounts of data to fight the coronavirus. That's raising privacy concerns (8/17/20)

Inbuilt biases and the problem of algorithms (8/17/20)

Facebook algorithm found to 'actively promote' Holocaust denial

(8/16/20)

Human Rights Commission warns government over 'dangerous' use of AI (8/15/20)

U.K. Regulator Issues Data Guidance for Companies Working With AI

(8/14/20)

Horrific AI surveillance experiment uses convicted felons as human guinea pigs (8/14/20)

Problematic study on Indiana parolees seeks to predict recidivism with AI (8/14/20)

NYPD Used Facial Recognition Technology In Siege Of Black Lives

Matter Activist's Apartment (8/14/20)

Facebook's AI for detecting hate speech is facing its biggest challenge yet (8/14/20)

The Quiet Growth of Race-Detection Software Sparks Concerns Over Bias (8/14/20)

Governments should close the AI trust gap with businesses (8/13/20)

The pandemic is speeding up automation, putting jobs in question (8/11/20)

Ethics Is More Important than Technology (8/10/20)

WOULD YOU TRUST A LAWYER BOT WITH YOUR LEGAL NEEDS? (8/10/20)

Research summary: Mass Incarceration and the Future of AI (8/9/20)

Researchers quantify bias in Reddit content sometimes used to train AI (8/7/20)

Hypotenuse AI wants to take the strain out of copywriting for e-commerce (8/7/20)

Can language models learn morality? (8/7/20)

AI Project Failure Rates Near 50%, But It Doesn't Have to Be That Way, Say Experts (8/7/20)

Here are a few ways GPT3 can go wrong (8/7/20)

AI models need to be 'interpretable' rather than just 'explainable' (8/6/20)

Researchers discover evidence of gender bias in major computer vision

APIs (8/6/20)

Researchers say 'The Whiteness of AI' in pop culture erases people of

color (8/6/20)

Police built an AI to predict violent crime. It was seriously flawed

(8/6/20)

Explainable AI: A guide for making black box machine learning models

explainable (8/6/20)

How a Popular Medical Device Encodes Racial Bias (8/5/20)

Evil AI: These are the 20 most dangerous crimes that artificial

intelligence will create (8/5/20)

Can Artificial Intelligence Help Increase Diversity in IT? (8/5/20)

Meet the computer scientist and activist who got Big Tech to stand

down (8/4/20)

New Zealand Has a Radical Idea for Fighting Algorithmic Bias:

Transparency (8/4/20)

Cloak your photos with this AI privacy tool to fool facial recognition

(8/4/20)

Home Office to scrap 'racist algorithm' for UK visa applicants (8/4/20)

The problem of underrepresented languages snowballs from data sets

to NLP models (8/4/20)

## On Diversity, Silicon Valley Failed to Think Different (8/3/20)

Why artificial intelligence models are often biased, according to the Google exec who heads Alphabet's internal tech incubator Jigsaw (8/3/20)

AI is struggling to adjust to 2020 (8/2/20)

The 6 unholy AI systems thou shalt not develop (7/31/20)

The problems AI has today go back centuries (7/31/20)

Researchers examine the ethical implications of AI in surgical settings (7/31/20)

AI-powered tool aims to help reduce bias and racially charged language on websites (7/30/20)

OPENAI'S LATEST BREAKTHROUGH IS ASTONISHINGLY POWERFUL, BUT STILL FIGHTING ITS FLAWS (7/30/20)

Over Half of Americans Do Not Trust Companies to Ethically Collect, Use or Sell Personal Data (7/29/20)

Google releases Model Card Toolkit to promote AI model transparency (7/29/20)

Service that uses AI to identify gender based on names looks incredibly biased (7/29/20)

Power, Harms, and Data (7/28/20)

Rite Aid deployed facial recognition systems in hundreds of U.S. stores

(7/28/20)

Even Google CEO Sundar Pichai agrees that it is imperative to embed

ethics into AI (7/27/20)

Researchers find evidence of bias in facial expression data sets

(7/25/20)

Here's why AI didn't save us from COVID-19 (7/24/20)

Four Steps for Drafting an Ethical Data Practice Blueprint (7/24/20)

An AI hiring firm says it can predict job hopping based on your

interviews (7/24/20)

An online image database will remove 600,000 pictures after an art

project revealed the system's racist bias. (9/24/20)

AI says men are lazy (7/24/20)

Google Ad Portal Equated “Black Girls” with Porn (7/23/20)

Facebook ignored racial bias research, employees say (7/23/20)

AI system detects posts by foreign ‘trolls’ on Facebook and Twitter

(7/22/20)

Blind to black people (7/21/20)

Dermatology faces a reckoning: Lack of darker skin in textbooks and

journals harms care for patients of color (7/21/20)

[Facebook Creates Teams to Study Racial Bias, After Previously Limiting Such Efforts \(7/21/20\)](#)

[Clients loved this designer's work. Turns out, he was an AI \(7/20/20\)](#)

[Tackling the misinformation epidemic with "In Event of Moon Disaster" \(7/20/20\)](#)

[Why Hundreds of Mathematicians Are Boycotting Predictive Policing \(7/20/20\)](#)

[WHY RACIAL BIAS STILL HAUNTS SPEECH-RECOGNITION AI \(7/17/20\)](#)

[Predictive policing algorithms are racist. They need to be dismantled. \(7/17/20\)](#)

[This grading algorithm is failing students \(7/16/20\)](#)

[MIT researchers find 'systematic' shortcomings in ImageNet data set \(7/15/20\)](#)

[Deepfake used to attack activist couple shows new disinformation frontier \(7/15/20\)](#)

[DARPA Pays \\$1 Million For An AI App That Can Predict An Enemy's Emotions \(7/15/20\)](#)

[An invisible hand: Patients aren't being told about the AI systems advising their care \(7/15/20\)](#)

[THE MICROSOFT POLICE STATE: MASS SURVEILLANCE, FACIAL RECOGNITION, AND THE AZURE CLOUD \(7/14/20\)](#)

Facial recognition technology is one of the most racist weapons in the police arsenal (7/14/20)

Massachusetts Edges Toward First State Facial Recognition Ban (7/14/20)

Op-Ed: Bias against African American English speakers is a pillar of systemic racism (7/14/20)

Why are Artificial Intelligence systems biased? (7/12/20)

DeepMind researchers propose rebuilding the AI industry on a base of anticolonialism (7/11/20)

Controversial Detroit facial recognition got him arrested for a crime he didn't commit (7/10/20)

Meet the Secret Algorithm That's Keeping Students Out of College (7/10/20)

AI for self-driving cars doesn't account for crime (7/10/20)

Automatic for the Bosses (7/9/20)

Deepfakes are becoming the hot new corporate training tool (7/7/20)

RACIAL EQUITY IN DATA INTEGRATION: HOW TO EXCLUDE RACIAL BIAS IN DATA (7/6/20)

How AI can empower communities and strengthen democracy (7/4/20)

Study: Only 18% of data science students are learning about AI ethics

(7/3/20)

9 emerging job roles for the future of AI (7/2/20)

MIT Takes Down Popular AI Dataset Due to Racist, Misogynistic Content (7/2/20)

Salesforce researchers claim new method mitigates AI models' gender bias (7/1/20)

MIT removes huge dataset that teaches AI systems to use racist, misogynistic slurs (7/1/20)

MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs (7/1/20)

Vivienne Ming on Understanding the Impact of Courage & Fear Using AI (6/30/20)

Vestager warns against predictive policing in Artificial Intelligence (6/30/20)

New mathematical idea reins in AI bias towards making unethical and costly commercial choices (6/30/20)

Machine learning systems, fair or biased, reflect our moral standards (6/30/20)

Montreal AI Ethics Institute suggests ways to counter bias in AI models (6/30/20)

Objective Algorithms Are a Myth (6/29/20)

[Here Is How The United States Should Regulate Artificial Intelligence](#)

(6/28/20)

[AI Weekly: A deep learning pioneer's teachable moment on AI bias](#)

(6/26/20)

[AI gatekeepers are taking baby steps toward raising ethical standards](#)

(6/26/20)

[Why Statistics Don't Capture The Full Extent Of The Systemic Bias In](#)

[Policing \(6/25/20\)](#)

[If Done Right, AI Could Make Policing Fairer \(6/25/20\)](#)

[Lessons from the PULSE Model and Discussion \(6/24/20\)](#)

[Wrongfully Accused by an Algorithm \(6/24/20\)](#)

[If AI is going to help us in a crisis, we need a new kind of ethics](#)

(6/24/20)

[California city bans predictive policing in U.S. first \(6/24/20\)](#)

[What a machine learning tool that turns Obama white can \(and can't\) tell us about AI bias \(6/23/20\)](#)

[Over 1,000 AI Experts Condemn Racist Algorithms That Claim to Predict Crime \(6/23/20\)](#)

[Police Forces in Canada Are Quietly Adopting Facial Recognition Tech](#)

(6/23/20)

The flaws that make today's AI architecture unsafe and a new approach that could fix it (6/22/20)

Big Tech Is Using the Pandemic to Push Dangerous New Forms of Surveillance (6/22/20)

US government doesn't know how it uses facial recognition in public housing (6/22/20)

The flaws that make today's AI architecture unsafe and a new approach that could fix it (6/22/20)

WE DON'T HAVE TO SETTLE FOR BIASED AI (6/18/20)

Here's an Algorithm for Defunding the Police (6/18/20)

Racial bias skews algorithms widely used to guide care from heart surgery to birth, study finds (6/17/20)

How is Face Recognition Surveillance Technology Racist? (6/16/20)

Riding Out Quarantine With a Chatbot Friend: 'I Feel Very Connected' (6/16/20)

Fresh concerns about AI bias in the age of COVID-19 (6/15/20)

The Liabilities of Artificial Intelligence Are Increasing (6/15/20)

Who Is Responsible When Autonomous Systems Fail? (6/15/20)

Mind the App—Considerations on the Ethical Risks of COVID-19 Apps (6/13/20)

Researchers find racial discrimination in ‘dynamic pricing’ algorithms used by Uber, Lyft, and others (6/12/20)

Researchers propose framework to measure AI’s social and environmental impact (6/12/20)

US government doesn't know how it uses facial recognition in public housing (6/11/20)

The Protests Prove the Need to Regulate Surveillance Tech (6/9/20)

How to detect unwanted bias in machine learning models (6/5/20)

This startup is using AI to give workers a “productivity score” (6/4/20)

Amazon “Stands in Solidarity” Against Police Racism While Selling Racist Tech to Police (6/3/20)

Of course technology perpetuates racism. It was designed that way. (6/3/20)

Microsoft researchers say NLP bias studies must consider role of social hierarchies like racism (6/1/20)

Gender Bias In Predictive Algorithms: How Applied AI Research Can Help Us Build A More Equitable Future (5/30/20)

Access Denied: Faulty Automated Background Checks Freeze Out Renters (5/28/20)

ACLU sues facial recognition firm Clearview AI, calling it a ‘nightmare scenario’ for privacy (5/28/20)

How people with Down syndrome are teaching Google (5/20/20)

[This robot can guess how you're feeling by the way you walk \(5/18/20\)](#)

[Algorithms associating appearance and criminality have a dark past \(5/15/20\)](#)

[AI and me: friendship chatbots are on the rise, but is there a gendered design flaw? \(5/7/20\)](#)

[Facebook's AI detects gender bias in text \(5/6/20\)](#)

[Do I sound sick to you? Researchers are building AI that would diagnose COVID-19 by listening to people talk. \(4/30/20\)](#)

[Aerial surveillance planes to begin flying over Baltimore Friday \(4/30/20\)](#)

[Algorithmic Risk Assessment and COVID-19: Why PATTERN Should Not Be Used \(4/30/20\)](#)

[Google Fixes Gender Bias in Google Translate \(Again\) \(4/29/20\)](#)

[Silicon Valley needs a new approach to studying ethics now more than ever \(4/24/20\)](#)

[The role of demographic data in addressing algorithmic bias \(4/24/20\)](#)

[StereoSet measures racism, sexism, and other forms of bias in AI language models \(4/22/20\)](#)

[AI researchers propose 'bias bounties' to put ethics principles into practice \(4/17/20\)](#)

Cambridge Researchers Tackle Neural Machine Translation's Gender Bias (4/16/20)

Google's AutoML Zero lets the machines create algorithms to avoid human bias (4/14/20)

Researchers find AI is bad at predicting GPA, grit, eviction, job training, layoffs, and material hardship 3/30/20

AI BIAS: A THREAT TO WOMEN'S LIVES? (3/26/20)

Trust at the center: Building an ethical AI framework (3/26/20)

Pardon the Intrusion #13: Policing using AI (3/23/20)

There Is a Racial Divide in Speech-Recognition Systems, Researchers Say (3/23/20)

Should I Worry About... the philosophy behind AI? (3/15/20)

Lie detectors have always been suspect. AI has made the problem worse. (3/13/20)

Could we forgive a machine? Study explores forgiveness in the context of robotics and AI (3/12/20)

ATIH: Natalie Garrett on Ethical Debt and Ethics Education in Tech (3/12/20)

Police Used Facial Recognition to Arrest Over 1,100 People in India Last Month (3/11/20)

A Crisis of Ethics in Technology Innovation (3/10/20)

[International Women's Day: how can algorithms be sexist? \(3/10/20\)](#)

[Algorithms that run our lives are racist and sexist. Meet the women trying to fix them \(3/10/20\)](#)

[Algorithms Learn Our Workplace Biases. Can They Help Us Unlearn Them? \(3/10/20\)](#)

[Maryland's face recognition system is one of the most invasive in the nation \(3/9/20\)](#)

[Why the AI we rely on can't get privacy right \(yet\) \(3/7/20\)](#)

[The case for an AI that puts nature and ethics first, not humans \(3/7/20\)](#)

[Before Clearview Became a Police Tool, It Was a Secret Plaything of the Rich \(3/5/20\)](#)

[Clearview AI: We Are 'Working to Acquire All U.S. Mugshots' From Past 15 Years \(3/5/20\)](#)

[This Small Company Is Turning Utah Into a Surveillance Panopticon \(3/4/20\)](#)

[Humans Are The Cause Of Bias In AI, But We're Also The Solution \(3/3/20\)](#)

[This Is The Year Of AI Regulations \(3/1/20\)](#)

[TOP 8 FUNNIEST AND SHOCKING AI FAILURES OF ALL TIME \(3/1/20\)](#)

[How To Combat The Dark Side Of AI \(2/28/20\)](#)

HOW FACIAL RECOGNITION TECHNOLOGY COULD CHANGECOLLEGE CAMPUSES COMPLETELY (2/28/20)Why is TikTok creating filter bubbles based on your race? (2/28/20)Pope Francis joins IBM and Microsoft in call for AI regulation(2/28/20)FIRST EVER DECISION OF A FRENCH COURT APPLYING GDPR TOFACIAL RECOGNITION (2/27/20)Clearview's Facial Recognition App Has Been Used By The JusticeDepartment, ICE, Macy's, Walmart, And The NBA (2/27/20)Scientists propose new regulatory framework to make AI safer(2/26/20)Clearview AI, The Company Whose Database Has Amassed 3 BillionPhotos, Hacked (2/26/20)This Filter Makes Your Photos Invisible to Facial Recognition (2/26/20)Suckers List: How Allstate's Secret Auto Insurance Algorithm SqueezesBig Spenders (2/25/20)AI Deception: When Your Artificial Intelligence Learns to Lie (2/24/20)Pentagon Adopts New Ethical Principles for Using AI in War (2/24/20)Study finds quarter of climate change tweets from bots (2/22/20)Privacy commissioners launch investigation into facial recognitiontechnology tested by Toronto police, other GTA forces (2/21/20)

[Google launches TensorFlow library for optimizing fairness constraints](#)

(2/21/20)

[Google AI will no longer use gender labels like 'woman' or 'man' on images of people to avoid bias](#) (2/20/20)

[Artificial Intelligence \(AI\) And The Law: Helping Lawyers While Avoiding Biased Algorithms](#) (2/18/20)

[Facial Expressions Don't Tell The Whole Story Of Emotion](#) (2/17/20)

[Mastercard working with transport partners on payments via gait or face biometrics](#) (2/17/20)

[Mastercard is pioneering new payment technology that identifies commuters by the way they walk](#) (2/17/20)

[EU's new AI rules will focus on ethics and transparency](#) (2/17/20)

[AI systems claiming to 'read' emotions pose discrimination risks](#) (2/16/20)

[The Future of Artificial Emotional Awareness](#) (2/13/20)

[Fake news? A.I. algorithm reveals political bias in the stories you read](#) (2/13/20)

[Cost Cutting Algorithms Are Making Your Job Search a Living Hell](#) (2/12/20)

[Attorneys in unemployment fraud cases join forces, call for state review of AI in government](#) (2/12/20)

The end of privacy as we know it (2/10/20)

Public bodies are secretly using AI for decisions on people's lives, warns standards watchdog (2/10/20)

An AI regulation strategy that could really work (2/8/20)

An Algorithm That Grants Freedom, or Takes It Away (2/6/20)

How Algorithmic Bias Hurts People With Disabilities (2/6/20)

Who owns your DNA? You should, according to this biodata bill of rights (2/6/20)

Google's ML-fairness-gym lets researchers study the long-term effects of AI's decisions (2/5/20)

Algorithms on social media need regulation, says UK's AI adviser (2/4/20)

EPIC Asks Federal Trade Commission To Regulate Use Of Artificial Intelligence In Pre-Employment Screenings (2/3/20)

Inside the future of online dating: AI swiping and concierge bots (1/31/20)

Why asking an AI to explain itself can make things worse (1/29/20)

Why Amazon's Ring and facial recognition technology are a clear and present danger to society (1/31/20)

An AI Epidemiologist Sent the First Warnings of the Wuhan Virus (1/25/20)

[New Jersey cops told to halt all use of controversial facial-recognition technology \(1/25/20\)](#)

[AI Can Do Great Things—if It Doesn't Burn the Planet \(1/21/20\)](#)

[The Secretive Company That Might End Privacy as We Know It \(1/18/20\)](#)

[Your online activity is now effectively a social ‘credit score’ \(1/17/20\)](#)

[Siri, Alexa and unconscious bias: the case for designing fairer AI assistants \(1/17/20\)](#)

[There's a new obstacle to landing a job after college: Getting approved by AI \(1/15/20\)](#)

[What Does Fairness in AI Mean? \(1/15/20\)](#)

[Are Your Students Bored? This AI Could Tell You \(1/13/20\)](#)

[Cambridge Votes To Ban Face Surveillance Technology \(1/13/20\)](#)

['Smile with your eyes': How to beat South Korea's AI hiring bots and land a job \(1/13/20\)](#)

[AI can now read emotions – should it? \(1/13/20\)](#)

[Troll Watch: AI Ethics \(1/11/20\)](#)

[Samsung's Neon AI has an ethics problem, and it's as old as sci-fi canon \(1/10/20\)](#)

[Artificial Intelligence Makes Bad Medicine Even Worse \(1/10/20\)](#)

Germany's plans for automatic facial recognition meet fierce criticism

(1/10/20)

Could an AI become your friend? (1/10/20)

Technology Can't Fix Algorithmic Injustice (1/9/20)

San Diego's massive, 7-year experiment with facial recognition

technology appears to be a flop (1/9/20)

THE 'ROBOT TAX' DEBATE HEATS UP (1/8/20)

AI Governance in 2019: A Year in Review in Japan (1/8/20)

Twitter bots and trolls promote conspiracy theories about Australian

bushfires (1/7/20)

The US just released 10 principles that it hopes will make AI safer

(1/7/20)

Dating apps need women. Advertisers need diversity. AI companies

offer a solution: Fake people (1/7/20)

Google's AI beats doctors at detecting breast cancer. (Except when it

doesn't.) (1/6/20)

AIRBNB CLAIMS ITS AI CAN PREDICT WHETHER GUESTS ARE

PSYCHOPATHS (1/4/20)

Rise of #MeTooBots: scientists develop AI to detect harassment in

emails (1/3/2020)

[Illinois says you should know if AI is grading your online job interviews](#)

(1/1/2020)

[Technology Can't Fix Algorithmic Injustice \(Jan 2020\)](#)

## 2019

### Peer-reviewed

[Detect Toxic Content to Improve Online Conversations \(2019\)](#)

[When Hate Speech Leads to Hateful Actions: A Corpus and Discourse](#)

[Analytic Approach to Linguistic Threat Assessment of Hate Speech](#)

(2019)

[Technology on the margins: AI and global migration management](#)

[from a human rights perspective \(Dec 2019\)](#)

[Lessons from Archives: Strategies for Collecting Sociocultural Data in](#)

[Machine Learning \(12/22/19\)](#)

[What's Next for AI Ethics, Policy, and Governance? A Global Overview](#)

(12/18/19)

[It's easy to fool yourself: Case studies on identifying bias and](#)

[confounding in bio-medical datasets \(12/12/19\)](#)

[Explainable Artificial Intelligence \(XAI\): Concepts, taxonomies,](#)

[opportunities and challenges toward responsible AI \(12/12/19\)](#)

[An Unethical Optimization Principle \(11/12/19\)](#)

[Measurement and Fairness \(12/11/19\)](#)

BERT has a Moral Compass: Improvements of ethical and moral values of machines (12/11/19)

Mapping the Ethicality of Algorithmic Pricing: A Review of Dynamic and Personalized Pricing (12/10/19)

Automatically Neutralizing Subjective Bias in Text (12/5/19)

On the Legal Compatibility of Fairness Definitions (11/25/19)

Mapping the Ethicality of Algorithmic Pricing: A Review of Dynamic and Personalized Pricing (11/25/19)

The Second Wave of Algorithmic Accountability (11/25/19)

The Risks of Using AI to Interpret Human Emotions (11/18/19)

The wrong kind of AI? Artificial intelligence and the future of labour demand (11/18/19)

The Debate on the Ethics of AI in Health Care: a Reconstruction and Critical Review (11/13/19)

In bot we trust: A new methodology of chatbot performance measures (11/1/19)

Principles alone cannot guarantee ethical AI (Nov 2019)

Release Strategies and the Social Impacts of Language Models (Nov 2019)

AI SYSTEMS AS STATE ACTORS (Nov 2019)

Evaluating the quality of voice assistants' responses to consumer health questions about vaccines: an exploratory comparison of Alexa, Google Assistant and Siri (Nov 2019)

Evaluating the Factual Consistency of Abstractive Text Summarization (10/28/19)

Fairness and Privacy in AI/ML Systems (10/28/19)

Dissecting racial bias in an algorithm used to manage the health of populations (10/25/19)

Toward a better trade-off between performance and fairness with kernel-based distribution matching (10/25/19)

In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions (10/23/19)

The Woman Worked as a Babysitter: On Biases in Language Generation (10/23/19)

Quantifying the Carbon Emissions of Machine Learning (10/21/19)

An Information-Theoretic Perspective on the Relationship Between Fairness and Accuracy (10/17/19)

Recommender systems and their ethical challenges (10/13/19)

Bias Detect Neutral Network (10/8/19)

A RIGHT TO REASONABLE INFERENCES: RE-THINKING DATA PROTECTION LAW IN THE AGE OF BIG DATA AND AI (10/5/19)

[A fairer way forward for AI in health care \(9/25/19\)](#)

[The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings \(9/25/19\)](#)

[Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead \(9/22/19\)](#)

[Pretrained AI Models: Performativity, Mobility, and Change \(9/7/19\)](#)

[The global landscape of AI ethics guidelines \(9/2/19\)](#)

['Computer says no': was your mortgage application rejected unfairly \(8/8/19\)](#)

[Selection Bias Explorations and Debias Methods for Natural Language Sentence Matching Datasets \(8/2/19\)](#)

[NHS AI Lab: why we need to be ethically mindful about AI for healthcare \(August 2019\)](#)

[Racial Bias in AI Isn't Getting Better and Neither Are Researchers' Excuses \(7/29/19\)](#)

[Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence \(7/18/19\)](#)

[A Causal Bayesian Networks Viewpoint on Fairness \(7/15/19\)](#)

[Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements \(7/13/19\)](#)

Artificial Intelligence Governance and Ethics: Global Perspectives

(7/4/19)

Artificial Intelligence: the global landscape of ethics guidelines

(6/24/19)

A Unified Framework of Five Principles for AI in Society (6/22/19)

“With great power comes great responsibility”: keeping public sector algorithms accountable (6/11/2019)

Notes on bias in the socio-material realization of AI technologies

(June, 2019)

Transparency as design publicity: explaining and justifying inscrutable algorithms (June, 2019)

Detecting Bias with Generative Counterfactual Face Attribute Augmentation (6/18/19)

Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology (6/11/19)

SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (6/9/20)

Evaluating Gender Bias in Machine Translation (6/3/19)

Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies and the Potential of Rational Discourse (Jun 2019)

Racial Bias in Hate Speech and Abusive Language Detection Datasets

(May 2019)

Ethical Dimensions of Visualization Research (May 2019)

Semantics derived automatically from language corpora necessarily contain human biases (5/25/19)

Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical (5/23/19)Tutorial: Socially Responsible Natural Language Processing (5/13/19)

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead (5/13/19)

Limits of Deepfake Detection: A Robust Estimation Viewpoint (5/9/19)

Error terrain analysis for machine learning: Tools and visualizations (5/6/19)

Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture (5/2/19)Evaluating the Underlying Gender Bias in Contextualized Word Embeddings (4/18/19)

Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes (4/3/19)

Biased Language in Patient Medical Records: Detection through Natural Language Processing (NLP) and Impact on Quality (4/2/19)

That's not fair! (April 2019)

The EU-Emotion Voice Database (April 2019)

DATA PROTECTION, ARTIFICIAL INTELLIGENCE AND COGNITIVE SERVICES IS THE GENERAL DATA PROTECTION REGULATION (GDPR)  
"ARTIFICIAL INTELLIGENCE-PROOF" ? (April 2019)

Tay is you. The attribution of responsibility in the algorithmic culture (3/31/19)

On Measuring Social Biases in Sentence Encoders (3/25/19)

Adversarial attacks on medical machine learning (3/22/19)

Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice (3/19/20)

DIRTY DATA, BAD PREDICTIONS: HOW CIVIL RIGHTS VIOLATIONS IMPACT POLICE DATA, PREDICTIVE POLICING SYSTEMS, AND JUSTICE  
(3/5/19)

Big data and discrimination (March 2019)

What We Really Mean When We Say "Ethics" (video) (February 2019)

Predictive Inequity in Object Detection (2/21/19)

Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions (2/14/19)

Is It Good to Cooperate? Testing the Theory of Morality-as-Cooperation in 60 Societies (2/8/19)

## Discrimination in the Age of Algorithms (2/7/19)

Private Accountability in the Age of Artificial Intelligence (2019)

No training required: Exploring random encoders for sentence classification (Jan 2019)

Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure (Jan 2019)

Fairness in representation: quantifying stereotyping as a representational harm (1/28/19)

Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment (1/23/19)

Attenuating Bias in Word Vectors (1/23/19)

Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions (1/23/19)

Identifying and Correcting Label Bias in Machine Learning (1/1/19)

Giving Algorithms a Sense of Uncertainty Could Make Them More Ethical (1/18/19)

Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? (1/7/19)

Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products (Jan 2019)

## Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning (Jan 2019)

### **Government, NGO, and expert publications**

AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement (Dec 2019)

Recommender systems and their ethical challenges (2019)

Beyond Bias: Contextualizing “Ethical AI” Within the History of Exploitation and Innovation in Medical Research (12/21/19)

NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software (12/19/19)

Fairness Indicators: Scalable Infrastructure for Fair ML Systems (12/11/19)

2019 AI Index Report (December 2019)

Disability, Bias, and AI (Nov 2019)

AI 360: Hold, Fold, or Double Down (Nov 2019)

The value of a shared understanding of AI models (Nov 2019)

Supporting rights-respecting AI (11/26/19)

The Second Wave of Algorithmic Accountability (11/25/19)

Opinion: AI For Good Is Often Bad (11/18/19)

The Ethical Dilemma at the Heart of Big Tech Companies (11/14/19)

Delivering the benefits of Custom Neural Voice (11/12/19)

AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense (Nov 2019)

How Machine Learning Pushes Us to Define Fairness (11/6/19)

Expanding Frameworks: An economic justice approach to digital privacy (11/6/19)

Using Tensorflow Fairness Indicators (11/1/19)

Fair ML for Health Resources (Oct 2019)

Artificial Intelligence Research Needs Responsible Publication Norms (10/24/19)

The AI Ethics Charter by Korea Artificial Intelligence Ethics Association (10/23/19)

Introducing AI Blindspot: A Call for Tech to Think Holistically and Spot Risks (10/22/19)

AI-enabled human rights monitoring (10/17/19)

Fairness in Clustering with Multiple Sensitive Attributes (10/11/19)

Austria's employment agency rolls out discriminatory algorithm, sees no problem (10/6/19)

New York City Police Department Surveillance Technology (10/4/19)

This machine read 3.5 million books then told us what it thought

about men and women (9/30/19)

Building ML models for everyone: understanding fairness in machine

learning (9/25/19)

Unpacking “Ethical AI”: A curated reading list (9/24/19)

LITIGATING ALGORITHMS 2019 US REPORT: New Challenges to

Government Use of Algorithmic Decision Systems (9/24/19)

STREET-LEVEL ALGORITHMS AND SEEING THE FOREST FOR THE

TREES IN AI (9/23/19)

Towards Fairer Datasets: Filtering and Balancing the Distribution of the

People Subtree in the ImageNet Hierarchy (9/17/19)

AI ethics and the limits of code(s) (9/16/19)

Face recognition, bad people and bad data (9/9/19)

A Draft Syllabus/Curriculum for an Ethics of AI Course (9/1/19)

How New A.I. Is Making the Law’s Definition of Hacking Obsolete

(8/21/19)

U.S. LEADERSHIP IN AI: A Plan for Federal Engagement in Developing

Technical Standards and Related Tools (8/9/19)

AI Justice: When AI Principles Are Not Enough (8/5/19)

WEF: Responsible Use of Technology (August 2019)

Research Summary: Social Robots and Empathy: The Harmful Effects of Always Getting What We Want (7/27/19)

AI & Global Governance: Human Rights and AI Ethics – Why Ethics Cannot be Replaced by the UDHR (7/19/19)

Gov.UK: Interim report: Review into bias in algorithmic decision-making (7/19/19)

Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns (7/17/19)

Gender diversity in AI research (7/17/19)

Open View: AI/ML Fairness (7/13/19)

IDC Survey Finds Artificial Intelligence to be a Priority for Organizations But Few Have Implemented an Enterprise-Wide Strategy (7/8/19)

Differential Privacy in the 2020 Decennial Census and the Implications for Available Data Products (7/8/19)

The Foundation of Responsible Artificial Intelligence (7/7/19)

AI Differential Privacy and Federated Learning (7/1/19)

Fluxus Landscape: An Expansive View of AI Ethics and Governance (6/26/19)

Five practical steps to make Artificial Intelligence (AI) interpretable (6/21/19)

[Researchers develop 'vaccine' against attacks on machine learning](#)

(6/20/19)

[Ethics of Emotion AI - Part 1 \(6/19/19\)](#)

[The future of women at work: Transitions in the age of automation](#)

(6/12/19)

[Ethical AI: Tensions and trade-offs \(6/11/19\)](#)

[Ethical Codex for Data- Based Value Creation \(6/10/19\)](#)

[AI is not as advanced as you might think \(6/10/19\)](#)

[Introducing the Principled Artificial Intelligence Project \(6/7/19\)](#)

[Does object recognition work for everyone? A new method to assess bias in CV systems \(6/7/19\)](#)

[Deepfakes and Synthetic Media: Updated Survey of Solutions against Malicious Usages \(June 2019\)](#)

[Emotional Artificial Intelligence guidelines for ethical use \(June 2019\)](#)

[Systemic Algorithmic Harms \(5/31/19\)](#)

[Responsible AI: A Global Policy Framework \(5/29/19\)](#)

[Accountable AI: Can the Law Reduce Bias and Increase Transparency? \(5/20/19\)](#)

[Tech workers call time on moving fast and breaking things \(May 2019\)](#)

Designing AI for Social Good: Seven Essential Factors (May 2019)

From What to How: An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices (May 2019)

GARBAGE IN, GARBAGE OUT: FACE RECOGNITION ON FLAWED DATA (5/16/19)

What the Fifth Industrial Revolution is and why it matters (5/15/19)

When Humans Attack: Re-thinking safety, security, and AI (5/14/19)

People + AI Guidebook (May 2019)

World Economic Forum's AI head on how to protect human rights without stifling innovation (4/29/19)

Recommender Systems and their Ethical Challenges (April 2019)

DATA PROTECTION, ARTIFICIAL INTELLIGENCE AND COGNITIVE SERVICES IS THE GENERAL DATA PROTECTION REGULATION (GDPR)  
"ARTIFICIAL INTELLIGENCE-PROOF"? (April 2019)

How AI can enable a sustainable future (April 2019)

Artificial Intelligence and Human Rights (April 2019)

DISCRIMINATING SYSTEMS: Gender, Race, and Power in AI (April 2019)

Gender, Race, and Power in AI (4/17/19)

From Principles to Action: How do we Implement Tech Ethics?

(4/17/19)

Untold History of AI: Algorithmic Bias Was Born in the 1980s

(4/15/19)

The Need to Regulate AI Implementation in Public Assistance Programs (4/12/19)

Laying down the law on AI: ethics done, now the EU must focus on human rights (4/8/19)

The Automated Administrative State (4/8/19)

UK to introduce world first online safety laws (4/8/19)

Artificial Intelligence: Australia's Ethics Framework: A discussion (4/5/19)

AI Ethics: Seven Traps (3/25/19)

Google: Perspectives on Issues in AI Governance (March 2019)

Agile Ethics: Managing Ethical Complexity in Technology (3/14/19)

BENEFITS & RISKS OF ARTIFICIAL INTELLIGENCE (March 2019)

Pentagon Seeks 'Ethical Principles' for AI Use (3/4/19)

Intel's Recommendations for the U.S. National Strategy on Artificial Intelligence (White Paper) (2/6/19)

Ethics of AI in Radiology: European and North American Multisociety Statement (February 2019)

In Favor of Developing Ethical Best Practices in AI Research (2/21/19)

The Environment for Ethical Action (2/18/19)

Council of Europe: Algorithmic Used To Manipulate Social And Political Behaviours (2/18/19)

How the Economics of Data Science is Creating New Sources of Value (2/16/19)

Public deliberation could help address AI's legitimacy problem in 2019 (2/8/19)

Start with values: How innovation can be bold, fast and responsible (1/31/19)

AI in Context: The Labor of Integrating New Technologies (1/30/19)

The ethical and political questions raised by AI (1/29/19)

Illinois Supreme Court Holds That Biometric Privacy Law Does Not Require Actual Harm for Private Suits (1/29/19)

Fairness in representation: quantifying stereotyping as a representational harm (1/28/19)

Ignorance and Distrust Prevail about What Companies and Governments Do with Personal Data (1/25/19)

Automation and artificial intelligence: How machines are affecting people and places (1/24/19)

Building Ethically Aligned AI (1/23/19)

The Era of “Move Fast and Break Things” Is Over [HBR] (1/22/19)

Ethics and technology in the Fourth Industrial Revolution (1/21/19)

Why AI is failing the next generation of women (1/18/19)

Creating AI Systems That Take Culture into Account (1/17/19)

Overcoming barriers to AI adoption (1/16/19)

Facebook Algorithms and Personal Data (Pew Survey) (1/16/19)

Case study: Algorithmic decision-making and accountability (Jan 2019)

Singapore Proposed Model Artificial Intelligence Governance Framework (Jan 2019)

Artificial Intelligence: American Attitudes and Trends (Jan 2019)

Future of work research papers (Jan 2019)

## **News/popular press**

Fight against facial recognition hits wall across the West (12/30/19)

Designing a Moral Compass for AI (12/30/19)

Artificial Intelligence Is Rushing Into Patient Care—And Could Raise Risks (12/24/19)

AI year in review: Opportunities grow, but ethics loom large

(12/23/19)

'The Algorithm Made Me Do It': Artificial Intelligence Ethics Is Still On

Shaky Ground (12/22/19)

AI is biased. Here's how scientists are trying to fix it (12/21/19)

Beyond Bias: Contextualizing "Ethical AI" Within the History of

Exploitation and Innovation in Medical Research (12/20/19)

Federal study of top facial recognition algorithms finds 'empirical

evidence' of bias (12/20/19)

THE INVENTION OF "ETHICAL AI": How Big Tech Manipulates

Academia to Avoid Regulation (12/20/19)

AI Is Biased. Here's How Scientists Are Trying to Fix It (12/19/19)

Ethics In AI: Why Values For Data Matter (12/18/19)

Human Rights Commission wants privacy laws adjusted for an AI

future (12/17/19)

Researchers were about to solve AI's black box problem, then the

lawyers got involved (12/17/19)

Facebook Ads Can Still Discriminate Against Women and Older

Workers, Despite a Civil Rights Settlement (12/13/19)

Emotion recognition technology should be banned, says an AI research

institute (12/13/19)

The AI community needs to take responsibility for its technology and its actions (12/13/19)

AI expert calls for end to UK use of ‘racially biased’ algorithms (12/12/19)

Emotion-detecting tech should be restricted by law - AI Now (12/12/19)

Stanford, Kyoto & Georgia Tech Model ‘Neutralizes’ Biased Language (12/11/19)

The Utility of Interpretable AI (12/10/19)

Microsoft tech expert warns of bias and sexism in artificial intelligence (12/10/19)

As AI moves into content creation, researchers aim to battle its biases (12/9/19)

How practitioners and academics think (and then forget) about fairness when building AI systems (12/6/19)

Biased Algorithms Are Easier to Fix Than Biased People (12/6/19)

Life Under the Algorithm: How a relentless speedup is reshaping the working class (12/4/19)

Human insight remains essential to beat the bias of algorithms (12/4/19)

UK GDPR Watchdog: Explain Your AI (12/3/19)

Portland plans to propose the strictest facial recognition ban in the country (12/2/19)

Chinese tech groups shaping UN facial recognition standards (12/1/19)

Dealing With Bias in Artificial Intelligence (Nov 2019)

An epidemic of AI misinformation (11/30/19)

Explain yourself, mister: Fresh efforts at Google to understand why an AI system says yes or no (11/29/19)

The first effort to regulate AI was a spectacular failure (11/26/19)

Tainted Data Can Teach Algorithms the Wrong Lessons (11/25/19)

Human biases are baked into algorithms. Now what? (11/25/19)

Scientists developed a new AI framework to prevent machines from misbehaving (11/25/19)

Can We Force AIs to Be Fair Towards People? Scientists Just Invented a Way (11/25/19)

Tuning Out Toxic Comments, With the Help of AI (11/21/19)

Stanford, UMass Amherst develop algorithms that train AI to avoid specific misbehaviors (11/21/19)

Researchers Want Guardrails to Help Prevent Bias in AI (11/21/19)

Dealing with bias in AI (11/19/19)

The Apple Card Didn't 'See' Gender—and That's the Problem

(11/19/19)

There's an easy way to make lending fairer for women. Trouble is, it's illegal. (11/15/19)

The Real AI Threat to Democracy (11/14/19)When Algorithms Decide Whose Voices Will Be Heard (11/12/19)

Google's secret cache of medical data includes names and full details of millions – whistleblower (11/12/19)

These Black Women Are Fighting For Justice In A World Of Biased Algorithms (11/12/19)

Apple co-founder Steve Wozniak says Apple Card discriminated against his wife (11/11/19)

'You sound worried': would you let an AI rephrase the tone of your emails? (11/6/19)

AI Can Make Bank Loans More Fair (11/6/20)

Rights group files federal complaint against AI-hiring firm HireVue, citing 'unfair and deceptive' practices (11/6/19)

Artificial Intelligence Can Be Biased. Here's What You Should Know. (11/5/19)

Microsoft's AI Research Draws Controversy Over Possible Disinformation Use (11/4/19)

How to train artificial intelligence that won't destroy the environment

(11/4/19)

AI is making literary leaps – now we need the rules to catch up

(11/2/19)

Pentagon's draft AI ethics guidelines fight bias and rogue machines

(11/2/19)

AI-Related Lawsuits Are Coming (11/1/20)

NYPD Built Bias Safeguards Into Pattern-Spotting AI System

(10/31/19)

Using AI to Eliminate Bias from Hiring (10/29/19)

Big Data and Racial Bias: Can That Ghost Be Removed from the Machine? (10/28/19)

Why did Microsoft fund an Israeli firm that surveils West Bank Palestinians? (10/28/19)

The hidden costs of AI (10/26/19)

A health care algorithm affecting millions is biased against black patients (10/24/19)

Artificial Intelligence Research Needs Responsible Publication Norms

(10/24/19)

How much “fake news” is on Twitter?

When Binary Code Won't Accommodate Nonbinary People

(10/23/19)

AN ALGORITHM FOR DE-BIASING AI SYSTEMS (10/23/19)After Backlash, Portland-Area Outback Steakhouse Cancels New Artificial Intelligence Surveillance Test (10/23/19)A face-scanning algorithm increasingly decides whether you deserve the job (10/22/19)These Startups Are Building Tools to Keep an Eye on AI (10/21/19)Military artificial intelligence can be easily and dangerously fooled (10/21/19)Facial recognition: This new AI tool can spot when you are nervous or confused (10/21/19)New framework makes AI systems more transparent without sacrificing performance (10/21/19)Fighting hate with AI-powered retorts (10/19/19)Can you make AI fairer than a judge? Play our courtroom algorithm game (10/17/19)Facial recognition AI can't identify trans and non-binary people (10/16/19)Zombie debts are hounding struggling Americans. Will you be next? (10/15/19)

Bias and Algorithmic Fairness: The modern business leader's new responsibility in a brave new world ruled by data. (10/13/19)

Ethical AI: Why governments, companies are a worried lot (10/10/19)

Rasa's conversational AI can selectively ignore bits of dialogue to improve its responses (10/9/19)

Beware of automated hiring (10/8/19)

How to operationalize AI ethics (10/7/19)

Austria's employment agency rolls out discriminatory algorithm, sees no problem (10/6/19)

Google causing more facial recognition problems, machine learning goes quantum and losing a job if an AI doesn't like your face (10/6/19)

Artificial stupidity: 'Move slow and fix things' could be the mantra AI needs (10/5/19)

An Open Source License That Requires Users to Do No Harm (10/4/19)

Congress Is Investigating the Military's Use of Facial Recognition (10/3/19)

The Unsettling Ways Tech Is Changing Your Personal Reality (10/3/19)

Bias and Algorithmic Fairness (10/3/19)

Blind Spots in AI Just Might Help Protect Your Privacy (10/2/19)

[British workers are deliberately sabotaging robots amid fears they will take their jobs, study finds \(9/29/19\)](#)

["The New Jim Code" – Ruha Benjamin on racial discrimination by algorithm \(9/26/19\)](#)

[Reddit and Gab's most toxic communities inadvertently train AI to combat hate speech \(9/25/19\)](#)

[Africa Is Building an A.I. Industry That Doesn't Look Like Silicon Valley \(9/25/19\)](#)

[The Viral App That Labels You Isn't Quite What You Think \(9/19/20\)](#)

[See how an AI system classifies you based on your selfie \(9/17/19\)](#)

[AI ethics and the limits of code\(s\) \(9/16/19\)](#)

[120 million workers will need to be retrained due to AI, says IBM study \(9/6/19\)](#)

[An AI-Run World Needs to Better Reflect People of Color \(9/6/19\)](#)

['Sense of urgency', as top tech players seek AI ethical rules \(9/2/19\)](#)

[How to Build an AI Ethics Committee \(8/30/19\)](#)

[Silicon Valley's Secret Philosophers Should Share Their Work \(8/27/19\)](#)

['Dangerous' AI offers to write fake news \(8/27/19\)](#)

[Amazon, Microsoft, 'putting world at risk of killer AI': study \(8/21/19\)](#)

Flawed Algorithms Are Grading Millions of Students' Essays (8/20/19)

An AI privacy conundrum? The neural net knows more than it says (8/19/19)

AI can read your emotions. Should it? (8/17/19)

'Stereotyping' emotions is getting in the way of artificial intelligence. Scientists say they've discovered a better way. (8/16/19)

Do Tech Companies Really Need to Snoop Into Private Conversations to Improve Their A.I.? (8/14/19)

Training the next generation of ethical techies (8/14/19)

Training bias in AI "hate speech detector" means that tweets by Black people are far more likely to be censored (8/14/19)

What your voice reveals about you (8/13/19)

Google's Artificial Intelligence Hate Speech Detector Has a 'Black Tweet' Problem (8/13/19)

Google's algorithm for detecting hate speech is racially biased (8/13/19)

GOOGLE'S HATE SPEECH-DETECTING AI IS BIASED AGAINST BLACK PEOPLE (8/12/19)

Facial recognition software mistook 1 in 5 California lawmakers for criminals, says ACLU (8/12/2019)

IBM Research launches explainable AI toolkit (8/8/19)

## AI Needs Your Data—and You Should Get Paid for It (8/8/19)

'Computer says no': was your mortgage application rejected unfairly  
(8/8/19)

## AI Ethics Guidelines Every CIO Should Read (8/7/19)

China has started a grand experiment in AI education. It could reshape  
how the world learns (8/2/19)

Artificial intelligence could improve healthcare for all—unless it doesn't  
(8/1/19)

## WE TESTED EUROPE'S NEW LIE DETECTOR FOR TRAVELERS – AND IMMEDIATELY TRIGGERED A FALSE POSITIVE (7/26/19)

AI 'emotion recognition' can't be trusted (7/25/19)

Fitbits and other wearables may not accurately track heart rates in  
people of color (7/24/19)

Microsoft and the learnings from its failed Tay artificial intelligence bot  
(7/24/19)

## HIGH-STAKES AI DECISIONS NEED TO BE AUTOMATICALLY AUDITED (7/18/19)

How companies can avoid ethics washing (7/17/19)

Facial Recognition Tech Is Growing Stronger, Thanks to Your Face  
(7/13/19)

Ethical Artificial Intelligence Becomes A Supreme Competitive Advantage (7/7/19)

Botched Adidas Twitter Campaign Spits out Racist and Anti-Semitic Arsenal Jerseys (7/2/19)

What Does an AI Ethicist Do? (6/24/19)

AN AI “VACCINE” CAN BLOCK ADVERSARIAL ATTACKS (6/20/19)

AI ain’t for everyone – who trusts bots, and why (6/14/29)

Credit Scores Could Soon Get Even Creepier and More Biased (6/13/19)

San Francisco says it will use AI to reduce bias when charging people with crimes (6/12/19)

THE NEXT BIG PRIVACY HURDLE? TEACHING AI TO FORGET (6/12/19)

AI is worse at identifying household items from lower-income countries (6/11/19)

Illinois Bill Aims to Limit Use of AI in Video Job Interviews (6/11/19)

How Big Tech funds the debate on AI ethics (6/6/19)

How companies like Google are dealing with the ethics of AI (6/4/19)

Stop saying "robots are coming for your job"; start saying "Your boss wants to replace you with a robot" (6/3/19)

[Ethics as a competitive advantage in the booming artificial intelligence industry \(6/3/19\)](#)

[This Creepy AI Predicts What You Look Like Based on Your Voice \(6/1/19\)](#)

[Artificial intelligence, the future of work, and inequality \(5/31/19\)](#)

[Addressing the Biases Plaguing Algorithms \(5/31/19\)](#)

[10 things we should all demand from Big Tech right now \(5/29/19\)](#)

[AI is only human \(5/28/19\)](#)

[To Fight Deepfakes, Researchers Built a Smarter Camera \(5/28/19\)](#)

[Amazon venturing into emotionally-savvy Artificial Intelligence at the cost of your Privacy \(5/28/19\)](#)

[When algorithms mess up, the nearest human gets the blame \(5/28/19\)](#)

[Top 5 Tools Data Scientists Can Use To Mitigate Biases In Algorithms \(5/28/19\)](#)

[America and its economic allies have announced five “democratic” principles for AI \(5/22/19\)](#)

[The Best Reason for Your City to Ban Facial Recognition \(5/17/19\)](#)

[Voice Recognition Still Has Significant Race and Gender Biases \(5/10/19\)](#)

The Legislation That Targets the Racist Impacts of Tech (5/7/19)

All the ways hiring algorithms can introduce bias (5/6/19)

Who to sue when a robot loses your fortune (5/5/19)

The danger of AI is weirder than you think (April 2019)

Why your board needs an AI council (4/28/19)

WILL ARTIFICIAL INTELLIGENCE ENHANCE OR HACK HUMANITY?  
(4/28/19)

FOR THE SUCCESSFUL ADOPTION OF AI, WE NEED MORE WOMEN  
LEADERS (4/24/19)

How AI can help close the gender pay gap and eliminate bias  
(4/24/19)

Microsoft is winning the techlash (4/24/19)

Shazeda Ahmed on the Messy Truth About Social Credit (4/23/19)

Ethics committee raises alarm over 'predictive policing' tool (4/20/19)

How killer robots overran the UN (4/19/19)

Some AI just shouldn't exist (4/19/19)

Emotionally intelligent AI will respond to how you feel (4/17/19)

A.I. Bias Isn't the Problem. Our Society Is (4/16/19)

[Google reportedly disbands review panel monitoring DeepMind Health AI \(4/15/19\)](#)

[Discrimination's Digital Frontier \(4/15/19\)](#)

[What is fair when it comes to AI bias? \(4/12/19\)](#)

[Do no evil: why we need a public conversation about AI ethics \(4/12/19\)](#)

[A new bill would force companies to check their algorithms for bias \(4/10/19\)](#)

[Microsoft's Brad Smith on How to Responsibly Deploy AI \(4/10/19\)](#)

[AI's leading developers share their fears about the tech developing too fast \(4/9/19\)](#)

[Why Artificial Intelligence Needs Democratic Governance \(4/9/19\)](#)

[How will AI change your life? AI Now Institute founders Kate Crawford and Meredith Whittaker explain. \(4/8/19\)](#)

[It's disturbingly easy to trick AI into doing something deadly \(4/8/19\)](#)

[Tech titans declare AI ethics concerns \(4/8/19\)](#)

[What we lost when we lost Google ATEAC \(4/7/19\)](#)

[UK businesses using artificial intelligence to monitor staff activity \(4/7/19\)](#)

[Real or artificial? Tech titans declare AI ethics concerns \(4/7/19\)](#)

[Hey Google, sorry you lost your ethics council, so we made one for you](#)

(4/6/19)

[Facebook's ad-serving algorithm discriminates by gender and race](#)

(4/5/19)

["Color-blindness" is a bad approach to solving bias in algorithms](#)

(4/3/19)

[Algorithms have gotten out of control. It's time to regulate them.](#)

(4/3/19)

[The problem with AI ethics \(4/3/19\)](#)

[The Pitfalls of Data's Gender Gap \(4/1/19\)](#)

['Bias deep inside the code': the problem with AI 'ethics' in Silicon Valley](#)

(3/29/19)

[Inmates in Finland are training AI as part of prison labor \(3/28/19\)](#)

[RISK ASSESSMENT: EXPLAINED \(3/27/19\)](#)

[Rethinking Privacy For The AI Era \(3/27/19\)](#)

[Ethical question takes center stage at Silicon Valley summit on artificial intelligence \(3/27/19\)](#)

[How malevolent machine learning could derail AI \(3/25/19\)](#)

[CAN AI BE A FAIR JUDGE IN COURT? ESTONIA THINKS SO \(3/25/19\)](#)

[How Pope Francis could shape the future of robotics \(3/24/19\)](#)

Human Contact Is Now a Luxury Good (3/23/19)

Our Software Is Biased Like We Are. Can New Laws Change That? (3/23/19)

The Dog That Did Not Bark In the Night: The Danger of Survivor Bias in AI (3/22/19)

Deciding how to decide: Six key questions for reducing AI's democratic deficit (3/22/19)

Warnings of a Dark Side to A.I. in Health Care (3/12/19)

When AI speaks on behalf of humans: Proposing ethical guidelines based on Google Duplex assistant (3/12/19)

AI, ethics and data bias - getting it right (3/20/19)

The world's first genderless AI voice is here. Listen now (3/19/19)

To Be Ethical, AI Must Become Explainable. How Do We Get There? (3/19/19)

As AI Spreads, Tech Needs 'Chief Bias Officers' (3/18/19)

Tackling Bias in Machine Learning (3/18/19)

Courts and police departments are turning to AI to reduce bias, but some argue it'll make the problem worse (3/17/19)

Is Microsoft AI Helping to Deliver China's 'Shameful' Xinjiang Surveillance State? (3/15/19)

Without Humans, A.I. Can Wreak Havoc (3/12/19)

The NYPD Now Has an Automated Tool to Help Them Recognize Patterns in Crimes (3/10/19)

The Achilles' Heel Of AI (3/7/19)

Does AI Ethics Have A Bad Name? (3/7/19)

Don't look now: why you should be worried about machines reading your emotions (3/6/19)

A new study finds a potential risk with self-driving cars: failure to detect dark-skinned pedestrians (3/6/19)

The infamous AI gaydar study was repeated – and, no, code can't tell if you're straight or not just from your face (3/5/19)

Should we be treating algorithms the same way we treat hazardous chemicals (3/5/19)

AI will reproduce and enshrine age-old biases—if we let it (3/5/19)

How AI Will Rewire Us (3/5/19)

Some self-driving car systems have trouble detecting darker skin, study says (3/5/19)

China's tech billionaires back ethical rules to guide development of AI and other technologies (3/4/19)

Seeking Ground Rules for A.I. (3/4/19)

[Apple is hiring an analyst to explain Siri complaints to executives](#)

(3/4/19)

[16 Uncomfortable Questions Everyone Needs to Ask About Artificial Intelligence](#) (3/3/19)

[Is Ethical A.I. Even Possible?](#) (3/1/19)

[Why AI is a threat to democracy—and what we can do to stop it](#)

(2/26/19)

[Chinese police test gait-recognition technology from AI start-up Watrix that identifies people based on how they walk](#) (2/26/19)

[Troubling Trends Towards Artificial Intelligence Governance](#) (2/25/19)

[AI researchers debate the ethics of sharing potentially harmful programs](#) (2/21/19)

[China Uses DNA to Track Its People, With the Help of American Expertise](#) (2/21/19)

[Pope Francis and Microsoft team up to promote prize for artificial intelligence](#) (2/19/19)

[OpenAI: Social science, not just computer science, is critical for AI](#) (2/19/19)

[Over 2000 people died after receiving Centrelink robo-debt notice, figures reveal](#) (2/18/19)

[AI in the UK: The only way is ethics](#) (2/18/19)

Bentham, Hobbes, and The Ethics of Artificial Intelligence (2/17/19)

Are the algorithms that power dating apps racially biased? (2/17/19)

New AI fake text generator may be too dangerous to release, say creators (2/14/19)

THE REAL REASON TECH STRUGGLES WITH ALGORITHMIC BIAS (2/12/19)

GOOGLE AND MICROSOFT WARN THAT AI MAY DO DUMB THINGS (2/11/19)

YouTube announces it will no longer recommend conspiracy videos (2/10/19)

How Vodafone's Chatbot Technology Is Helping It Cut Jobs (2/8/19)

Women Stand Against Social Injustice In AI (2/7/19)

I Cut the 'Big Five' Tech Giants From My Life. It Was Hell (2/7/19)

Microsoft warns investors that its artificial-intelligence tech could go awry and hurt its reputation (2/6/19)

Goodbye, trolley problem. This is Silicon Valley's new ethics test. (2/5/19)

SUPPOSEDLY 'FAIR' ALGORITHMS CAN PERPETUATE DISCRIMINATION (2/5/19)

Microsoft warned investors that biased or flawed AI could hurt the company's image (2/5/19)

## Five Takeaways from an AI Fairness Conference (2/2/19)

JUSTICE BY THE NUMBERS: MEET THE STATISTICIAN TRYING TO FIX BIAS IN CRIMINAL JUSTICE ALGORITHMS (2/1/19)

Guidelines for human-AI interaction design (2/1/19)

Automated background checks are deciding who's fit for a home (2/1/19)

Getting efficient with “What-happens-if ...” (2/1/19)

A.I. Could Worsen Health Disparities (1/31/19)

Google’s head of translation on fighting bias in language and why AI loves religious texts (1/30/19)

This Is Your Brain Off Facebook (1/30/19)

Prisons Are Building Databases of Inmates' Voice Prints (1/30/19)

PRISONS ACROSS THE U.S. ARE QUIETLY BUILDING DATABASES OF INCARCERATED PEOPLE'S VOICE PRINTS (1/30/19)

New York Insurers Can Evaluate Your Social Media Use—if They Can Prove Why It's Needed (1/30/19)

San Francisco proposal would ban government facial recognition use in the city (1/29/19)

ML Integrity: Four Production Pillars For Trustworthy AI (1/29/19)

IBM releases Diversity in Faces, a dataset of over 1 million annotations to help reduce facial recognition bias (1/29/19)

The Unnatural Ethics of AI Could Be Its Undoing (1/29/19)

How to lift the veil off hidden algorithms (1/28/19)

SECURITY ISN'T ENOUGH. SILICON VALLEY NEEDS 'ABUSABILITY' TESTING (1/28/19)

Being Ethical: How to Beat Bias in AI and Be Transparent (1/28/19)

MIT hopes to automatically 'de-bias' face detection AI (1/27/19)

Speaking Black Dialect in Courtrooms Can Have Striking Consequences (1/25/19)

Bias in, bias out: the Stanford scientist out to make AI less white and male (1/25/19)

The Hidden Automation Agenda of the Davos Elite (1/25/19)

MIT researchers: Amazon's Rekognition shows gender and ethnic bias (updated) (1/24/19)

Amazon Is Pushing Facial Technology That a Study Says Could Be Biased (1/24/19)

Quarter of U.S. jobs could be jeopardized by AI, research shows (1/24/19)

A.I. and Automation Will Hit Low-Skill Jobs and Trump Swing States Hardest (1/24/19)

[Workers in heartland states most at risk of losing jobs to AI, new study finds \(1/24/19\)](#)

[Automation is a bigger threat to inland California workers, study finds \(1/24/19\)](#)

[Can we make artificial intelligence ethical? \(1/23/19\)](#)

[Artificial Intelligence and Ethics \(Jan/Feb 2019\)](#)

[The AI Arms Race Means We Need AI Ethics \(1/22/19\)](#)

[He Said, She Said: Addressing Gender in Neural Machine Translation \(1/22/19\)](#)

[What makes AI ethicists “the top hire companies need to succeed”? \(1/21/19\)](#)

[A.I. Is Sending People to Jail—and Getting It Wrong \(1/21/19\)](#)

[The goal is to automate us': welcome to the age of surveillance capitalism \(1/20/19\)](#)

[What makes AI ethicists “the top hire companies need to succeed”? \(1/21/19\)](#)

[Why Is AI And Machine Learning So Biased? The Answer Is Simple Economics \(1/20/19\)](#)

[Giving algorithms a sense of uncertainty could make them more ethical \(1/18/19\)](#)

[How This Chicago Company Is Using Predictive Analytics To Decrease Bias \(1/17/19\)](#)

US military trusted more than Google, Facebook to develop AI: survey

(1/14/19)

The weaponization of artificial intelligence (1/14/19)

AI can wreak havoc if left unchecked by humans (1/14/19)

Facial and emotional recognition; how one man is advancing artificial intelligence (1/13/19)

One day your voice will control all your gadgets, and they will control you (1/11/19)

How a Google Researcher Is Making AI Easier to Understand (1/10/19)

Americans want to regulate AI but don't trust anyone to do it (1/10/19)

Never mind killer robots—here are six real AI dangers to watch out for in 2019 (1/7/19)

The Future Of Recruiting In The Age of Automation And Artificial Intelligence (1/4/19)

2019 Is the Year to Stop Talking About Ethics and Start Taking Action (1/4/19)

How a Feel-Good AI Story Went Wrong in Flint (1/3/19)

New App Helps People Remember Faces (1/1/19)

# 2018

## Peer-reviewed

IMPLEMENTING ETHICS INTO ARTIFICIAL INTELLIGENCE: A CONTRIBUTION, FROM A LEGAL PERSPECTIVE, TO THE DEVELOPMENT OF AN AI GOVERNANCE REGIME (2018)

Ethics as an Escape from Regulation: From ethics-washing to ethics shopping (2018)

The Bias Bias in Behavioral Economics (12/31/18)

Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making (Dec 2018)

Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions (12/16/18)

Improving fairness in machine learning systems: What do industry practitioners need? (12/13/18)

Racial Influence on Automated Perceptions of Emotions (12/6/18)

Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data (12/1/18)

Artificial Intelligence in the Context of Crime and Criminal Justice (Dec 2018)

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science (Dec 2018)

Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved (11/27/18)

50 Years of Test (Un)fairness: Lessons for Machine Learning

(11/25/18)

An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and RecommendationsEven If AI Can Cure Loneliness – Should It? (11/9/18)Some Requests for Machine Learning Research from the East African Tech Scene (11/8/18)Fairness and Abstraction in Sociotechnical Systems (11/7/18)Consumer-Lending Discrimination in the Era of FinTech (Oct 2018)The dangers of automating social programs (Oct 2018)AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations (10/28/18)Learning Adversarially Fair and Transferable Representations (10/22/18)Gender-Aware Natural Language Translation (10/8/18)Model Cards for Model Reporting (10/5/18)Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations (10/1/18)Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice (September 2018)

Language Modeling Teaches You More Syntax than Translation Does:

Lessons Learned Through Auxiliary Task Analysis (9/26/18)

TRANSPARENCY AND EXPLANATION IN DEEP REINFORCEMENT

LEARNING NEURAL NETWORKS (9/17/18)

World Economic Forum - The Future of Work 2018 Report (Sept 2018)

A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity (9/10/18)

Assessing Gender Bias in Machine Translation -- A Case Study with Google Translate (9/6/18)

How AI can be a force for good (8/24/18)

Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices (8/19/18)

The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning (8/14/18)

Building Safer AGI by introducing Artificial Stupidity (8/11/18)

A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics (7/2/18)

GENDER SHADES: INTERSECTIONAL ACCURACY DISPARITIES IN COMMERCIAL GENDER CLASSIFICATION (July 2018)

A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions (July 2018)

Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction (6/30/18)

A review of possible effects of cognitive biases on interpretation of rule-based machine learning models (6/27/18)

Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure (6/15/18)

AI and the economy (6/13/18)

Fairness definitions explained (May 2018)

Why is my classifier discriminatory? (5/30/18)

ENHANCING THE ACCURACY AND FAIRNESS OF HUMAN DECISION MAKING (5/25/18)

To Build Truly Intelligent Machines, Teach Them Cause and Effect (5/15/18)

The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards (5/9/18)

Addressing Age-Related Bias in Sentiment Analysis (April 2018)

Let's talk about race: Identity, chatbots, and AI (April 2018)

Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making (4/21/18)

Modeling and Simultaneously Removing Bias via Adversarial Neural Networks (4/18/18)

Word embeddings quantify 100 years of gender and ethnic stereotypes (4/17/18)

Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions (4/10/18)

It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process (3/29/18)

DATASHEETS FOR DATASETS (3/23/18)

Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science (2/23/18)

DELAYED IMPACT OF FAIR MACHINE LEARNING (3/12/18)

A Reductions Approach to Fair Classification (3/6/18)

Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making (2/318)

Measuring and Mitigating Unintended Bias in Text Classification (2/1/18)

AN AI RACE FOR STRATEGIC ADVANTAGE: RHETORIC AND RISKS (February 2018)

Transparent Model Distillation (1/26/18)

Mitigating Unwanted Biases with Adversarial Learning (1/22/18)

The accuracy, fairness, and limits of predicting recidivism (1/17/18)

AI: from rational agents to socially responsible agents (2018)

## Government, NGO, and expert publications

Discrimination, artificial intelligence, and algorithmic decision-making (2018)

Detecting Bias in Amazon reviews (2018)

AI Index 2018 Annual Report (December 2018)

AI Narratives Report (12/11/18)

The AI Maturity Playbook (12/11/18)

Artificial Intelligence and the Future of Humans (Pew Survey) (12/10/18)

AI Now 2018 Report (12/6/18)

AI Now 2018 Recommendations (12/6/18)

AI Fairness for People with Disabilities: Point of View (11/26/18)

Public Attitudes Toward Computer Algorithms (Pew Survey) (11/16/18)

Gender, Race and Power: Outlining a New AI Research Agenda (11/15/18)

AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations (11/2/18)

Portrayals and perceptions of AI and why they matter (Nov 2018)

## Applying AI for Social Good: McKinsey Report (Nov 2018)

### WHY DOES ARTIFICIAL INTELLIGENCE DISCRIMINATE? (10/24/18)

Declaration on Ethics and Data Protection in Artificial Intelligence by  
40th International Conference of Data Protection and Privacy  
Commissioners (10/23/18)

Big data and AI: Ethical and societal implications (10/16/18)

Governing Artificial Intelligence: Upholding Human Rights & Dignity  
(10/10/18)

In Advanced and Emerging Economies Alike, Worries About Job  
Automation (Pew: Survey) (9/13/18)

Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System (Sept 2018)

Cognitive diversity: AI and the future of work (Sept 2018)

KPMG: AI Risk and Controls Matrix (Sept 2018)

Playing with fairness (Sept. 2018)

An ethics checklist for data scientists

Brookings survey finds divided views on artificial intelligence for  
warfare, but support rises if adversaries are developing it (8/29/18)

Factsheets for AI Services (8/22/18)

Ethical OS: A guide to anticipating the future impact of today's technology (2018)

Humans Plus Robots: Why the Two Are Better Than Either One Alone (7/12/18)

Critical Perspectives on Artificial Intelligence and Human Rights (6/19/18)

Data ethics framework (6/13/18)

Analyzing & Preventing Unconscious Bias in Machine Learning (6/12/18)

How good are Google's new AI ethics principles? (6/7/18)

Google's AI Principles (6/6/18)

Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems (6/5/18)

The State of Machine Learning (ML) Adoption in the Enterprise - O'Reilly Survey (June 2018)

The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems (5/16/18)

Designing ethically pt. 2 (5/4/18)

Designing ethically pt. 1 (5/2/18)

Skill shift: Automation and the future of the workforce (May 2018)

Addressing Age-Related Bias in Sentiment Analysis (4/21/18)

Algorithmic Accountability: A Primer (4/18/18)

Algorithmic impact assessments: A practical framework for public agency accountability (4/9/18)

How to Prevent Discriminatory Outcomes in Machine Learning (3/12/18)

PERSISTENT SURVEILLANCE'S CYNICAL ATTEMPT TO PROFIT OFF BALTIMORE'S TRAUMA (3/1/18)

21 definitions of fairness and their politics [\[pdf\]](#) [\[video\]](#) (2/20/18)

The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation (2/20/18)

Face Off: Law Enforcement Use of Face Recognition Technology (2/12/18)

The problem with building a “fair” system (1/30/18)

Eight Futures of Work: Scenarios and their Implications (January 2018)

## **News/popular press**

The Verge 2018 tech report card: AI (12/30/18)

Are AI toys ethical? (12/27/18)

The Welfare State Is Committing Suicide by Artificial Intelligence (12/25/18)

The case for taking AI seriously as a threat to humanity (12/23/18)

HELLO 2019: 3 Predictions On How AI Will Learn and Impact Society (12/21/18)

The prison-reforming First Step Act has a critical software bug (12/21/18)

'Kill your foster parents': Amazon's Alexa talks murder, sex in AI experiment (12/21/18)

The rise of A.I. could hurt women's careers in a major way (12/21/18)

Amazon error allowed Alexa user to eavesdrop on another home (12/20/18)

AI makers get political (12/19/18)

Google AI Principles updates, six months in (12/18/18)

Super Human or Less Human? (12/16/18)

It's time for a Bill of Data Rights (12/14/18)

A.I. 'bias' could create disastrous results, experts are working out how to fight it (12/14/18)

Nine charts that really bring home just how fast AI is growing (12/12/18)

The invisible workers of the AI era (12/12/18)

[Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret \(12/10/18\)](#)

[2018 in Review: 10 AI Failures \(12/10/18\)](#)

[Fixing Bias In Algorithms Is Possible, And This Scientist Is Doing It \(12/9/18\)](#)

[120 predictions for AI for 2019 \(12/9/18\)](#)

[The Seductive Diversion of 'Solving' Bias in Artificial Intelligence \(12/8/18\)](#)

[Bank of America Tech Chief Defines Responsible AI Projects \(12/5/18\)](#)

[Big tech has your kid's data – and you probably gave it to them \(12/5/18\)](#)

[Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias \(Dec 2018\)](#)

[Technologist Vivienne Ming: 'AI is a human right' \(12/7/18\)](#)

[New Research Reveals 75 Percent of Customers Still Favour Live Agent Support for Customer Service vs 25 Percent Self-Service and Chatbots \(12/6/18\)](#)

[Are Silicon Valley's Autoreplies Taking Over Our Minds? \(12/6/18\)](#)

[Facial recognition has to be regulated to protect the public, says AI report \(12/6/18\)](#)

[Facial recognition: It's time for action \(12/6/18\)](#)

AI has a culturally-biased worldview that Google has a plan to change

(12/2/18)

10 Fundamental Insights about the Tech-Driven Future for Humanity

and why women, POC, and other underrepresented people in tech

should lead it (11/29/18)

Why AI Needs To Reflect Society (11/29/18)

INSTAGRAM WILL USE AI TO DESCRIBE IMAGES FOR VISUALLY

IMPAIRED USERS (11/28/18)

10 Most Important AI Research Papers for 2018 (11/27/18)

UK police wants AI to stop violent crime before it happens (11/26/18)

Tech giants offer empty apologies because users can't quit (11/26/18)

THE DANGEROUS JUNK SCIENCE OF VOCAL RISK ASSESSMENT

(11/25/18)

How Cheap Labor Drives China's A.I. Ambitions (11/25/18)

Wanted: The 'perfect babysitter.' Must pass AI scan for respect and

attitude (11/23/18)

"Siri, is artificial intelligence biased?" (11/20/18)

How to use science fiction to teach tech ethics (11/19/18)

One of the Fathers of AI Is Worried About Its Future (11/19/18)

[A new DNA sequencing service wants to reward you for sharing your data \(11/15/18\)](#)

[Google 'betrays patient trust' with DeepMind Health move \(11/14/18\)](#)

[Harvard researchers want to school Congress about AI \(11/14/18\)](#)

[If You Drive in Los Angeles, the Cops Can Track Your Every Move \(11/13/18\)](#)

[Don't Believe Every AI You See \(11/13/18\)](#)

[Algorithmic Bias or Fairness: The Importance of the Economic Context \(11/13/18\)](#)

[In the Age of A.I., Is Seeing Still Believing? \(11/12/18\)](#)

[The newest Jim Crow \(11/8/18\)](#)

[Breaking down AI's trustability challenges \(11/8/18\)](#)

[Google Workers' Walkout Signals Crisis of Faith in Company Culture \(11/2/18\)](#)

[How to Stop Technology From Becoming a Digital Frankenstein \(10/31/18\)](#)

[Should a Self-Driving Car Kill the Baby or the Grandma? Depends on Where You're From \(10/29/18\)](#)

[4 human-caused biases we need to fix for machine learning \(10/27/18\)](#)

The Digital Gap Between Rich and Poor Kids Is Not What We Expected

(10/26/18)

A Dark Consensus About Screens and Kids Begins to Emerge in Silicon

Valley (10/26/18)

IBM explores the intersection of AI, ethics—and Pac-Man (10/25/18)

The AI cold war that threatens us all (10/23/18)

No Innovation Without Representation (10/23/18)

Companies are on the hook if their hiring algorithms are biased

(10/22/18)

No, A.I. Won't Solve the Fake News Problem (10/20/18)

When AI Misjudgment Is Not an Accident (10/19/18)

An Ethical Awakening in AI (10/18/18)

Weapons of Micro Destruction: How Our 'Likes' Hijacked Democracy

(10/17/18)

When Alexa Can't Understand You (10/16/18)

Black Mirror, Light Mirror: Teaching Technology Ethics Through

Speculation (10/15/18)

MICROSOFT'S NADELLA SAYS AI CAN MAKE THE WORLD MORE

INCLUSIVE (10/15/18)

[WIRED25: Ethical AI: Intel's Genevieve Bell On Living with Artificial Intelligence \(10/15/18\)](#)

[A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI \(10/11/18\)](#)

[Building trust in AI applications \(10/11/18\)](#)

[Amazon scraps secret AI recruiting tool that showed bias against women \(10/9/18\)](#)

[WHEN TECH KNOWS YOU BETTER THAN YOU KNOW YOURSELF \(10/4/18\)](#)

[Think computers are less biased than people? Think again. \(10/3/18\)](#)

[Why Technology Favors Tyranny \(Oct. 2018\)](#)

[Build a minimum ethical product \(9/30/18\)](#)

[5 takeaways on the state of AI from Disrupt SF \(9/28/18\)](#)

[Senators introduce the 'Artificial Intelligence in Government Act' \(9/26/18\)](#)

[Taking algorithms to court \(9/24/18\)](#)

[Senators are asking whether artificial intelligence could violate US civil rights laws \(9/21/18\)](#)

[Artificial intelligence hates the poor and disenfranchised \(9/21/18\)](#)

[Artificial Intelligence: The Robots Are Now Hiring \(9/20/18\)](#)

IBM launches cloud tool to detect AI bias and explain automated decisions (9/19/18)

The exploitation, injustice, and waste powering our AI (9/18/18)

How AI is already affecting jobs (9/17/18)

The Human Promise of the AI Revolution (9/14/18)

High Time to Regulate Face Recognition A.I. (9/13/18)

What Algorithms Know About You Based on Your Grocery Cart (9/13/18)

Google Knows Where You've Been, but Does It Know Who You Are? (9/12/18)

Many voice assistant users have 'trust issues' with their device, study claims (9/12/18)

Ethics + Data Science (9/10/18)

Underneath all the AI hype is the likelihood it threatens the poor, says this former Microsoft and Google exec (9/8/18)

Left Unchecked, Artificial Intelligence Can Become Prejudiced All On Its Own (9/7/18)

Artificial Intelligence is greater concern than climate change or terrorism, says new head of British Science Association (9/6/18)

From headlines to headway: Defining data ethics and its impact on data workers (9/5/18)

IBM collaborated with the NYPD on an AI system that can search for people by race (9/6/18)

Everyday Ethics for Artificial Intelligence (9/5/18)

Don't believe the algorithm (9/5/18)

Silicon Valley Thinks Everyone Feels the Same Six Emotions (9/5/18)

California just replaced cash bail with algorithms (9/4/18)

The New AI Tech Turning Heads in Video Manipulation (9/3/18)

Japan developing 'pre-crime' artificial intelligence to predict money laundering and terror attacks (9/1/18)

How Not to Teach Ethics (Sept 2018)

Franken-algorithms: The Deadly Consequences of Unpredictable Code (8/30/18)

Now is the time to start thinking about AI's impact on xenophobia (8/28/18)

Watson is helping heal America's broken criminal-sentencing system (8/25/18)

To Build Trust In Artificial Intelligence, IBM Wants Developers To Prove Their Algorithms Are Fair (8/22/18)

Technologists are trying to fix the "filter bubble" problem that tech helped create (8/22/18)

Who needs democracy when you have data? (8/20/18)

Bias In Maternal AI Could Hurt Expectant Black Mothers (8/17/18)

AI could make dodgy lip sync dubbing a thing of the past (8/17/18)

AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind (8/16/18)

The Role of Trust in an Era of AI Bots (8/15/18)

Children are susceptible to peer pressure from robots (8/15/18)

Yuval Noah Harari on what the year 2050 has in store for humankind (8/12/18)

Confronting demons of the computer age (8/12/18)

Tech companies use “persuasive design” to get us hooked.

Psychologists say it's unethical (8/8/18)

What HBR Gets Wrong About Algorithms and Bias (8/7/18)

Case studies in data ethics (8/7/18)

Why artificial intelligence will have very human frailties (8/3/18)

Using artificial intelligence to fix Wikipedia's gender problem (8/3/18)

Facial Recognition Is the Perfect Tool for Oppression (8/2/18)

Five of the scariest predictions about artificial intelligence (8/1/18)

The U.K. Wants to Become the World Leader in Ethical A.I. (8/1/18)

Fostering a Human-Centered Approach to Artificial Intelligence

(7/31/18)

Data's day of reckoning (7/31/18)

The Future of Artificial Intelligence Depends on Trust (7/31/18)

Microsoft's politically correct chatbot is even worse than its racist one

(7/31/18)

This VR Founder Wants to Gamify Empathy to Reduce Racial Bias

(7/20/18)

The accent gap (7/19/18)

Health Insurers Are Vacuuming Up Details About You – And It Could

Raise Your Rates (7/17/18)

Alexa alternatives have a secret weapon: Privacy (7/14/18)

UX for AI: Trust as a Design Challenge (7/12/18)

The Economics Of Artificial Intelligence - How Cheaper Predictions Will

Change The World (7/10/18)

'I was shocked it was so easy': meet the professor who says facial

recognition can tell if you're gay (1/7/18)

London police chief 'completely comfortable' using facial recognition

with 98 percent error rate (7/5/18)

## 4 ways AI can change democracy for the better (7/4/18)

Technology Alone Can't Preserve Endangered Languages (6/30/18)

BIAS EVERYWHERE: Tech companies just woke up to a big problem with their AI (6/30/18)

AI, Ain't I A Woman? (6/28/18)

When impact measures fail (And what to do about it) (6/26/18)

Techstars: AI startups must be wary of 'move fast and break things' mantra (6/26/18)

Facial recognition software is not ready for use by law enforcement (6/25/18)

Unmasking A.I.'s Bias Problem (6/25/18)

Bias detectives: the researchers striving to make algorithms fair (6/20/18)

Bing researchers develop a novel way of collecting high-quality AI training data (6/18/18)

UK report warns DeepMind Health could gain 'excessive monopoly power' (6/15/18)

A Tough Week for Tech Workers, and It Won't Be the Last (6/15/18)

Plum uses AI to hire people 'that never would have been discovered through a traditional hiring process' (6/13/18)

[When it Comes to AI and Weapons, the Tech World Needs Philosophers \(6/12/18\)](#)

[This tool lets you see—and correct—the bias in an algorithm \(6/12/18\)](#)

[Tackling the Ethical Challenges of Slippery Technology \(6/11/18\)](#)

[Accenture wants to beat unfair AI with a professional toolkit \(6/9/18\)](#)

[Silicon Valley insiders say Facebook, Snapchat, and Twitter are using 'behavioural cocaine' to turn people into addicts \(6/5/18\)](#)

[The digital poorhouse \(6/4/18\)](#)

[Are you scared yet? Meet Norman, the psychopathic AI \(6/2/18\)](#)

[Artificial Intelligence: Have No Fear \(June 2018\)](#)

[China has turned Xinjiang into a police state like no other \(5/31/18\)](#)

[Does China's digital police state have echoes in the West? \(5/31/18\)](#)

[Pymetrics open-sources Audit AI, an algorithm bias detection tool \(5/31/18\)](#)

[Machine Un-Learning: Why Forgetting Might Be the Key to AI \(5/31/18\)](#)

[Who is writing the future? Designing infrastructure for ethical AI \(5/31/18\)](#)

[The \(holy\) ghost in the machine: Catholic thinkers tackle the ethics of artificial intelligence \(5/26/18\)](#)

Navigating the risks of artificial intelligence and machine learning in low income countries (5/24/18)

Enhanced AI Tools, Enhanced AI Trust (5/24/18)

How Advances in AI and Automation Will Upend Our Traditional Models of Economic Development (5/20/18)

Finkle floats an ethical AI plan (5/18/18)

Artificial Intelligence: What's Human Rights Got To Do With It? (5/14/18)

Algorithms are making the same mistakes assessing credit scores that humans did a century ago (5/14/18)

Society needs a reboot for the Fourth Industrial Revolution (5/14/18)

Aided by Palantir, the LAPD used predictive policing to monitor specific people and neighborhoods (5/11/18)

This company audits algorithms to see how biased they are (5/9/18)

Facebook says it has a tool to detect bias in its artificial intelligence (5/3/18)

Data Violence and How Bad Engineering Choices Can Damage Society (4/30/18)

Revealed: how bookies use AI to keep gamblers hooked (4/30/18)

PwC: Lack of trust in AI assistants like Alexa could hinder adoption (4/29/18)

[It's time to address the reproducibility crisis in AI \(4/24/18\)](#)

[What you need to know about Facebook and ethics \(4/19/18\)](#)

[Artificial intelligence must be 'for common good' \(4/16/18\)](#)

[When algorithms surprise us \(4/13/18\)](#)

[Microsoft has given up 'significant sales' over concerns that the customer will use AI for evil, says a top scientist \(4/13/18\)](#)

[Democracy vs. the Algorithm \(4/12/18\)](#)

[Human bias is a huge problem for AI. Here's how we're going to fix it \(4/10/18\)](#)

[Drawing the Ethical Line on Weaponized Deep Learning Research \(4/5/18\)](#)

[Leading AI researchers threaten Korean university with boycott over its work on 'killer robots' \(4/4/18\)](#)

[7 Short-Term AI ethics questions \(4/4/18\)](#)

[A.I. Engineers Must Open Their Designs To Democratic Control \(4/2/18\)](#)

[What happens when an algorithm cuts your health care \(3/21/18\)](#)

[OpenAI Wants to Make Safe AI, but That May Be an Impossible Task \(3/15/18\)](#)

[What Developers Really Think About AI And Bias \(3/15/18\)](#)

Fun fact of the day: Voice recognition tech is naturally sexist (3/14/18)

The case for fairer algorithms (3/14/18)

Google's DeepMind Has An Idea For Stopping Biased AI (3/13/18)

How to Make A.I. That's Good for People (3/7/18)

The ART of AI—Accountability, Responsibility, Transparency (3/4/18)

Is Artificial Intelligence the Ultimate Test for Privacy? (3/2/18)

New Report on Emerging AI Risks Paints a Grim Future (2/21/18)

Interpreting machine learning models (2/20/18)

"We're in a diversity crisis": cofounder of Black in AI on what's poisoning algorithms in our lives (2/14/18)

Google's DeepMind Has An Idea For Stopping Biased AI (2/13/18)

He Predicted The 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse. (2/11/18)

Facial Recognition Is Accurate, if You're a White Guy (2/9/18)

SHOULD DATA SCIENTISTS ADHERE TO A HIPPOCRATIC OATH? (2/8/18)

Dealing with Imbalanced Classes in Machine Learning (2/2/18)

WHY DO WE KEEP GENDERING OUR AI ASSISTANTS? (1/26/18)

[Algorithms are making American inequality worse \(1/26/18\)](#)

[Sorry, Alexa Is Not a Feminist \(1/24/18\)](#)

[A Popular Algorithm Is No Better at Predicting Crimes Than Random People \(1/17/18\)](#)

[A CHILD ABUSE PREDICTION MODEL FAILS POOR FAMILIES \(1/15/18\)](#)

[Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech \(1/12/18\)](#)

[Do algorithms reveal sexual orientation or just expose our stereotypes? \(1/11/18\)](#)

[Why bots go bad: Curbing transgressive tendencies in AI \(1/3/18\)](#)

[Can an Algorithm Tell When Kids Are in Danger? \(1/2/18\)](#)

[The Impact of Machine Learning on Economics \(Jan 2018\)](#)

[Artificial Intelligence Trends to Watch in 2018 - CB Insights](#)

## 2017

### Peer-reviewed

[Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences \(12/2/17\)](#)

[Are Algorithms Building the New Infrastructure of Racism? How we use big data can reinforce our worst biases—or help fix them \(12/2/17\)](#)

Avoiding discrimination through causal reasoning (12/2017)

Ethical issues in research using datasets of illicit origin (Nov 2017)

The real risks of artificial intelligence (2017)

Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation (10/17/17)

Is that social bot behaving unethically? (Sept 2017)

Toward algorithmic transparency and accountability (Sept 2017)

Algorithmic decision making and the cost of fairness (8/13/17)

Human decisions and machine predictions (2017)

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints (7/29/17)

Decoupled classifiers for fair and efficient machine learning (2017)

Algorithmic decision making and the cost of fairness (6/10/17)

Responses to Critiques on Machine Learning of Criminality Perceptions (5/26/17)

Semantics derived automatically from language corpora necessarily contain human biases (5/25/17)

Evaluating Quality of Chatbots and Intelligent Conversational Agents (4/15/17)

Ethical by Design: Ethics Best Practices for Natural Language

Processing (4/4/17)

Goal-Oriented Design for Ethical Machine Learning and NLP (4/4/17)

Social Bias in Elicited Natural Language Inferences (April 2017)

Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. (April 2017)

Decision Making With Quantized Priors Leads to Discrimination (Feb 2017)

Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration (2017)

Refractive Surveillance: Monitoring Customers to Manage Workers (submitted 2/8/17)

## **Government, NGO, and expert publications**

IEEE Standards Association Ethically Aligned Design guidelines v2 (12/2017)

Montreal Responsible AI Declaration (11/17)

WHAT'S NEXT FOR MACHINE LEARNING: ETHICS AND EXPLAINABILITY IN AI FOR FRAUD (11/2017)

AI Now 2017 Report (10/2017)

Bots Won't Just Help Us Buy Stuff. They'll Help Us Become Better Versions of Ourselves (6/1/17)

## Physiognomy's New Clothes (5/6/17)

When Good Intentions Backfire ...And why we need a hacker mindset (2/15/17)

## **News/popular press**

ATIH: Natalie Garrett on Ethical Debt and Ethics Education in Tech (3/12/20)

Can we teach morality to machines? Three perspectives on ethics for artificial intelligence (12/19/17)

Inside China's vast new experiment in social ranking (12/14/17)

New York City's Bold, Flawed Attempt to Make Algorithms Accountable (12/12/17)

The trouble with bias (12/10/17)

ARTIFICIAL INTELLIGENCE SEEKS AN ETHICAL CONSCIENCE (12/7/17)

Researchers Combat Gender and Racial Bias in Artificial Intelligence (12/4/17)

Global AI Dialogue Series: Observations from the China-US Workshop in Beijing (12/2/2017)

Ethics at scale (11/29/17)

Google Translate's gender bias pairs "he" with "hardworking" and "she" with "lazy, and other examples (11/29/17)

Can A.I. Be Taught to Explain Itself? (11/21/17)

This Chatbot Is Trying Hard To Look And Feel Like Us (11/15/17)

Trustworthy AI? Yes, by design (11/8/17)

Who's the fairest of them all? Not AI (11/8/17)

When Data Science Destabilizes Democracy and Facilitates Genocide (11/2/17)

The 10 Top Recommendations for the AI Field in 2017 (10/18/17)

What Silicon Valley gets wrong about universal basic income (10/16/17)

Can we teach robots ethics? (10/15/17)

Asking the right questions about AI (10/11/17)

Forget killer robots - Bias is the real AI danger (10/3/17)

Artificial Empathy Systems and Design Research at Scale (10/2/17)

Artificial intelligence is about the people, not the machines (9/30/17)

How to recognize exclusion in AI (9/26/17)

Instagram uses 'I will rape you' post as Facebook ad in latest algorithm mishap (9/21/17)

Artificial intelligence can make our societies more equal. Here's how (9/21/17)

Training soft skills into AI technology (9/19/17)

AI research is in desperate need of an ethical watchdog (9/18/17)

New AI can guess whether you're gay or straight from a photograph (9/7/17)

How to regulate artificial intelligence (9/1/17)

Turns out Algorithms are racist (8/31/17)

The IRS Is Mining Taxpayer Data On Social Media In Violation Of Federal Privacy Law (8/28/17)

New app scans your face and tells companies whether you're worth hiring (8/25/17)

Machines taught by photos learn a biased view of women (8/21/17)

AI's biggest challenge is human, not technological. (8/19/17)

AI programs are learning to exclude African-American voices (8/16/17)

Teaching AI systems to behave themselves (8/13/17)

Why we desperately need women to design AI (8/4/17)

Biased AI is a threat to civil liberties. The ACLU has a plan to fix it. (7/25/17)

Technology Is Biased Too. How Do We Fix It? (7/20/17)

Why Google's PAIR initiative to take bias out of AI will never be complete (7/18/17)

AI and the future of ethics (7/1/17)

Volvo admits its self-driving cars are confused by kangaroos (6/30/17)

When a Computer Program Keeps You in Jail (6/13/17)

To live in harmony with AI we must create a modern Magna Carta (7/6/17)

Corporate Surveillance in Everyday Life (June 2017)

A discussion about AI's conflicts and challenges (6/17/17)

Elon Musk and linguists say that AI is forcing us to confront the limits of human language (6/14/19)

Facebook taught bots to negotiate (and lie) like humans (6/14/17)

This backflipping noodle has a lot to teach us about AI safety (6/14/17)

Don't Grade Teachers With a Bad Algorithm (5/15/17)

The dark secrets at the heart of AI (4/11/17)

Will Using Artificial Intelligence To Make Loans Trade One Kind Of Bias For Another? (3/31/17)

We need to talk about Accessibility on Chatbots (1/30/17)

## Algorithms and bias: What lenders need to know (1/20/17)

# 2016

## Peer Reviewed

Policing Predictive Policing (2016)

Man is to computer programmer as woman is to homemaker?

Debiasing word embeddings (2016)

Equality of Opportunity in Supervised Learning (10/7/2016)

The Other Question: Socialbots and the Question of Ethics

(12/1/2016)

Designing AI systems that obey our laws and values (2016)

Generating Visual Explanations (Oct 2016)

On the (im)possibility of fairness (9/23/16)

The Ethics of Algorithms: Mapping the Debate (July 2016)

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

(2/16/18)

A Confidence-Based Approach for Balancing Fairness and Accuracy

(1/22/16)

## Government, NGO, and expert publications

A Guide to Solving Social Problems with Machine Learning (12/8/16)

Racism and inequity are products of design. They can be redesigned.

(11/15/16)

Big Data's Disparate Impact (2016)

## News/popular press

Facebook Privacy: Social Network Buys Data From Third-Party Brokers

To Fill In User Profiles (12/28/16)

Top 10 AI failures of 2016 (12/2/2016)

Facebook Lets Advertisers Exclude Users by Race (10/28/16)

The ethics of good design: A principle for the connected age

(11/20/16)

The perpetual line-up (10/18/16)

There is a blind spot in AI (10/13/16)

A beauty contest was judged by AI and the robots didn't like dark skin

(9/8/2016)

How algorithms rule our working lives (9/1/16)

Chicago's predictive policing tool just failed a major test (8/19/16)

Artificial Intelligence Has a 'Sea of Dudes' Problem (6/23/16)

Machine Bias (5/23/2016)

Is Amazon same-day delivery service racist? (4/23/16)

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day (3/24/2016)

Uber seems to offer better service in areas with more white people (3/10/16)

## 2015

Race effects on eBay (Oct 2015)

Women less likely to be shown ads for high-paid jobs on Google, study shows (7/28/15)

Questioning the Fairness of Targeting Ads Online (7/7/15)

Automated Experiments on Ad Privacy Settings (4/18/15)

What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. (2015)

## 2014

Social attributions from faces bias human choices (Nov 2014)

How big data is unfair (9/26/14)

How to Build a Strong Ethical Culture at Your [Government] Agency (7/14/14)

## 2013

Data brokers selling lists of rape victims, AIDS patients (12/19/13)

## Behavioral Signal Processing: Deriving Human Behavioral Informatics From Speech and Language (9/11/13)

Linguistic Models for Analyzing and Detecting Biased Language (August 2013)

Discrimination in Online Ad Delivery (1/28/13)

## **2012**

Fairness through awareness (Jan 2012)

## **2011**

A Framework for Ethical Reasoning (12/2/11)

Testing a Bayesian Measure of Representativeness Using a Large Image Database (2011)

## **2010**

Perceptual cues in non-verbal vocal expressions of emotion (4/29/10)

## **2009**

HP Investigates Claims of 'Racist' Computers (12/22/09)

Camera Misses the Mark on Racial Sensitivity (5/19/09)

## **2008**

Understanding Information Ethics and Policy: Integrating Ethical Reflection and Critical Thinking into Policy Development (2008)

# 2004

## Ethical Fading: The Role of Self-Deception in Unethical Behavior

(7/2004)



Content

Terms

Privacy

Cookie Preferences

Press

---

© Copyright 2020 Salesforce.com, inc. All rights reserved. Rights of ALBERT EINSTEIN are used with permission of The Hebrew University of Jerusalem. Represented exclusively by Greenlight.

# The Contests of Tech Ethics: A Sociotechnical Approach to Ethics and Technology in Action

Ben Green  
[bzgreen@umich.edu](mailto:bzgreen@umich.edu)  
Michigan Society of Fellows  
Gerald R. Ford School of Public Policy

## Abstract

Recent controversies related to topics such as fake news, privacy, and algorithmic bias have prompted increased public scrutiny of digital technologies and soul-searching among many of the people associated with their development. In response, the tech industry, academia, civil society, and governments have rapidly increased their attention to “ethics” in the design and use of digital technologies (“tech ethics”). Yet almost as quickly as ethics discourse has proliferated across the world of digital technologies, the limitations of these approaches have also become apparent: tech ethics is vague and toothless, is subsumed into corporate logics and incentives, and has a myopic focus on individual engineers and technology design rather than on the structures and cultures of technology production. As a result of these limitations, many have grown skeptical of tech ethics and its proponents, charging them with “ethics-washing”: promoting ethics research and discourse to defuse criticism and government regulation without committing to ethical behavior. By looking at how ethics has been taken up in both science and business in superficial and depoliticizing ways, I recast tech ethics as a terrain of contestation where the central fault line is not *whether* it is desirable to be ethical, but *what* “ethics” entails and *who* gets to define it. This framing highlights the significant limits of current approaches to tech ethics and the importance of studying the formulation and real-world effects of tech ethics. In order to identify and develop more rigorous strategies for reforming digital technologies and the social relations that they mediate, I describe a sociotechnical approach to tech ethics, one that reflexively applies many of tech ethics’ own lessons regarding digital technologies to tech ethics itself.

## Table of Contents

1	Introduction: The Crisis of Conscience .....	3
2	The Rise of Tech Ethics .....	6
2.1	Tech Industry .....	7
2.2	Academia .....	7
2.3	Civil Society.....	8
2.4	Government.....	9
3	The Limits of Tech Ethics.....	10
3.1	Tech ethics principles are abstract and toothless .....	10
3.2	Tech ethics is subsumed into corporate logics and incentives.....	11
3.3	Tech ethics has a myopic focus on individual engineers and technology design .....	14
3.4	Tech ethics has become an avenue for ethics-washing.....	16
4	The Contestation of Tech Ethics .....	18
4.1	Ethics in Science .....	20
4.2	Corporate Ethics and Co-optation.....	21
4.3	The Future of Tech Ethics.....	23
5	A Sociotechnical Approach to Tech Ethics .....	24
5.1	Determinism.....	25
5.2	Solutionism .....	25
5.3	Objectivity and Neutrality.....	26
5.4	Sociotechnical Systems.....	27

## **1 Introduction: The Crisis of Conscience**

If digital technology production in the beginning of the 2010s was characterized by the brash spirit of Facebook's motto “move fast and break things” and the superficial assurances of Google’s motto “don’t be evil,” digital technology toward the end of the decade was characterized by a “crisis of conscience” (Marantz, 2019) about these and other technologies’ perils. While many of digital technology’s harms were visible early in the decade (not to mention long before), it was not until stories of this technology causing harm reached a critical mass that they became salient in the public eye. Consider just a few of the cases that have prompted this crisis of conscience within tech and the associated “techlash”—the growing animosity of the public toward major technology companies—which in 2018 was deemed by both Oxford Dictionaries (Oxford Languages, 2018) and the Financial Times (Foroohar, 2018) to be one of the words of the year.

*Fake News:* Throughout the 2016 U.S. presidential election between Donald Trump and Hillary Clinton, social media was plagued with fraudulent stories that went viral: “FBI confirms evidence of huge underground Clinton sex network,” “Hillary sold weapons to ISIS,” and more (Emery Jr., 2016; Ritchie, 2016). Social media platforms—Facebook in particular—came under scrutiny for their milquetoast response to fake news, allowing these stories to spread widely without providing any indication that they were not true. Given that fake news spreads further, faster, and more broadly than true news (Vosoughi, Roy, and Aral, 2018), social media business models of prioritizing user engagement appear to be partially responsible for the spread of fake news. In turn, numerous commentators—including Hillary Clinton herself—blamed Facebook for Donald Trump’s presidential election victory (Blake, 2018; Graham, 2017; Read, 2016; Solon, 2016). Later reporting revealed that Facebook’s leadership actively resisted sharing more information about Russian efforts to propagate fake news, instead prioritizing the company’s business strategies (Perlroth, Frenkel, and Shane, 2018).

*Cambridge Analytica:* In 2018, *The New York Times* and *The Guardian* reported that the voter-profiling firm Cambridge Analytica had harvested information from the Facebook profiles of more than 50 million (later revealed to be 87 million) people, without their knowledge or permission, in order to target political ads to benefit Donald Trump’s 2016 presidential campaign (Cadwalladr and Graham-Harrison, 2018; Lapowsky, 2018; Rosenberg, Confessore, and Cadwalladr, 2018).

Cambridge Analytica had acquired this data by exploiting the sieve-like nature of Facebook's privacy policy. These revelations raised new questions about how carefully Facebook protects user data and privacy: despite having learned about this data harvesting by 2015, the company did not alert users and took only cursory actions to protect the data from further misuse. After the Cambridge Analytica story was reported, Congress summoned Mark Zuckerberg to testify about Facebook's practices (Kang et al., 2018) and a concerted effort arose among Facebook users to delete their profiles (Hsu, 2018).

*Military and ICE Contracts:* In 2018, the technology website Gizmodo revealed that Google was working with the U.S. Department of Defense (DoD) to develop artificial intelligence software that could analyze drone footage (Cameron and Conger, 2018). This effort, known as Project Maven, was part of a \$7.4 billion investment in AI by the DoD in 2017 (Cameron and Conger, 2018) and represented an opportunity for Google to gain billions of dollars in future defense contracts (Tiku, 2019). The project set off intense debate within Google, as many engineers expressed concern about facilitating drone strikes and began organizing to assert that "Google should not be in the business of war" (Shane and Wakabayashi, 2018). Project Maven, along with similar stories about tech companies partnering with the Trump administration—such as reports that Palantir was developing software for Immigration and Customs Enforcement (ICE) to facilitate deportations (Woodman, 2017)—prompted new organizing among tech workers and computer science students in opposition to tech industry contracts with U.S. defense and intelligence agencies, often centered around the slogans #TechWontBuildIt and #NoTechForICE (Conger and Metz, 2018; Glaser and Oremus, 2018; Goldberg, 2020; Mijente, 2019).

*Algorithmic Bias:* In 2016, ProPublica revealed that an algorithm used in criminal courts was biased against Black defendants, mislabeling them as future criminals at twice the rates of white defendants (Angwin et al., 2016). The dispute about whether the algorithm in question was, in fact, biased (Dieterich, Mendoza, and Brennan, 2016; Kleinberg, Mullainathan, and Raghavan, 2016) helped build interest in the emerging field of "algorithmic fairness." Through books such as Cathy O'Neil's *Weapons of Math Destruction* (O'Neil, 2017), Virginia Eubanks' *Automating Inequality* (Eubanks, 2018a), and Safiya Noble's *Algorithms of Oppression* (Noble, 2018) as well as articles demonstrating the biases in algorithms used in contexts ranging from facial recognition

(Buolamwini and Gebru, 2018) to hiring (Dastin, 2018), the public began to recognize algorithms as both fallible and discriminatory—potentially the source of more harm than good.

These and other tech-related controversies were a shock to many both inside and outside the world of technology, as they arrived in an era of widespread (elite) optimism about the beneficence of technology. Yet these controversies also brought public attention to what scholars in fields such as STS<sup>1</sup>, philosophy of science, critical data and algorithm studies, information science, human-computer interaction, and law have long argued: technology is shaped by social forces, technology structures society in often deleterious ways, and technology cannot solve every social problem. Broadly speaking, these fields bring a “sociotechnical” approach to studying technologies (digital and otherwise), analyzing how technologies shape, are shaped by, and interact with society. The sociotechnical frame emphasizes the ways in which social actors and technological artifacts are intertwined as part of unified—rather than discrete—systems and serve to simultaneously and mutually constitute one another (Bijker and Law, 1994; Jasanoff, 2004; Suchman et al., 1999). As tech scandals mounted, these fields’ critical and sociotechnical insights, long ignored by most technologists and technology journalists, were newly recognized (or in some form recreated).

Many in the tech sector and academia diagnosed these ills as the result of an inattention to ethics: a lack of training in ethical reasoning for engineers and a dearth of ethical principles in engineering practice, which in turn led to the development of unethical technologies (Fiesler, 2018b; Karoff, 2019; Marantz, 2019; Raicu, 2017; Zunger, 2018). In response, academics, technologists, companies, governments, and more have embraced a broad set of goals often characterized with the label “tech ethics”: the introduction of considerations around ethics and social responsibility into digital technology education, research, development, use, and governance. In the span of just a few years, tech ethics has become a dominant discourse discussed in technology companies, academia, civil society organizations, and governments. For those committed to combating an array of injustices connected to digital technologies, the rise of tech ethics has produced a range of responses: on the one hand, excitement that technologists are increasingly considering their social responsibilities and impacts; on the other hand, frustration regarding the limited scope and impacts of what tech ethics discourse and practice has thus far entailed.

---

<sup>1</sup> Referring to the related areas of a) science and technology studies and b) science, technology, and society.

This essay summarizes these developments and debates in tech ethics. I first describe the primary forms of tech ethics and summarize the central critiques made against these efforts, which focus on tech ethics' abstract and toothless nature, corporate logics, narrow and individualistic focus, and capacity for strategic use ("ethics-washing"). Against the backdrop of these debates, I then turn to describing tech ethics itself as a terrain of contestation, where the central debate is not over *whether* ethics is desirable but over *what* ethics entails and *who* obtains the authority to define it. These debates suggest the need for a sociotechnical approach to tech ethics that explores the social construction and real-world effects of tech ethics. I introduce such an approach through four frames: determinism, solutionism, objectivity and neutrality, and sociotechnical systems.

## 2 The Rise of Tech Ethics

Although scholars, activists, and others have long considered the ethics and social impacts of technology, attention to developing and promoting technology ethics has rapidly grown across the tech industry, academia, civil society, and government. Across these sectors, one common treatment of tech ethics is through statements of ethical principles. One analysis of 36 prominent AI principles documents showed the rapid rise in these statements, from 2 in 2014 to 16 in 2018 (Fjeld et al., 2020). These documents tend to cover very similar themes, particularly fairness and non-discrimination, privacy, accountability, and transparency and explainability (Fjeld et al., 2020). Many of the documents also reference human rights, with some taking international human rights as the framework for analyzing and promoting tech ethics (Fjeld et al., 2020).

The broad and diverse nature of the work connected with tech ethics can make it difficult to define where the boundaries of "tech ethics" begin and end. While many people and institutions directly embrace the label, others are pursuing related efforts without any direct reference to tech ethics. As tech ethics becomes a more widespread frame, however, it is often used as a catchall to refer to any effort to study or improve the social impacts of technology. And although sometimes directly tied to the philosophical discipline of ethics (i.e., moral philosophy), the "ethics" in tech ethics is more frequently tied to applied forms of ethics such as codes of ethics, research ethics, and the lived impacts of digital technologies. In order to evaluate the term and the debates that surround it, I will largely restrict my discussions of "tech ethics" to those people and organizations

that explicitly embrace the label (and its close analogues or derivatives, such as AI ethics and algorithmic fairness).

## *2.1 Tech Industry*

The most pervasive treatment of tech ethics within tech companies has come in the form of ethics principles and ethics oversight bodies. Companies like Microsoft, Google, and IBM have developed and publicly shared AI ethics principles, which include statements such as “AI systems should treat all people fairly” and “AI should [...] Be socially beneficial” (IBM, 2018; Microsoft, 2018; Pichai, 2018). These principles are often supported through dedicated ethics teams and advisory boards within companies, with such bodies in place at companies including Microsoft, Google, Facebook, Alphabet subsidiary DeepMind, and policing technology company Axon (Legassick and Harding, 2017; Nadella, 2018; Novet, 2018; Vincent and Brandom, 2018; Walker, 2018). Companies such as Google, Accenture, and Clifford Chance have also begun offering tech ethics consulting services (Simonite, 2020; Accenture, n.d.; Clifford Chance, n.d.).

As part of these efforts, the tech industry has formed several coalitions aimed at advancing a common dialogue about safe and ethical artificial intelligence. In 2015, Elon Musk and Sam Altman (the then-president of the tech incubator Y Combinator) created OpenAI, an independent research organization that aims to develop socially beneficial artificial intelligence and mitigate the “existential threat” presented by AI, with more than \$1 billion in donations from major tech executives and companies (Dowd, 2017). A year later, Amazon, Facebook, DeepMind, IBM, and Microsoft founded the Partnership on AI (PAI), a nonprofit coalition to shape best practices in AI development, advance public understanding of AI, and support socially beneficial applications of AI (Finley, 2016; Hern, 2016).

## *2.2 Academia*

Computer and information science programs at universities have rapidly increased their emphasis on ethics training in curricula. While some universities have taught computing ethics courses (within both computer science and other fields) for many years (Grosz et al., 2019; Reich et al., 2020; Shilton et al., 2017), the emphasis on ethics within computing education has increased dramatically in recent years (Fiesler, Garrett, and Beard, 2020). When information science

professor Casey Fiesler tweeted a link to an editable spreadsheet of tech ethics classes in November 2017, it quickly grew to a crowdsourced list of more than 200 courses (Fiesler, 2018a). This plethora of courses represents a dramatic shift in computer science training and culture, with ethics becoming a popular topic of discussion and study after being largely ignored by the mainstream of the field just a few years prior.

Research in computer science and related fields has also become increasingly focused on the ethics and social impacts of computing. This trend is observable in the recent explosion of conferences and workshops related to computing ethics. The ACM Conference on Fairness, Accountability, and Transparency (FAccT, formerly FAT\*) and the AAAI/ACM Conference on AI, Ethics, and Society (AIES) both held their first annual meetings in February 2018 and have since grown rapidly. Through 2019, there had been more than 30 workshops related to fairness and ethics at major computer science conferences (ACM FAccT Conference, 2020). Many universities have supported these curricular and research efforts through the creation of new institutes focused on the social implications of technology. 2017 alone saw the launch of the AI Now Institute at NYU (AI Now Institute, 2017), the Princeton Dialogues on AI and Ethics (Sharlach, 2019), and the MIT/Harvard Ethics and Governance of Artificial Intelligence Initiative (MIT Media Lab, 2017). More recently formed centers include the MIT College of Computing (MIT News Office, 2018); the Stanford Institute for Human-Centered Artificial Intelligence (Adams, 2019); and the University of Michigan Center of Ethics, Society, and Computing (Marowski, 2020).

### *2.3 Civil Society*

Numerous civil society organizations have also coalesced around tech ethics, with strategies that include grantmaking and developing principles and toolkits. Organizations such as the MacArthur and Ford Foundations have begun exploring and making grants in tech ethics (Robinson and Bogen, 2016). For instance, the Omidyar Network, Mozilla Foundation, Schmidt Futures, and Craig Newmark Philanthropies partnered on the Responsible Computer Science Challenge, which awarded \$3.5 million between 2018 and 2020 to support efforts to embed ethics into undergraduate computer science education (Mozilla, 2018). Many foundations also contribute to the research, conferences, and institutes that have emerged in recent years.

Other organizations have been created or have expanded their scope to consider the implications and governance of digital technologies. For example, the American Civil Liberties Union (ACLU) has begun hiring technologists and is increasingly engaged in debates and legislation related to new technology. Organizations such as Data & Society, Upturn, the Center for Humane Technology, and Tactical Tech study the social implications of technology and advocate for improved technology governance and design practices.

Many advocates call for engineers to follow an ethical oath modeled after the Hippocratic Oath, an ethical oath taken by physicians (Eubanks, 2018a; O'Neil, 2017; Patil, 2018; Simonite, 2018). In 2018, for instance, the organization Data for Democracy partnered with Bloomberg and the data platform provider BrightHive to develop a code of ethics for data scientists, developing 20 principles that include “I will respect human dignity” and “It’s my responsibility to increase social benefit while minimizing harm” (Data4Democracy, 2018). Former U.S. Chief Data Scientist DJ Patil described the event as the “Constitutional Convention” for data science (Eubanks, 2018b). A related effort, produced by the Institute for the Future and the Omidyar Network, is the Ethical OS Toolkit, a set of prompts and checklists to help technology developers “anticipat[e] the future impact of today’s technology” and “not [...] regret the things you will build” (The Institute for the Future and Omidyar Network, 2018).

#### *2.4 Government*

Many governments have also taken up the mantle of tech ethics, developing commissions and principles dedicated to the topic. In the United States, for example, the National Science Foundation formed a Council for Big Data, Ethics, and Society (Council for Big Data, 2014), the National Science and Technology Council published a report about AI that emphasized ethics (National Science and Technology Council, 2018), and the Department of Defense adopted ethical principles for AI (U.S. Department of Defense, 2020). Elsewhere, governing bodies in Dubai (Smart Dubai, 2018), Europe (European Commission High-Level Expert Group on Artificial Intelligence, 2019), Japan (Integrated Innovation Strategy Promotion Council, 2019), and Mexico (Martinho-Truswell et al., 2018), as well as international organizations such as the OECD (Organisation for Economic Co-operation and Development, 2019) have all put forward documents stating principles for ethical AI development. A 2019 analysis of global AI ethics

guidelines found 84 such documents (with more than a third from the U.S. and U.K. and none from Africa or South America) espousing a common set of principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin, Ienca, and Vayena, 2019).

### **3 The Limits of Tech Ethics**

Despite the rapid adoption of “ethics” as an analytic frame for digital technologies, critical analyses of these early efforts have indicated that tech ethics suffers from several core limitations. First, the actual principles espoused by tech ethics statements are too abstract and lacking in enforcement mechanisms to reliably spur ethical behavior in practice. Second, as ethics is incorporated into tech companies, ethical ideals are subsumed into corporate logics and incentives. Third, by emphasizing the design decisions of individual engineers, tech ethics overlooks the structural forces that shape technology’s harmful social impacts. Collectively, these issues suggest that the emphasis on ethics represents a strategy of technology companies “ethics-washing” their behavior with a façade of ethics while largely continuing with business-as-usual.

#### *3.1 Tech ethics principles are abstract and toothless*

Tech ethics codes deal in universal principles (Greene, Hoffmann, and Stark, 2019). In 2016, for example, Accenture published a report explicitly outlining “a universal code of data ethics” (Accenture, 2016). Professional computing societies also present ethical commitments in a highly abstract form, encouraging computing professionals, for instance, “to be ever aware of the social, economic, cultural, and political impacts of their actions” and to “contribute to society and human well-being” (Stark and Hoffmann, 2019). Ethics codes in computing and information science are notably lacking in explicit commitments to normative principles (Stark and Hoffmann, 2019).

The emphasis on universal principles papers over the fault lines of debate and disagreement that spurred the emergence of tech ethics as a widespread discourse in the first place. The recent upsurge in tech ethics was prompted by a spate of tech scandals and ensuing critiques of digital technologies. Yet the principles that have been developed embody an almost eerie level of agreement: two 2019 reports on global AI ethics guidelines similarly emphasized a “global convergence” (Jobin, Ienca, and Vayena, 2019) and a “consensus” (Fjeld et al., 2020) in the principles espoused. Although these documents tend to reflect a common set of global principles,

the actual interpretation and implementation of these principles raise substantive conflicts (Jobin, Ienca, and Vayena, 2019). The superficial consensus around abstract ideals may thus be hindering substantive deliberation regarding whether the chosen values are appropriate, how those values should be balanced in different contexts, and what those values actually entail in practice.

This level of abstraction and consensus is particularly troubling due to a lack of mechanisms to enact or enforce the principles embodied in tech ethics principles. When framed at such a high level of abstraction, values such as fairness and respect are unable to guide specific actions (Mittelstadt, 2019). Nor are these principles binding. The ethics oversight boards in companies such as Google and Axon and the ethics principles in companies and governments around the world lack the independent authority to veto projects or require certain behaviors (Harwell, 2018; Jobin, Ienca, and Vayena, 2019; Knight, 2019). Similarly, professional computing organizations such as the IEEE and ACM lack the power to meaningfully sanction individuals who violate their codes of ethics (Mittelstadt, 2019). Moreover, unlike fields such as medicine, which has a strong and established emphasis on professional ethics, computing lacks a common aim or fiduciary duty to unify disparate actors around shared ethical practices (Mittelstadt, 2019). All told, “Principles alone cannot guarantee ethical AI” (Mittelstadt, 2019).

### *3.2 Tech ethics is subsumed into corporate logics and incentives*

Digital technology companies have embraced ethics as a matter of corporate concern, stemming in large part from the desire to avoid reputational or financial harm. In recent SEC filings, both Alphabet and Microsoft noted the potential harms they could face for products that are deemed unethical (Simonite, 2019). Corporate tech ethics thus emphasizes the presentation of ethical behavior for scrutinizing audiences. An ethnography of ethics work in Silicon Valley found that “[p]erforming, or even showing off, the seriousness with which a company takes ethics becomes a more important sign of ethical practices than real changes to a product” (Metcalf, Moss, and boyd, 2019). For instance, after an effort at Twitter to reduce online harassment stalled, an external researcher involved in the effort noted, “The impression I came away with from this experience is that [Twitter was] more sensitive to deflecting criticism than in solving the problem of harassment” (Seetharaman, 2020).

Ethics is also framed by companies in terms of its direct alignment with business strategy. A software engineer at LinkedIn described algorithmic fairness as being profitable for companies, arguing, “If you’re very biased, you might only cater to one population, and eventually that limits the growth of your user base, so from a business perspective you actually want to have everyone come on board, so it’s actually a good business decision in the long run” (Johnson, 2019). Similarly, one of the people behind the Ethical OS toolkit described being motivated to produce “a tool that helps you think through [societal] consequences and makes sure what you’re designing is good for the world and good for your longer-term bottom line” (Pardes, 2018).

Finding this alignment between ethics and business is an important task for those charged with promoting ethics in tech companies. Recognizing that “[market] success trumps ethics,” individuals focused on ethics in Silicon Valley feel pressure to align ethical principles with corporate revenue sources (Metcalf, Moss, and boyd, 2019). As one senior researcher in a tech company notes, “the [ethics] system that you create has to be something that people feel adds value and is not a massive roadblock that adds no value, because if it is a roadblock that has no value, people literally won’t do it, because they don’t have to” (Metcalf, Moss, and boyd, 2019). When ethical ideals are at odds with a company’s bottom line, they are met with resistance (Marantz, 2019).

The emphasis on business strategy creates significant conflicts with ethics in technology companies, whose business models often rely on extractive and exploitative practices that were partially responsible for the scandals that prompted the “techlash” in the first place. Indeed, announcements by Facebook and Twitter that they would invest in privacy, security, and combating misinformation and abusive behavior led to rapid declines in the companies’ stock values (Neate, 2018; Phillips, 2018). Moreover, even as tech companies espouse a devotion to ethics, they continue to develop products and services that raise ethical red flags but promise significant profits. For example, even after releasing AI ethics principles that include safety, privacy, and inclusiveness (Microsoft, 2018) and committing not to “deploy facial recognition technology in scenarios that we believe will put these freedoms at risk” (Smith, 2018), Microsoft invested in AnyVision, an Israeli facial recognition company that supports military surveillance of Palestinians in the West Bank (Solon, 2019). Microsoft claimed that AnyVision complied with its

ethics principles (even while former AnyVision employees admitted that the company did not meet Microsoft's ethical standards), suggesting either that these principles are so flexible as to allow such applications or that they are followed only when doing so is convenient.

These examples indicate that tech ethics is being subsumed into existing tech company logics and business practices rather than meaningfully challenging or changing those logics and practices (even if some individuals within companies do want to create meaningful change). This absorption allows companies to take up the mantle of caring about ethics without making substantive changes to their processes or business strategies. The goal in companies is to find practices “which the organization is *not yet doing* but is *capable of doing*” (Metcalf, Moss, and boyd, 2019), indicating an effort to find relatively costless reforms that provide the veneer of ethical behavior. Ethics vision statements “co-opt the language of some critics,” taking critiques grounded in a devotion to equity and social justice and turning them into ethical discussions that are akin to “conventional business ethics” (Greene, Hoffmann, and Stark, 2019). As they integrate these principles into new practices and products, tech companies “are learning to speak and perform ethics rather than make the structural changes necessary to achieve the social values underpinning the ethical fault lines that exist” (Metcalf, Moss, and boyd, 2019).

These limits to corporate tech ethics are exemplified by Google’s firing of Timnit Gebru (and later Meg Mitchell) following the company’s concerns about a paper examining the limitations and harms of large language models, which are central to Google’s business (Hao, 2020). Despite Gebru’s and Mitchell’s supposed charge as co-leads of Google’s ethical AI team, Google objected to a paper they had written with several internal and external colleagues about ethical concerns related to these AI models, suggesting that the authors were insufficiently attentive to recent technical advances that mitigate these concerns (Hao, 2020). Soon after, journalists revealed that Google had expanded its review of papers that discuss “sensitive topics,” telling researchers, for instance, to “take great care to strike a positive tone” regarding Google’s technologies and products (Dave and Dastin, 2020). Thus, even as Google publicly advertised its care for the ethics of its technologies, internally the company was acting strongly to curtail criticisms that it deemed threatening to its core business interests.

### *3.3 Tech ethics has a myopic focus on individual engineers and technology design*

Tech ethics typically emphasizes the roles and responsibilities of engineers, paying relatively little attention to the broader environments in which these individuals work. Although professional codes in computing and related fields state general commitments to the public, profession, and one's employer, “the morality of a profession’s or an employer’s motives are not scrutinized” (Stark and Hoffmann, 2019). Similarly, ethics within computer science curricula tends to focus on ethical decision making for individual engineers (Silbey, 2018).

From this individualistic frame comes an emphasis on appealing to the good intentions of engineers, with the assumption that better design practices and procedures will lead to better technology. Ethics becomes a matter of individual engineers and managers “doing the right thing” “for the right reasons” (Metcalf, Moss, and boyd, 2019). Efforts to provide ethical guidance for tech CEOs rest on a similar logic: “if a handful of people have this much power—if they can, simply by making more ethical decisions, cause billions of users to be less addicted and isolated and confused and miserable—then, isn’t that worth a shot?” (Marantz, 2019). The broader public beyond technical experts is not seen as having a role in defining ethical concerns or shaping the responses to these concerns (Greene, Hoffmann, and Stark, 2019).

Tech ethics therefore centers debates about *how* to build better technology rather than *whether* or *in what form* to build technology (let alone who gets to make such decisions). Underlying ethics work across academia, civil society, and tech companies is an assumption that artificial intelligence and machine learning are “inevitable,” such that “‘better building’ is the only ethical path forward” (Greene, Hoffmann, and Stark, 2019). In turn, tech ethics diagnoses the harmful social consequences of technology as treatable through technical and procedural solutions (Metcalf, Moss, and boyd, 2019). Ethics teams within tech companies have developed and shared numerous ethics and fairness toolkits, including Datasheets at Microsoft (Gebru et al., 2018), Model Cards at Google (Mitchell et al., 2019), AI Fairness 360 at IBM (Varshney, 2018), the Fairness Tool at Accenture (Peters, 2018), and Fairness Flow at Facebook (Gershgorn, 2018).

Although efforts such as these stand to remedy certain harms that result from digital technology, approaches to tech ethics that focus on the design decisions of engineers omit much of what

scholarship in STS, law, and other fields has diagnosed regarding the source of social harms connected to digital technologies. For example, scholars and journalists have articulated the social harms associated with business models that rely on collecting massive amounts of data about the public (Schneier, 2016; Zuboff, 2018), companies that wield monopolistic power (Khan, 2017; Wu, 2018), technologies that are built through the extraction of natural resources and the abuse of workers (Crawford and Joler, 2018; Dobbe and Whittaker, 2019; Evans, 2019; Gray and Suri, 2019), and the exclusion of women, minorities, and non-technical experts from decisions with significant social impacts (Jasanoff, 2006; West, Whittaker, and Crawford, 2019). To the extent that efforts based in individual design decisions are taken as the heart of what it means to “do” ethics, they bear the “risk of a premature foreclosure of the fundamentally open-ended and irresolvable questions that underlie human value commitments” into a set of procedures that must be followed (Metcalf, Moss, and boyd, 2019).

The focus on improving design also relies on a narrow theory of change for how to reform technology. Regardless of their intentions and the design frameworks at their disposal, individual engineers typically have little power to shift corporate strategy. Executives can limit which teams know about a project and what they know about it, enforcing strict secrecy and segmenting project teams to prevent knowledge and internal dissent about controversial projects (Conger and Metz, 2018; Gallagher, 2018). Even when engineers do know about and protest projects, the result is often them resigning or being replaced rather than the company changing course (Conger and Metz, 2018; Simonite, 2018). Instead of efforts focused explicitly on design and ethics, many improvements in tech systems and new regulations have been the result of collective action among tech workers as well as external pressure and organizing from activists, journalists, workers, and scholars (Crawford et al., 2019; Haskins, 2020).

These structural conditions place significant barriers on the benefits that design-oriented tech ethics will be able to achieve. As MIT anthropologist Susan Silbey notes in regard to teaching engineering ethics, “while we might want to acknowledge human agency and decision-making at the heart of ethical action, [...] we blind ourselves to the structure of those choices—*incentives, content, and pattern*—if we focus too closely on the individual and ignore the larger pattern of opportunities and motives that channel the actions we call ethics” (Silbey, 2018). By ignoring the

contexts of technology production and how technology interacts with society, we risk tinkering on the margins of technology, developing a lingo and practices that come to define ethical behavior while leaving in place the structures that generated the controversies spurring the rise of tech ethics.

### *3.4 Tech ethics has become an avenue for ethics-washing*

As evidence of tech ethics' limitations has grown, many have critiqued tech ethics as a strategic effort among technology companies to quell public scrutiny rather than as a noble effort to take responsibility for technology's social impacts. This strategy has been labeled "ethics-washing" (i.e., "ethical white-washing"): using the language of ethics to paint a superficial portrait of ethical behavior in order to avoid heightened public backlash and the introduction of regulations that would require substantive concessions (Metzinger, 2019; Nemitz, 2018; Wagner, 2018).

In other words, ethics discourse has become a convenient way for tech companies to defuse criticism and regulation by creating structures for self-governance without any commitment to meaningfully altering their behavior. As an ethnography of ethics in Silicon Valley found, "It is a routine experience at 'ethics' events and workshops in Silicon Valley to hear ethics framed as a form of self-regulation necessary to stave off increased governmental regulation" (Metcalf, Moss, and boyd, 2019). Recognizing this strategy casts important "flaws" of tech ethics as features rather than bugs: by focusing public attention on the actions of individual engineers and particular technical limitations (such as algorithmic bias), companies perform a sleight-of-hand that shifts structural questions about power and profit out of view.

Thomas Metzinger, a philosopher who served on the European Commission's High-Level Expert Group (HLEG) on Artificial Intelligence to develop AI ethics guidelines (European Commission High-Level Expert Group on Artificial Intelligence, 2019), provides a particularly striking account of ethics-washing in action (Metzinger, 2019). The HLEG on AI contained only four ethicists out of 52 total people and was dominated by representatives from industry. Metzinger's own work to develop "Red Lines" that AI applications should not cross was "watered down" by industry representatives eager for a "positive vision" for AI. All told, Metzinger describes the HLEG's guidelines as "lukewarm, short-sighted and deliberately vague" and concludes that the tech

industry is “using ethics debates as elegant public decorations for a large-scale investment strategy” (Metzinger, 2019).

Tech companies have further advanced this “ethics-washing” agenda through funding academic research and events. Many of the scholars writing about tech policy and ethics are funded by Google, Microsoft, and others, yet often do not disclose this funding (Google Transparency Project, 2017; Williams, 2019). Tech companies also provide funding for prominent academic conferences, including the ACM Conference on Fairness, Accountability, and Transparency (FAccT); and the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES); and the Privacy Law Scholars Conference (PLSC). Even if these funding practices do not directly influence the research output of individual scholars, they allow tech companies to shape the broader academic and public discourse regarding tech ethics, raising up certain voices and conversations at the expense of others.<sup>2</sup>

Further debate about the value and impacts of tech ethics erupted in December 2019, spurred by an article written by MIT graduate student Rodrigo Ochigame (Ochigame, 2019) in the wake of the revelations that the MIT Media Lab—a center that conducts research and provides grants related to tech ethics—had received secret funding from Jeffrey Epstein, the financier charged with sex trafficking of minors (Farrow, 2019; Tracy and Hsu, 2019). Describing his experiences working in Joi Ito’s AI ethics group at the Media Lab and collaborating with the Partnership on AI, Ochigame articulated how “the discourse of ‘ethical AI’ [...] was aligned strategically with a Silicon Valley effort seeking to avoid legally enforceable restrictions of controversial technologies” (Ochigame, 2019). Ochigame described witnessing firsthand how the Partnership on AI made recommendations that “aligned consistently with the corporate agenda” by reducing political questions about the criminal justice system to matters of technical consideration. A central part of this effort was tech companies strategically funding researchers and conferences in order to generate a widespread discourse about “ethical” technology. Finding that “the corporate lobby’s effort to shape academic research was extremely successful,” Ochigame concluded that “[b]ig tech money and direction proved incompatible with an honest exploration of ethics.”

---

<sup>2</sup> The integrity of academic tech ethics has been further called into question due to funding from other sources beyond tech companies (Domínguez et al., 2019; Farrow, 2019; Mboya, 2019).

Some believed that Ochigame oversimplified the story, failing to fully credit the many people behind tech ethics and the reforms that this movement have prompted (Darling, 2019; Epstein, 2019; Sinders, 2019). On this view, all of the work by scholars and activists increasing attention to technological harms and pushing for legislation was worthy of praise as more than just corporate capture. Yet many of the people (often activists and scholars of color) centrally involved in efforts to expose and resist the harms of digital technology see their work as distinct from tech ethics, which represents the narrow domain of efforts typically promulgated by tech companies. Safiya Noble described Ochigame's article as "All the way correct and worth the time to read" (Noble, 2019). Lilly Irani and Ruha Benjamin expressed similar sentiments, noting that "AI ethics is not a m[o]v[e]m[e]nt" (Irani, 2019) and that "many of us don't frame our work as 'ethical AI'" (Benjamin, 2019).

This debate in response to Ochigame's article exposed the fault lines at the heart of tech ethics. While tech ethics is often framed as encompassing the broad societal debates about the social impacts of technology, many of the people at the forefront of those efforts see tech ethics as the narrower (typically industry-led) efforts to promote "ethics" in technology. Where, then, do the bounds of tech ethics lie? Who is behind tech ethics: tech companies, or the activists and scholars who are highlighting harmful technologies and advocating for reforms? What are the appropriate strategies for pursuing ethical technology: through reforming design practices, or through reforming the political and economic forces that generate and govern technology? And, perhaps most fundamentally, what is the value of tech ethics: does it provide valuable reforms (even if incremental ones), or is it an active hindrance to achieving more just technology? The answers to these contested questions will shape the future of efforts to pursue more just digital infrastructures and technologies.

#### **4 The Contestation of Tech Ethics**

The debates described in the previous section reveal that the central question regarding tech ethics is not over *whether* it is desirable to be ethical, but over *what* "ethics" entails and *who* gets to define it. In this sense, ethics is an "essentially contested concept" (Gallie, 1955), where the central debates regard the meaning—rather than desirability—of ethics. Akin to the debates described

above, the “ethics” in tech ethics tends to take on four overlapping yet often conflicting meanings: moral justice, corporate values, legal risk, and compliance (Moss and Metcalf, 2020).

Such definitional debates have significant stakes. In the context of antidiscrimination law, Kimberlé Crenshaw describes the “definitional tension” between an “expansive view” of antidiscrimination that strives to eradicate the oppression of Blacks and a “restrictive view” of antidiscrimination that aims only to prevent discrimination based explicitly on race (Crenshaw, 1988). In any given instance, “specific interpretations proceed largely from the world view of the interpreter” (Crenshaw, 1988). Similarly, different interpretations of “tech ethics” suggest drastically different paths forward: tech ethics embodies a conflict between an expansive view that aims to remedy a broad range of injustices associated with digital technologies and a restrictive view that aims to limit a specific set of tangible harms caused by these technologies. Because of these conflicting accounts co-existing within the same term, to call for tech ethics is, in Crenshaw’s terms, “to demand nothing specific” (Crenshaw, 1988).

In this sense, tech ethics is itself a terrain of contestation whose impacts hinge on who is able to demarcate the bounds of legitimate authority regarding what ethics entails in relation to technology and who is responsible for technologies’ social impacts. Whether it be technology companies projecting procedural toolkits as solutions to ethical dilemmas, computer scientists reducing normative questions into mathematical metrics, academic tech ethics institutes being funded by billionaires and led primarily by white men (Gershgorn, 2019), or tech ethics principles being disseminated predominantly by the U.S. and Western Europe, the contestation of tech ethics centers on certain actors attempting to acquire or maintain authority over what it means for technology to be “ethical,” at the expense of others. These efforts embody practices of “boundary-work,” which involves drawing boundaries between legitimate and illegitimate sources of authority (Gieryn, 1983). Drawing these boundaries can help scientists acquire intellectual authority and maintain professional autonomy, often in ways that exclude women, minorities, the Global South, and other publics (Collins, 2000; Haraway, 1988; Visvanathan, 2005).

Examples of how ethics has been implemented in two other domains—science and business—shed light on the nature and stakes of present debates about tech ethics.

#### *4.1 Ethics in Science*

Many areas of science have embraced ethics in recent decades in response to debates about the social implications of emerging research and applications. Despite the seeming promise of ethics in science, however, existing approaches to promoting ethics in science are limited in their ability both to raise debates about the structure and values of science and to promote democratic governance of science. Science ethics bodies suffer from limited “ethical imaginations” and are often primarily concerned with “keep[ing] the wheels of research turning while satisfying publics that ethical standards are being met” (Jasanoff, 2016). “Ethical analysis that does not advance such instrumental purposes tends to be downgraded as not worthy of public support” (Jasanoff, 2016).

Rather than interrogating fundamental questions about the purposes of research or who gets to shape that research, ethics has become increasingly institutionalized, instrumentalized, and professionalized, with an emphasis on filling out forms and checking off boxes (Jasanoff, 2016; Reardon, 2013). In turn, “systems of ethics [...] play key roles in eliding fundamental social and political issues” that inhere in the conception and development of research (Reardon, 2011). For instance, efforts to introduce ethics into genetic research throughout the 1990s and 2000s treated ethics “as something that could be added onto science—and not something that was unavoidably implicit in it,” thus obscuring the “bioconstitutional” questions about racial groups at the heart of genetic testing projects (Reardon, 2011). Because “ethical choices inhered in efforts to study human genetic variation, regardless of any explicit effort to *practice ethics*,” these research projects “bypass[ed] responsibility for their roles in co-constituting natural and moral orderings of human difference, despite efforts to address ethics at the earliest stages of research design” (Reardon, 2011).

The turn to ethics can also entail an explicit effort among scientists to defuse external scrutiny and to develop a regime of self-governance. For example, biologists in the 1970s, frightened by calls for greater public participation in genetic engineering, organized a conference at the Asilomar Conference Center in California. The scientific community at Asilomar pursued two, intertwined goals. In order to present a unified and responsible public image, the Asilomar organizers restricted the agenda to eschew discussions of the most controversial applications of genetic engineering

(biological warfare and human genetic engineering). And in order to convince the American public and politicians to allow biologists to self-govern their pursuit of genetic engineering, the Asilomar attendees “redefined the genetic engineering problem as a technical one” that only biologists could credibly discuss (Wright, 2001). Although often hailed as a remarkable occasion of scientific self-sacrifice for the greater good, accounts from the conference itself demonstrate that “[s]elf-interest, not altruism, was most evident at Asilomar,” as not making any sacrifices and appearing self-serving would have only invited stringent, external regulation (Wright, 2001).

Tech ethics mirrors many of these attributes in scientific ethics. As with ethics in other fields of science, tech ethics involves a significant emphasis on institutionalized design practices, often entailing checklists and worksheets. Mirroring ethics in genetic research, the emphasis on ethical design in computer science departments and tech companies treats ethics as something that can be *added on* to digital technologies by individual engineers, overlooking the epistemologies and economic structures that underlie these technologies and their harms. Furthermore, tech companies and computer scientists focusing on technical challenges such as algorithmic fairness mirror the strategic efforts of molecular biologists at Asilomar to reframe moral questions as technical questions in order to retain self-regulation.<sup>3</sup> The removal of red lines in the European Commission’s High-Level Expert Group on AI bears a striking resemblance to the exclusion of controversial topics from the agenda at Asilomar.

#### *4.2 Corporate Ethics and Co-optation*

Codes of ethics have long been employed by groups of experts (e.g., doctors and lawyers) to codify a profession’s expectations regarding culture and behavior and to shore up the profession’s public reputation (Abbott, 1983; Metcalf, 2014). Similarly, companies across a wide range of sectors have embraced ethics codes, particularly over the past half century and typically in response to public perceptions of unethical behavior (Wood and Rimmer, 2003).

Yet it has long been clear that the public benefits of corporate ethics codes are minimal. While ethics codes can help make a group appear ethical, on their own they do little to promote a culture

---

<sup>3</sup> In an ironic parallel, the Future of Life Institute organized an Asilomar Conference on Beneficial AI in 2017, leading to the development of 23 “Asilomar AI Principles” (Future of Life Institute, 2017).

of ethical behavior (Wood and Rimmer, 2003). The primary goal of business ethics has instead been the “inherently unethical” motivation of corporate self-preservation: to reduce public and regulatory scrutiny by promoting a visible appearance of ethical behavior (Cressey and Moore, 1983; Wood and Rimmer, 2003). This emphasis on corporate reputation and profit is facilitated by ethics codes making universal moral claims that “are extremely important as claims but extremely vague as rules” and emphasizing individual actors and behaviors, leading to a narrow, “one-case-at-a-time approach to control and discipline” (Abbott, 1983). Ethics codes in the field of information systems have long exhibited a notable lack of explicit moral obligations for computing professionals (Oz, 1992; Stark and Hoffmann, 2019).

Business ethics is indicative of the broader phenomenon of co-optation: an institution incorporating elements of external critiques from groups such as social movements—often gaining the group’s support and improving the institution’s image—with meaningful action on that group’s demands or providing that group with decision-making authority (Gamson, 1975; Selznick, 1948; Trumpy, 2014). The increasing centrality of companies as the target of social movements has led to a particular form of co-optation called “corporatization,” in which “corporate interests come to engage with ideas and practices initiated by a social movement and, ultimately, to significantly shape discourses and practices initiated by the movement” (King and Busa, 2017). Through this process, large corporate actors in the United States have embraced “diluted and deradicalized” elements of social movements “that could be scaled up and adapted for mass markets” (King and Busa, 2017). Two factors make movements particularly susceptible to corporatization: heterogeneity—movement factions that are willing to work with companies gain influence through access to funding—and materiality—structural changes get overlooked in favor of easily commodifiable technological “fixes.” By participating in movement-initiated discourses, companies are able to present themselves as part of the solution rather than part of the problem, and in doing so can avoid more restrictive government regulations.

Tech ethics embodies many of the attributes of corporate ethics, particularly the significant efforts to shore up legitimacy and avoid external regulation. Abstract and individualized tech ethics codes reproduce the virtue signaling and self-preservation behind traditional business ethics. And in a notable example of co-optation and corporatization, technology companies have promoted tech

ethics as a diluted and commoditized version of tech-critical discourses that originated largely from outside of technology circles. Because societal efforts to improve technology are often aimed at companies and include both heterogeneity and materiality, it is particularly vulnerable to corporatization. Through this process of corporatization, tech companies are using tech ethics to present themselves as part of the solution rather than part of the problem and are using funding to empower the voices of certain scholars and academic communities. The power of this influence can be seen in the expanding scope of work that is published and discussed at conferences, workshops, and other events under the banner of “tech ethics.” Even scholars who do not embrace the tech ethics label are increasingly subsumed into this category, either lumped into it by others or compelled into it as opportunities to publish research, impact policymakers, and receive grants are increasingly shifting to the terrain of “ethics.”

#### *4.3 The Future of Tech Ethics*

Drawing on these two examples of ethics in action leads to two conclusions about tech ethics. First, the parallels to ethics in science and business indicate that current tech ethics discourse is likely to enable technologists and technology companies to label themselves as “ethical” without substantively altering their practices. Although there are multiple notions of tech ethics currently coexisting—many individuals and groups, including some within tech companies, are pursuing expansive forms of tech ethics—the influence of tech companies makes it likely that their narrow vision of “tech ethics” will subsume any others. Furthermore, many of the most prominent voices regarding tech ethics are white men who want to claim expertise and thought leadership while ignoring the work of established fields and scholars, many of whom are women and people of color (Irani and Chowdhury, 2019; Mozilla, 2020). To the extent that tech ethics continues to follow the path of science ethics, business ethics, and corporatization, tech companies and others will be able to define ethics in such a way that technologies and technology companies are deemed “ethical” even while continuing to produce significant social harm. For the tech companies that dominate tech ethics discourse, such narrowing is precisely the point.

Second, rather than treating “ethics” as well-defined and inherently desirable, those striving for substantive and structural improvements in digital technologies must look to the formulation and real-world effects of “tech ethics.” The examples of science and business ethics indicate that ethics

in practical contexts can be quite distinct from the normative demands and moral inquiry that gave rise to the embrace of ethics in the first place. Just as digital technology is often applied as a solution to social problems, tech ethics is today being applied as a solution to sociotechnical problems. As with technology itself, tech ethics is a tool with particular affordances that is being developed to serve social purposes. If “technologies can be assessed only in their relations to the sites of their production and use” (Suchman et al., 1999), then so too, we might say, tech ethics can be assessed only with regard to how it is conceived and how it affects the world in practice.

Thus, rather than presenting a unifying and beneficent set of principles and practices, tech ethics has emerged as a central site of struggle regarding the future of digital architectures, governance, and economies. As was hinted at in the varied responses to Ochigame’s critique of ethics-washing, many of the more radical critics of technology and inequality see themselves as outside of (if not in opposition to) the dominant strains of tech ethics. Discourses and practices of resistance (Anti-Eviction Mapping Project, 2020), defense (Lewis et al., 2018), abolition (Hamid, 2020; Stop LAPD Spying Coalition and Free Radicals, 2020), and decentering technology (Gangadharan and Niklas, 2019) have also emerged (often from activists and communities rather than academics) in response to the social injustices meted out through digital technologies. Although some may see all of these efforts as falling under the broad umbrella of an expansive tech ethics, these nascent movements embody distinct aspirations from the narrow mainstream of tech ethics. Conflating them under a common label risks giving tech ethics the imprimatur of radical, justice-oriented work even as its core tenets and practices eschew such commitments. Efforts to resist oppressive technological architectures and to study or support such efforts must be attentive to these porous and slippery boundaries.

## 5 A Sociotechnical Approach to Tech Ethics

Given these dynamics of contestation surrounding tech ethics, ethics will not, on its own, provide a salve for technology’s social harms. Nonetheless, taking a reflexive (Bloor, 1991) approach that applies core elements of tech ethics’ analyses of digital technology to tech ethics itself contains the seeds of a more robust approach to pursuing a more just society, digital and otherwise. One dimension of tech ethics (and critiques of tech ethics) is a focus on technology’s social impacts and on how technology interacts with society—in other words, approaching digital technologies

through a sociotechnical analysis. This sociotechnical approach, drawing on STS, has much to offer engineering ethics, shedding light on the responsibilities of engineers and how artifacts shape and are shaped by the social world (Johnson and Wetmore, 2007). This suggests the value of an approach to tech ethics that mirrors the sociotechnical approach to technology—one that can inform our understanding of the responsibilities of those behind tech ethics and how tech ethics shapes and is shaped by the social world. With this aim in mind, it is fruitful to consider tech ethics through the lens of four sociotechnical frames: determinism, solutionism, objectivity and neutrality, and sociotechnical systems.

### *5.1 Determinism*

A central component of a sociotechnical approach to technology is rejecting technological determinism: the belief that technology evolves autonomously and determines social outcomes (Dafoe, 2015; Marx and Smith, 1994). Such an approach has been rejected through scholarship demonstrating that even as technology plays some role in shaping society, so too, simultaneously, does society shape technology (Bijker and Law, 1994; Jasanoff, 2004; Pinch and Bijker, 1987; Winner, 1986). Ethics in digital technology is today being treated through a similar sort of “ethical determinism,” with an underlying assumption that adopting “ethics” will lead to ethical technologies. Yet in both science and business there is a long history demonstrating that such an embrace of ethics is not sufficient to prompt substantive changes in behavior, and there is emerging evidence of the same limitation with regard to tech ethics. As with technology, ethics does not emerge in a vacuum and does not on its own determine sociotechnical outcomes. What is needed, then, is a new approach that looks to the sociotechnical complexities of tech ethics: how it shapes the development and governance of technologies (and collective understandings of technologies) as well as how ethical discourses and practices are themselves shaped by a variety of social forces.

### *5.2 Solutionism*

Closely intertwined with a belief in technological determinism is the practice of technological solutionism: the expectation that technology can and will solve all social problems (Morozov, 2014). A great deal of scholarship has demonstrated how technological solutions not only typically fail to provide the intended solutions, but also can exacerbate the problems they are intended to solve (Ames, 2019; Green, 2019; Morozov, 2014; Toyama, 2015). Yet even as tech ethics debates

have highlighted how technology is not always the answer to social problems, a common response has been to embrace an “ethical solutionism”: promoting ethics principles and practices as the solution to these sociotechnical problems. A notable example (at the heart of many tech ethics agendas) is the response to algorithmic discrimination through algorithmic fairness, which often centers narrow mathematical definitions of fairness but leaves in place the structural and systemic conditions that generate a great deal of algorithmic harms (Green, 2020; Hoffmann, 2019). Efforts to introduce ethics across science, business, and technology function similarly, providing an addendum of ethical language and practices on top of existing structures and epistemologies which themselves go largely uninterrogated. Thus, just as technical specifications of algorithmic fairness are insufficient to guarantee fair algorithms, tech ethics principles are insufficient to guarantee ethical technologies. Ethics principles, toolkits, and other mechanisms are just one component of what must be a broad array of approaches to improve technology.

### *5.3 Objectivity and Neutrality*

Debates about biased and harmful technologies have led to an increasingly widespread recognition in computer science and the tech industry of notable insights from STS: engineers are not objective and technology is not neutral. It is clear that improving digital technologies requires grappling with the normative commitments of engineers and incorporating more voices into the design of technology (Costanza-Chock, 2020; Green and Viljoen, 2020). Yet even as tech ethics emphasizes principles such as fairness and inclusiveness, the range of perspectives remains quite narrow and ethics is treated as an objective, universal body of principles (Fjeld et al., 2020; Greene, Hoffmann, and Stark, 2019; Jobin, Ienca, and Vayena, 2019). The consensus around particular ethical principles may therefore say less about the objective universality of these ideals than about the narrow range of perspectives that have been given voice regarding tech ethics. In many cases, white and male former technology company employees are cast to the front lines of public influence regarding tech ethics (Irani and Chowdhury, 2019; Mozilla, 2020). Rather than treating tech ethics as the search for objective and universal moral principles, it is necessary to grapple with the standpoints and power of different actors, the normative principles embodied in different ethical frameworks, and potential mechanisms for adjudicating between conflicting ethical commitments.

#### *5.4 Sociotechnical Systems*

A key result of treating technologies as embedded within sociotechnical systems is expanding the frame of analysis beyond the technical artifact itself: rather than technical features fully determining an artifact's social impacts, the artifact and the social world "co-produce" social outcomes (Jasanoff, 2004). Technologies are not discrete objects that can be properly evaluated in the abstract from the context of their use (Suchman et al., 1999). Indeed, many of the animating concerns and critiques of tech ethics connect to harms that arise when digital technologies are integrated into society without proper attention to the social context and human interactions that will shape its impacts (Green and Chen, 2019; Green and Viljoen, 2020; Rose, 2019; Vincent, 2016; Vosoughi, Roy, and Aral, 2018). Nonetheless, efforts to promote ethical technology typically focus on the internal characteristics of tech ethics—which principles to promote, for instance—with little attention to the impacts that these efforts will or will not have when integrated into larger settings such as a tech company or computer science curriculum. In turn, tech ethics has had limited effects on technology production and has played a sometimes-legitimizing role for technology companies. But just as the "unintended consequences" of technology often represents a failure to consider how artifacts might be used or abused in practice (Jasanoff, 2016), these limited impacts of tech ethics must not be seen as having been impossible to predict, particularly given the precedents of science ethics, business ethics, and corporate co-optation. Rather than treating tech ethics as being defined by its formal characteristics such as the principles espoused, any attempts to promote more ethical technology must operate with an eye toward the many factors that will shape the real-world impacts of tech ethics efforts. This includes considering how central framings (such as ethics) could be redefined and wielded by different actors as well as how to robustly embody moral principles in the procedures involved in developing, evaluating, and deploying technology.

## Acknowledgments

I thank Elettra Bietti, Anna Lauren Hoffmann, Jenny Korn, Kathy Pham, and Luke Stark for their comments on this essay. I also thank the Harvard STS community, particularly Sam Weiss Evans, for feedback and direction on an earlier iteration of this chapter.

## References

- Abbott, Andrew. 1983. "Professional Ethics." *American Journal of Sociology* 88 (5):855-885.
- Accenture. 2016. Universal principles of data ethics. [https://www.accenture.com/\\_acnmedia/pdf-24/accenture-universal-principles-data-ethics.pdf](https://www.accenture.com/_acnmedia/pdf-24/accenture-universal-principles-data-ethics.pdf).
- Accenture. N.d. AI Ethics & Governance. <https://www.accenture.com/us-en/services/applied-intelligence/ai-ethics-governance>.
- ACM FAccT Conference. 2020. ACM FAccT Network. <https://facctconference.org/network/>.
- Adams, Amy. 2019. Stanford University launches the Institute for Human-Centered Artificial Intelligence. *Stanford News*. [https://news.stanford.edu/2019/03/18/stanford\\_university\\_launches\\_human-centered\\_ai/](https://news.stanford.edu/2019/03/18/stanford_university_launches_human-centered_ai/).
- AI Now Institute. 2017. The AI Now Institute Launches at NYU to Examine the Social Effects of Artificial Intelligence. <https://ainowinstitute.org/press-release-ai-now-launch>.
- Ames, Morgan G. 2019. *The Charisma Machine: The Life, Death, and Legacy of One Laptop per Child*: MIT Press.
- Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Anti-Eviction Mapping Project. 2020. *Counterpoints: A San Francisco Bay Area Atlas of Displacement & Resistance*. PM Press.
- Benjamin, Ruha. 2019. <https://twitter.com/ruha9/status/1208831999940714496>.
- Bijker, Wiebe E., and John Law, eds. *Shaping Technology/Building Society: Studies in Sociotechnical Change*. MIT press, 1994.
- Blake, Aaron. 2018. A new study suggests fake news might have won Donald Trump the 2016 election. *The Washington Post*. <https://www.washingtonpost.com/news/the-fix/wp/2018/04/03/a-new-study-suggests-fake-news-might-have-won-donald-trump-the-2016-election/>.

- Bloor, David. 1991. *Knowledge and Social Imagery*: University of Chicago Press.
- Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research.
- Cadwalladr, Carole, and Emma Graham-Harrison. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>.
- Cameron, Dell, and Kate Conger. 2018. Google Is Helping the Pentagon Build AI for Drones. *Gizmodo*. <https://gizmodo.com/google-is-helping-the-pentagon-build-ai-for-drones-1823464533>.
- Clifford Chance. N.d. Tech Group. <https://www.cliffordchance.com/hubs/tech-group-hub/tech-group.html>.
- Collins, Patricia Hill. 2000. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*: Routledge.
- Conger, Kate, and Cade Metz. 2018. Tech Workers Now Want to Know: What Are We Building This For? *The New York Times*. <https://www.nytimes.com/2018/10/07/technology/tech-workers-ask-censorship-surveillance.html>.
- Costanza-Chock, Sasha. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*: MIT Press.
- Council for Big Data, Ethics, and Society. 2014. Council for Big Data, Ethics, and Society. <https://bdes.datasociety.net>.
- Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. 2019. AI Now 2019 Report. [https://ainowinstitute.org/AI\\_Now\\_2019\\_Report.pdf](https://ainowinstitute.org/AI_Now_2019_Report.pdf).
- Crawford, Kate, and Vladan Joler. 2018. Anatomy of an AI System: The Amazon Echo as an anatomical map of human labor, data and planetary resources. <https://anatomyof.ai>.
- Crenshaw, Kimberlé Williams. 1988. "Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law." *Harvard Law Review* 101 (7):1331-1387.

- Cressey, Donald R., and Charles A. Moore. 1983. "Managerial Values and Corporate Codes of Ethics." *California Management Review* 25 (4):53-77. doi: 10.2307/41165032.
- Dafoe, Allan. 2015. "On Technological Determinism: A Typology, Scope Conditions, and a Mechanism." *Science, Technology, & Human Values* 40 (6):1047-1076. doi: 10.1177/0162243915579283.
- Darling, Kate. 2019. [https://twitter.com/grok\\_/status/1208434972564037633](https://twitter.com/grok_/status/1208434972564037633).
- Dastin, Jeffrey. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.
- Data4Democracy. 2018. Ethics Resources. <https://github.com/Data4Democracy/ethics-resources>.
- Dave, Paresh, and Jeffrey Dastin. 2020. Google told its scientists to 'strike a positive tone' in AI research - documents. *Reuters*. <https://www.reuters.com/article/us-alphabet-google-research-focus/google-told-its-scientists-to-strike-a-positive-tone-in-ai-research-documents-idUSKBN28X1CB>.
- Dieterich, William, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpoint Inc. Research Department*. [http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica\\_Commentary\\_Final\\_070616.pdf](http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf).
- Dobbe, Roel, and Meredith Whittaker. 2019. AI and Climate Change: How they're connected, and what we can do about it. *AI Now Institute*. <https://medium.com/@AINowInstitute/ai-and-climate-change-how-theyre-connected-and-what-we-can-do-about-it-6aa8d0f5b32c>.
- Domínguez, Alonso Espinosa, Remy Bassett-Audain, Husayn Karimi, Berenice Estrada, Claire Isabel Webb, Ruth Perry, Sally Haslanger, Jonathan King, Kevin Leonardo, Sarah Aladetan, Agnes Fury Cameron, Yarden Katz, Andrew Bolton, Lauren Surface, Kade Crockford, Katherine McConachie, Subrata Ghoshroy, and Alice Pote. 2019. Celebrating war criminals at MIT's 'ethical' College of Computing. *The Tech*. <https://thetech.com/2019/02/14/celebrating-war-criminals>.
- Dowd, Maureen. 2017. Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse. *Vanity Fair*. <https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x>.

- Emery Jr., C. Eugene. 2016. Evidence ridiculously thin for sensational claim of huge underground Clinton sex network. *PolitiFact*.  
<https://www.politifact.com/factchecks/2016/nov/04/conservative-daily-post/evidence-ridiculously-thin-sensational-claim-huge-/>.
- Epstein, Greg. 2019. <https://twitter.com/gregmepstein/status/1208798637221974016>.
- Eubanks, Virginia. 2018a. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*: St. Martin's Press.
- Eubanks, Virginia. 2018b. A Hippocratic Oath for Data Science. <https://virginia-eubanks.com/2018/02/21/a-hippocratic-oath-for-data-science/>.
- European Commission High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. <https://ec.europa.eu/futurium/en/ai-alliance-consultation>.
- Evans, Will. 2019. Ruthless Quotas at Amazon Are Maiming Employees. *The Atlantic*.  
<https://www.theatlantic.com/technology/archive/2019/11/amazon-warehouse-reports-show-worker-injuries/602530/>.
- Farrow, Ronan. 2019. How an Élite University Research Center Concealed Its Relationship with Jeffrey Epstein. *The New Yorker*. <https://www.newyorker.com/news/news-desk/how-an-elite-university-research-center-concealed-its-relationship-with-jeffrey-epstein>.
- Fiesler, Casey. 2018a. Tech Ethics Curricula: A Collection of Syllabi.  
<https://medium.com/@cfiesler/tech-ethics-curricula-a-collection-of-syllabi-3eedfb76be18>.
- Fiesler, Casey. 2018b. What Our Tech Ethics Crisis Says About the State of Computer Science Education. *How We Get To Next*. <https://howwegettonext.com/what-our-tech-ethics-crisis-says-about-the-state-of-computer-science-education-a6a5544e1da6>.
- Fiesler, Casey, Natalie Garrett, and Nathan Beard. 2020. "What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis." The 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20).
- Finley, Clint. 2016. Tech Giants Team Up to Keep AI From Getting Out of Hand. *Wired*.  
<https://www.wired.com/2016/09/google-facebook-microsoft-tackle-ethics-ai/>.
- Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Christopher Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based

Approaches to Principles for AI. *Berkman Klein Center Research Publication No. 2020-1.*  
<https://cyber.harvard.edu/publication/2020/principled-ai>.

Foroohar, Rana. 2018. Year in a Word: Techlash. *Financial Times*.  
<https://www.ft.com/content/76578fba-fca1-11e8-ac00-57a2a826423e>.

Future of Life Institute. 2017. Beneficial AI 2017. <https://futureoflife.org/bai-2017/>.

Gallagher, Ryan. 2018. Google Shut Out Privacy and Security Teams From Secret China Project. *The Intercept*. <https://theintercept.com/2018/11/29/google-china-censored-search/>.

Gallie, Walter B. 1955. "Essentially Contested Concepts." *Proceedings of the Aristotelian Society* 56:167-198.

Gamson, Howard. 1975. *The Strategy of Social Protest*: The Dorsey Press.

Seeta Peña Gangadharan & Jędrzej Niklas. 2019. "Decentering technology in discourse on discrimination." *Information, Communication & Society*, 22:7, 882-899. doi: 10.1080/1369118X.2019.1593484.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. "Datasheets for Datasets." *arXiv preprint arXiv:1803.09010*.

Gershgorn, Dave. 2018. Facebook says it has a tool to detect bias in its artificial intelligence. *Quartz*. <https://qz.com/1268520/facebook-says-it-has-a-tool-to-detect-bias-in-its-artificial-intelligence/>.

Gershgorn, Dave. 2019. Stanford's new AI institute is inadvertently showcasing one of tech's biggest problems. *Quartz*. <https://qz.com/1578617/stanfords-new-diverse-ai-institute-is-overwhelmingly-white-and-male/>.

Gieryn, Thomas F. 1983. "Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists." *American Sociological Review* 48 (6):781-795. doi: 10.2307/2095325.

Glaser, April, and Will Oremus. 2018. "A Collective Aghastness": Why Silicon Valley workers are demanding their employers stop doing business with the Trump administration. *Slate*. <https://slate.com/technology/2018/06/the-tech-workers-coalition-explains-how-silicon-valley-employees-are-forcing-companies-to-stop-doing-business-with-trump.html>.

Goldberg, Emma. 2020. 'Techlash' Hits College Campuses. *The New York Times*.  
<https://www.nytimes.com/2020/01/11/style/college-tech-recruiting.html>.

Google Transparency Project. 2017. Google Academics Inc.  
<https://www.techtransparencyproject.org/sites/default/files/Google-Academics-Inc.pdf>.

Graham, Jefferson. 2017. Hillary Clinton — tech has to fix fake news. *USA Today*.  
<https://www.usatoday.com/story/tech/talkingtech/2017/05/31/hrc-tech-has-fix-fake-news/102357904/>.

Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*: Houghton Mifflin Harcourt.

Green, Ben. 2019. *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future*: MIT Press.

Green, Ben. 2020. "The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain.

Green, Ben, and Yiling Chen. 2019. "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments." Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA.

Green, Ben, and Salomé Viljoen. 2020. "Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain.

Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark. 2019. "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." Proceedings of the 52nd Hawaii International Conference on System Sciences.

Grosz, Barbara J., David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. 2019. "Embedded EthiCS: Integrating Ethics Broadly Across Computer Science Education." *Communications of the ACM* 62 (8):54-61.

Hamid, Sarah. 2020. "Community Defense: Sarah T. Hamid on Abolishing Carceral Technologies." *Logic Magazine*. <https://logicmag.io/care/community-defense-sarah-t-hamid-on-abolishing-carceral-technologies/>.

Hao, Karen. 2020. We read the paper that forced Timnit Gebru out of Google. Here's what it says.

*MIT Technology Review*  
<https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/>.

- Haraway, Donna. 1988. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." *Feminist Studies* 14 (3):575-599.
- Harwell, Drew. 2018. Facial recognition may be coming to a police body camera near you. *The Washington Post*. <https://www.washingtonpost.com/news/the-switch/wp/2018/04/26/facial-recognition-may-be-coming-to-a-police-body-camera-near-you/>.
- Haskins, Caroline. 2020. The Los Angeles Police Department Says It Is Dumping A Controversial Predictive Policing Tool. *BuzzFeed News*. <https://www.buzzfeednews.com/article/carolinehaskins1/los-angeles-police-department-dumping-predpol-predictive>.
- Hern, Alex. 2016. 'Partnership on AI' formed by Google, Facebook, Amazon, IBM and Microsoft. *The Guardian*. <https://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms>.
- Hoffmann, Anna Lauren. 2019. "Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse." *Information, Communication & Society* 22 (7):900-915. doi: 10.1080/1369118X.2019.1573912.
- Hsu, Tiffany. 2018. For Many Facebook Users, a 'Last Straw' That Led Them to Quit. *The New York Times*. <https://www.nytimes.com/2018/03/21/technology/users-abandon-facebook.html>.
- IBM. 2018. Everyday Ethics for Artificial Intelligence. <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>.
- Integrated Innovation Strategy Promotion Council. 2019. AI for Everyone: People, Industries, Regions and Governments. <https://www8.cao.go.jp/cstp/english/humancentricai.pdf>.
- Irani, Lilly. 2019. <https://twitter.com/gleemie/status/1208793442509152258>.
- Irani, Lilly, and Rumman Chowdhury. 2019. "To Really 'Disrupt,' Tech Needs to Listen to Actual Researchers." *Wired*. <https://www.wired.com/story/tech-needs-to-listen-to-actual-researchers/>.
- Jasanoff, Sheila. 2004. "The idiom of co-production." In *States of Knowledge: The Co-Production of Science and the Social Order*, edited by Sheila Jasanoff, 1-12. Routledge.
- Jasanoff, Sheila. 2006. "Technology as a Site and Object of Politics." In *The Oxford Handbook of Contextual Political Analysis*, edited by Robert E. Goodin and Charles Tilly.

- Jasanoff, Sheila. 2016. *The Ethics of Invention: Technology and the Human Future*: W. W. Norton & Company.
- Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The global landscape of AI ethics guidelines." *Nature Machine Intelligence* 1 (9):389-399. doi: 10.1038/s42256-019-0088-2.
- Johnson, Deborah G., and Jameson M. Wetmore. 2007. "STS and Ethics: Implications for Engineering Ethics." In *The Handbook of Science and Technology Studies, Third Edition*, edited by Edward J. Hackett, Olga Amsterdamska, Michael E. Lynch and Judy Wajcman. MIT Press.
- Johnson, Khari. 2019. How to operationalize AI ethics. *VentureBeat*.  
<https://venturebeat.com/2019/10/07/how-to-operationalize-ai-ethics/>.
- Kang, Cecilia, Tiffany Hsu, Kevin Roose, Natasha Singer, and Matthew Rosenberg. 2018. Mark Zuckerberg Testimony: Day 2 Brings Tougher Questioning. *The New York Times*.
- Karoff, Paul. 2019. Embedding ethics in computer science curriculum. *The Harvard Gazette*.  
<https://news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/>.
- Khan, Lina M. 2017. "Amazon's Antitrust Paradox." *The Yale Law Journal* 126 (3):564-907.
- King, Leslie, and Julianne Busa. 2017. "When corporate actors take over the game: the corporatization of organic, recycling and breast cancer activism." *Social Movement Studies* 16 (5):549-563. doi: 10.1080/14742837.2017.1345304.
- Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent trade-offs in the fair determination of risk scores." *arXiv preprint arXiv:1609.05807*.
- Knight, Will. 2019. Google appoints an "AI council" to head off controversy, but it proves controversial. *MIT Technology Review*.  
<https://www.technologyreview.com/2019/03/26/136376/google-appoints-an-ai-council-to-head-off-controversy-but-it-proves-controversial/>.
- Lapowsky, Issie. 2018. Facebook Exposed 87 Million Users to Cambridge Analytica. *Wired*.  
<https://www.wired.com/story/facebook-exposed-87-million-users-to-cambridge-analytica/>.

- Legassick, Sean, and Verity Harding. 2017. Why we launched DeepMind Ethics & Society. *DeepMind Blog.* <https://deepmind.com/blog/announcements/why-we-launched-deepmind-ethics-society>.
- Lewis, Tamika, Seeta Peña Gangadharan, Mariella Saba, and Tawanna Petty. (2018). *Digital Defense Playbook: Community Power Tools for Reclaiming Data*. Our Data Bodies.
- Marantz, Andrew. 2019. Silicon Valley's Crisis of Conscience. *The New Yorker*. <https://www.newyorker.com/magazine/2019/08/26/silicon-valleys-crisis-of-conscience>.
- Marowski, Steve. 2020. Artificial intelligence researchers create ethics center at University of Michigan. *MLive*. <https://www.mlive.com/news/ann-arbor/2020/01/artificial-intelligence-researchers-create-ethics-center-at-university-of-michigan.html>.
- Martinho-Truswell, Emma, Hannah Miller, Isak Nti Asare, André Petheram, Richard Stirling, Constanza Gómez Mont, and Cristina Martínez. 2018. Hacia una Estrategia de IA en México: Aprovechando la Revolución de la IA (Towards an AI Strategy in Mexico: Leveraging the AI Revolution). [https://docs.wixstatic.com/ugd/7be025\\_ba24a518a53a4275af4d7ff63b4cf594.pdf](https://docs.wixstatic.com/ugd/7be025_ba24a518a53a4275af4d7ff63b4cf594.pdf).
- Marx, Leo, and Merritt Roe Smith. 1994. "Introduction." In *Does Technology Drive History?: The Dilemma of Technological Determinism*, edited by Merritt Roe Smith and Leo Marx. MIT Press.
- Mboya, Arwa. 2019. Why Joi Ito needs to resign. *The Tech*. <https://thetech.com/2019/08/29/joi-ito-needs-to-resign>.
- Metcalf, Jacob. 2014. Ethics Codes: History, Context, and Challenges. <https://bdes.datasociety.net/wp-content/uploads/2016/10/EthicsCodes.pdf>.
- Metcalf, Jacob, Emanuel Moss, and danah boyd. 2019. "Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics." *Social Research* 86 (2):449-476.
- Metzinger, Thomas. 2019. Ethics washing made in Europe. *Der Tagesspiegel*. <https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html>.
- Microsoft. 2018. Microsoft AI principles. <https://www.microsoft.com/en-us/ai/responsible-ai>.
- Mijente. 2019. 1,200+ Students at 17 Universities Launch Campaign Targeting Palantir. <https://notechforice.com/20190916-2/>.

- MIT Media Lab. 2017. MIT Media Lab to participate in \$27 million initiative on AI ethics and governance. *MIT News*. <https://news.mit.edu/2017/mit-media-lab-to-participate-in-ai-ethics-and-governance-initiative-0110>.
- MIT News Office. 2018. MIT reshapes itself to shape the future. *MIT News*. <http://news.mit.edu/2018/mit-reshapes-itself-stephen-schwarzman-college-of-computing-1015>.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. "Model Cards for Model Reporting." Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA.
- Mittelstadt, Brent. 2019. "Principles alone cannot guarantee ethical AI." *Nature Machine Intelligence* 1 (11):501-507. doi: 10.1038/s42256-019-0114-4.
- Morozov, Evgeny. 2014. *To Save Everything, Click Here: The Folly of Technological Solutionism: PublicAffairs*.
- Moss, Emanuel, and Jacob Metcalf. 2020. Too Big a Word. *Data & Society: Points*. <https://points.datasociety.net/too-big-a-word-13e66e62a5bf>.
- Mozilla. 2018. Announcing a Competition for Ethics in Computer Science, with up to \$3.5 Million in Prizes. *The Mozilla Blog*. <https://blog.mozilla.org/blog/2018/10/10/announcing-a-competition-for-ethics-in-computer-science-with-up-to-3-5-million-in-prizes/>.
- Mozilla. 2020. <https://twitter.com/mozilla/status/1308542908291661824>.
- Nadella, Satya. 2018. Embracing our future: Intelligent Cloud and Intelligent Edge. *Microsoft News Center*. <https://news.microsoft.com/2018/03/29/satya-nadella-email-to-employees-embracing-our-future-intelligent-cloud-and-intelligent-edge/>.
- National Science and Technology Council. 2018. Preparing for the Future of Artificial Intelligence. [https://obamawhitehouse.archives.gov/sites/default/files/whitehouse\\_files/microsites/ostp/NSTC/preparing\\_for\\_the\\_future\\_of\\_ai.pdf](https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf).
- Neate, Rupert. 2018. Twitter stock plunges 20% in wake of 1m user decline. *The Guardian*. <https://www.theguardian.com/technology/2018/jul/27/twitter-share-price-tumbles-after-it-loses-1m-users-in-three-months>.

- Nemitz, Paul. 2018. "Constitutional democracy and technology in the age of artificial intelligence." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133). doi: 10.1098/rsta.2018.0089.
- Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*: NYU Press.
- Noble, Safiya Umoja. 2019. <https://twitter.com/safiyanoble/status/1208812440403660800>.
- Novet, Jordan. 2018. Facebook forms a special ethics team to prevent bias in its A.I. software. *CNBC*. <https://www.cnbc.com/2018/05/03/facebook-ethics-team-prevents-bias-in-ai-software.html>.
- O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*: Broadway Books.
- Ochigame, Rodrigo. 2019. The Invention of “Ethical AI”: How Big Tech Manipulates Academia to Avoid Regulation. *The Intercept*. <https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/>.
- Organisation for Economic Co-operation and Development. 2019. Recommendation of the Council on Artificial Intelligence. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.
- Oxford Languages. 2018. Word of the Year 2018: Shortlist. *Oxford Languages*. <https://languages.oup.com/word-of-the-year/2018-shortlist/>.
- Oz, Effy. 1992. "Ethical Standards for Information Systems Professionals: A Case for a Unified Code." *MIS Quarterly*:423-433.
- Pardes, Arielle. 2018. Silicon Valley Writes a Playbook to Help Avert Ethical Disasters. *Wired*. <https://www.wired.com/story/ethical-os/>.
- Patil, D.J. 2018. A Code of Ethics for Data Science. <https://medium.com/@dpatil/a-code-of-ethics-for-data-science-cda27d1fac1>.
- Perlroth, Nicole, Sheera Frenkel, and Scott Shane. 2018. Facebook Exit Hints at Dissent on Handling of Russian Trolls. *The New York Times*. <https://www.nytimes.com/2018/03/19/technology/facebook-alex-stamos.html?mtrref=undefined>.

- Peters, Adele. 2018. This tool lets you see—and correct—the bias in an algorithm. *Fast Company*.  
<https://www.fastcompany.com/40583554/this-tool-lets-you-see-and-correct-the-bias-in-an-algorithm>.
- Phillips, Matt. 2018. Facebook's Stock Plunge Shatters Faith in Tech Companies' Invincibility. *The New York Times*. <https://www.nytimes.com/2018/07/26/business/facebook-stock-earnings-call.html>.
- Pichai, Sundar. 2018. AI at Google: our principles. <https://www.blog.google/technology/ai/ai-principles/>.
- Pinch, Trevor J., and Wiebe E. Bijker. 1987. "The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other." In *The Social Construction of Technological Systems*, edited by Wiebe E. Bijker, Thomas P. Hughes and Trevor Pinch. MIT Press.
- Raicu, Irina. 2017. Rethinking Ethics Training in Silicon Valley. *The Atlantic*.  
<https://www.theatlantic.com/technology/archive/2017/05/rethinking-ethics-training-in-silicon-valley/525456/>.
- Read, Max. 2016. Donald Trump Won Because of Facebook. *New York Magazine*.  
<https://nymag.com/intelligencer/2016/11/donald-trump-won-because-of-facebook.html>.
- Reardon, Jenny. 2011. "Human Population Genomics and the Dilemma of Difference." In *Reframing Rights: Bioconstitutionalism in the Genetic Age*, edited by Sheila Jasanoff, 217-238.
- Reardon, Jenny. 2013. "On the Emergence of Science and Justice." *Science, Technology, & Human Values* 38 (2):176-200. doi: 10.1177/0162243912473161.
- Reich, Rob, Mehran Sahami, Jeremy M. Weinstein, and Hilary Cohen. 2020. "Teaching Computer Ethics: A Deeply Multidisciplinary Approach." Proceedings of the 51st ACM Technical Symposium on Computer Science Education, Portland, OR, USA.
- Ritchie, Hannah. 2016. Read all about it: The biggest fake news stories of 2016. *CNBC*.  
<https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html>.
- Robinson, David, and Miranda Bogen. 2016. Data Ethics: Investing Wisely in Data at Scale. *Upturn*. [https://www.upturn.org/static/reports/2016/data-ethics/files/Upturn - Data%20Ethics\\_v.1.0.pdf](https://www.upturn.org/static/reports/2016/data-ethics/files/Upturn - Data%20Ethics_v.1.0.pdf).

- Rose, Kevin. 2019. The Making of a YouTube Radical. *The New York Times*.  
<https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html>.
- Rosenberg, Matthew, Nicholas Confessore, and Carole Cadwalladr. 2018. How Trump Consultants Exploited the Facebook Data of Millions. *The New York Times*.  
<https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>.
- Schneier, Bruce. 2016. *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*: W. W. Norton & Company.
- Seetharaman, Deepa. 2020. Jack Dorsey's Push to Clean Up Twitter Stalls, Researchers Say. *The Wall Street Journal*. <https://www.wsj.com/articles/jack-dorseys-push-to-clean-up-twitter-stalls-researchers-say-11584264600>.
- Selznick, Philip. 1948. "Foundations of the Theory of Organization." *American Sociological Review* 13 (1):25-35. doi: 10.2307/2086752.
- Shane, Scott, and Daisuke Wakabayashi. 2018. 'The Business of War': Google Employees Protest Work for the Pentagon. *The New York Times*.  
<https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html>.
- Sharlach, Molly. 2019. Princeton collaboration brings new insights to the ethics of artificial intelligence. <https://www.princeton.edu/news/2019/01/14/princeton-collaboration-brings-new-insights-ethics-artificial-intelligence>.
- Shilton, Katie, Michael Zimmer, Casey Fiesler, Arvind Narayanan, Jake Metcalf, Matthew Bietz, and Jessica Vitak. 2017. We're Awake — But We're Not At the Wheel. *PERVADE: Pervasive Data Ethics*. <https://medium.com/pervade-team/were-aware-but-we-re-not-at-the-wheel-7f0a7193e9d5>.
- Silbey, Susan S. 2018. How Not to Teach Ethics. *MIT Faculty Newsletter*.  
<https://web.mit.edu/fnl/volume/311/silbey.html>.
- Simonite, Tom. 2018. Should Data Scientists Adhere to a Hippocratic Oath? *Wired*.  
<https://www.wired.com/story/should-data-scientists-adhere-to-a-hippocratic-oath/>.
- Simonite, Tom. 2019. Google and Microsoft Warn That AI May Do Dumb Things. *Wired*.  
<https://www.wired.com/story/google-microsoft-warn-ai-may-do-dumb-things/>.

- Simonite, Tom. 2020. Google Offers to Help Others With the Tricky Ethics of AI. *Wired*.  
<https://www.wired.com/story/google-help-others-tricky-ethics-ai/>.
- Sinders, Caroline. 2019. <https://twitter.com/carolinesinders/status/1208443559998873601>.
- Smart Dubai. 2018. AI Ethics Principles & Guidelines. [https://www.smartdubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf?sfvrsn=d4184f8d\\_6](https://www.smartdubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf?sfvrsn=d4184f8d_6).
- Smith, Brad. 2018. Facial recognition: It's time for action. *Microsoft On The Issues*.  
<https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/>.
- Solon, Olivia. 2016. Facebook's failure: did fake news and polarized politics get Trump elected? *The Guardian*. <https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-election-conspiracy-theories>.
- Solon, Olivia. 2019. Why did Microsoft fund an Israeli firm that surveils West Bank Palestinians? *NBC News*. <https://www.nbcnews.com/news/all/why-did-microsoft-fund-israeli-firm-surveils-west-bank-palestinians-n1072116>.
- Stark, Luke, and Anna Lauren Hoffmann. 2019. "Data Is The New What?: Popular Metaphors & Professional Ethics in Emerging Data Cultures." *Journal of Cultural Analytics*. doi: 10.22148/16.036.
- Stop LAPD Spying Coalition and Free Radicals. 2020. The Algorithmic Ecology: An Abolitionist Tool for Organizing Against Algorithms. <https://medium.com/@stoplapdspying/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms-14fcbd0e64d0>.
- Suchman, Lucy, Jeanette Blomberg, Julian E. Orr, and Randall Trigg. 1999. "Reconstructing Technologies as Social Practice." *American Behavioral Scientist* 43 (3):392-408. doi: 10.1177/00027649921955335.
- The Institute for the Future, and Omidyar Network. 2018. Ethical OS Toolkit. <https://ethicalos.org>.
- Tiku, Nitasha. 2019. Three Years of Misery Inside Google, the Happiest Company in Tech. *Wired*.  
<https://www.wired.com/story/inside-google-three-years-misery-happiest-company-tech/>.
- Toyama, Kentaro. 2015. *Geek Heresy: Rescuing Social Change from the Cult of Technology*: PublicAffairs.
- Tracy, Marc, and Tiffany Hsu. 2019. Director of M.I.T.'s Media Lab Resigns After Taking Money From Jeffrey Epstein. *The New York Times*.

<https://www.nytimes.com/2019/09/07/business/mit-media-lab-jeffrey-epstein-joichi-ito.html>.

Trumpy, Alexa J. 2014. "Subject to Negotiation: The Mechanisms Behind Co-Optation and Corporate Reform." *Social Problems* 55 (4):480-500. doi: 10.1525/sp.2008.55.4.480.

U.S. Department of Defense. 2020. DOD Adopts Ethical Principles for Artificial Intelligence. <https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>.

Varshney, Kush R. 2018. Introducing AI Fairness 360. *IBM Research Blog*. <https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/>.

Vincent, James. 2016. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>.

Vincent, James, and Russell Brandom. 2018. Axon launches AI ethics board to study the dangers of facial recognition. *The Verge*. <https://www.theverge.com/2018/4/26/17285034/axon-ai-ethics-board-facial-recognition-racial-bias>.

Visvanathan, Shiv. 2005. "Knowledge, justice and democracy." In *Science and Citizens: Globalization and the Challenge of Engagement.*, edited by Melissa Leach, Ian Scoones and Brian Wynne. Zed Books.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359 (6380):1146-1151. doi: 10.1126/science.aap9559.

Wagner, Ben. 2018. "Ethics as Escape From Regulation: From Ethics-Washing to Ethics-Shopping?" In *Being Profiling. Cogitas Ergo Sum*, edited by Emre Bayamlioglu, Irina Baraliuc, Liisa Albertha Wilhelmina Janssens and Mireille Hildebrandt. Amsterdam University Press.

Walker, Kent. 2018. Google AI Principles updates, six months in. *The Keyword*. <https://www.blog.google/technology/ai/google-ai-principles-updates-six-months/>.

West, Sarah Myers, Meredith Whittaker, and Kate Crawford. 2019. Discriminating Systems: Gender, Race, and Power in AI. <https://ainowinstitute.org/discriminatingsystems.pdf>.

Williams, Oscar. 2019. How Big Tech funds the debate on AI ethics. *New Statesman*. <https://www.newstatesman.com/science-tech/technology/2019/06/how-big-tech-funds-debate-ai-ethics>.

- Winner, Langdon. 1986. *The Whale and the Reactor: A Search for Limits in an Age of High Technology*: University of Chicago Press.
- Wood, Greg, and Malcolm Rimmer. 2003. "Codes of Ethics: What Are They Really and What Should They Be?" *International Journal of Value-Based Management* 16 (2):181-195.
- Woodman, Spencer. 2017. Palantir Provides the Engine for Donald Trump's Deportation Machine. *The Intercept*. <https://theintercept.com/2017/03/02/palantir-provides-the-engine-for-donald-trumps-deportation-machine/>.
- Wright, Susan. 2001. "Legitimizing Genetic Engineering." *Perspectives in Biology and Medicine* 44 (2):235-247.
- Wu, Tim. 2018. *The Curse of Bigness: Antitrust in the New Gilded Age*: Columbia Global Reports.
- Zuboff, Shoshana. 2018. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*: PublicAffairs.
- Zunger, Yonatan. 2018. Computer science faces an ethics crisis. The Cambridge Analytica scandal proves it. *The Boston Globe*. <https://www.bostonglobe.com/ideas/2018/03/22/computer-science-faces-ethics-crisis-the-cambridge-analytica-scandal-proves/IzaXxl2BsYBtwM4nxezgcP/story.html>.



Government  
Digital Service

**DATA**  
**ETHICS**  
**FRAMEWORK**

# Data Ethics Framework

© Crown copyright 2020

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit [nationalarchives.gov.uk/doc/open-government-licence/version/3](https://nationalarchives.gov.uk/doc/open-government-licence/version/3)

This publication is also available on our website at [www.gov.uk/government/publications/data-ethics-framework](https://www.gov.uk/government/publications/data-ethics-framework)

# Data Ethics Framework

## What is it for?

The Data Ethics Framework guides appropriate and responsible data use in government and the wider public sector. It helps public servants understand ethical considerations, address these within their projects, and encourages responsible innovation.

## How to use it?

Teams should work through the framework together throughout the process of planning, implementing, and evaluating a new project. Each part of the framework is designed to be regularly revisited throughout your project, especially when any changes are made to your data collection, storage, analysis or sharing processes.

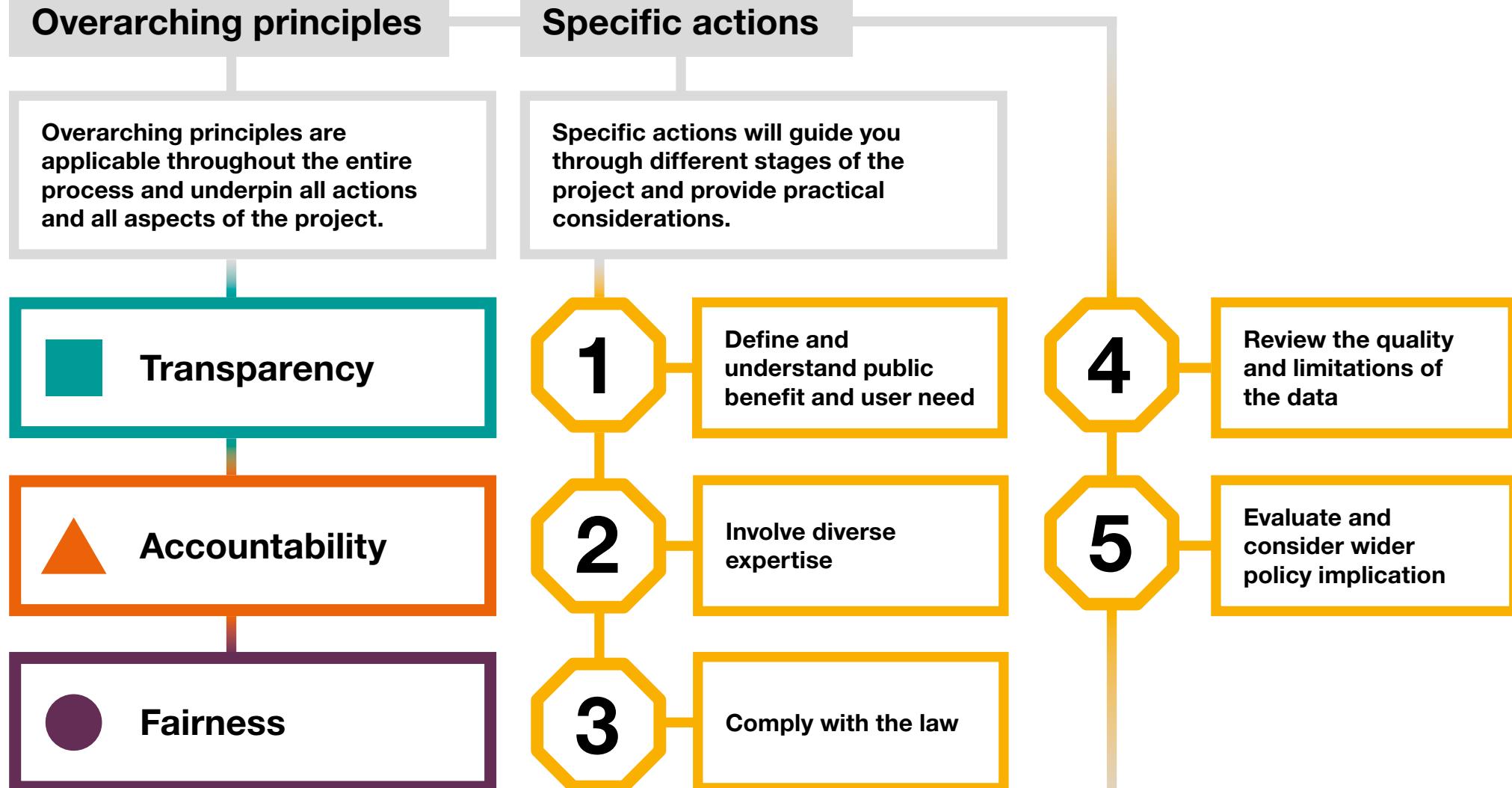
## Who is it for?

This guidance is aimed at anyone working directly or indirectly with data in the public sector, including data practitioners (statisticians, analysts and data scientists), policymakers, operational staff and those helping to produce data-informed insight.

## Structure

The framework is split into **overarching principles** and **specific actions**. Overarching principles are applicable throughout the entire process and underpin all actions and all aspects of the project. Specific actions will guide you through different stages of the project and provide practical considerations.

In addition, the framework provides specific actions you can take at each stage of the project to advance **transparency**, **accountability**, and **fairness**. These are marked within each section.

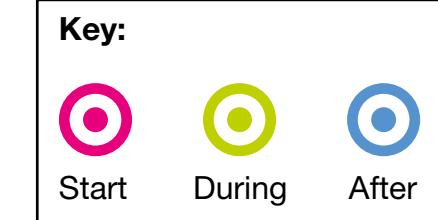


# Overarching principles

## Transparency

Transparency means that your actions, processes and data are made open to inspection by publishing information about the project in a complete, open, understandable, easily-accessible, and free format. In your work with and on data and AI, use the available guidance, e.g. the Open Government Playbook, to ensure transparency throughout the entirety of your process.

Transparency					
Score					
0	1	2	3	4	5
Information about the project, its methods, and outcomes is not publicly available					
					Information about the project, its methods, and outcomes is widely available to the public





## Accountability

Accountability means that there are effective governance and oversight mechanisms for any project. Public accountability means that the public or its representatives are able to exercise effective oversight and control over the decisions and actions taken by the government and its officials, in order to guarantee that government initiatives meet their stated objectives and respond to the needs of the communities they are designed to benefit.

Accountability					
Score					
0	1	2	3	4	5
Mechanisms for scrutiny, governance, or peer review for the project haven't been established					
					Long-term oversight and public scrutiny mechanisms are built into the project cycle

### Key:



Start



During



After

## Fairness

Fairness — It is crucial to eliminate your project's potential to have unintended discriminatory effects on individuals and social groups. You should aim to mitigate biases which may influence your model's outcome and ensure that the project and its outcomes respect the dignity of individuals, are just, non-discriminatory, and consistent with the public interest, including human rights and democratic values.

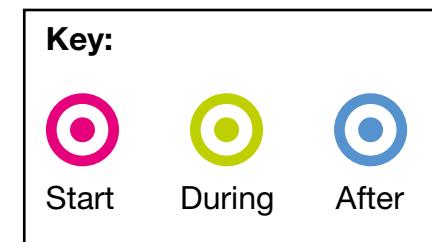
You can read more about fairness and its different types in the [Understanding artificial intelligence ethics and safety](#) guide developed by the Government Digital Service and the Office for Artificial Intelligence.

You can read more about the standards regarding human rights in AI and machine learning in the [Toronto Declaration](#) developed by Access Now and Amnesty International.

Fairness					
Score					
0	1	2	3	4	5
There is a significant risk that the project will result in harm or detrimental and discriminatory effects for the public or certain groups					
					The project promotes just and equitable outcomes, has negligible detrimental effects, and is aligned with human rights considerations
Key:					
			Start	During	After

# Self-assessment tables for the 3 overarching principles.

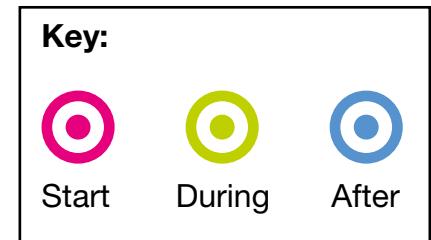
Please use the following self assessment to mark, track and share your overall progress over time against the following 3 overarching principles.



	Principles	Score					
		0	1	2	3	4	5
	Transparency						
	Accountability						
	Fairness						

# Self-assessment tables for the 5 specific actions.

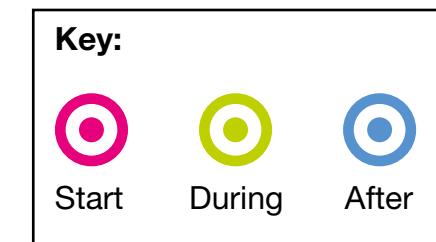
Please use the following self assessment to mark, track and share your overall progress over time against the following 5 specific actions.



	<b>Specific actions</b>	<b>Score</b>					
		0	1	2	3	4	5
<b>1</b>	Define public benefit and user need						
<b>2</b>	Involve diverse expertise						
<b>3</b>	Comply with the law						
<b>4</b>	Check the quality and limitations of the data						
<b>4.1</b>	Check the quality and limitations of the model						
<b>5</b>	Evaluate and consider wider policy implications						

# Self-assessment tables for the 3 overarching principles.

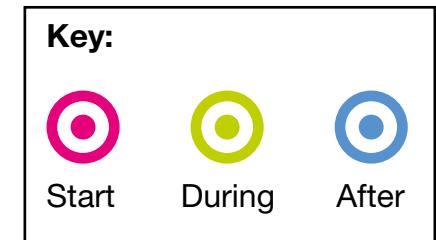
Please use the following self assessment to mark, track and share your overall progress over time against the following 3 overarching principles.



	Principles	Score					
		0	1	2	3	4	5
	Transparency						
	Accountability						
	Fairness						

# Self-assessment tables for the 5 specific actions.

Please use the following self assessment to mark, track and share your overall progress over time against the following 5 specific actions.



		<b>Score</b>					
<b>Specific actions</b>		0	1	2	3	4	5
<b>1</b>	Define public benefit and user need						
<b>2</b>	Involve diverse expertise						
<b>3</b>	Comply with the law						
<b>4</b>	Check the quality and limitations of the data						
<b>4.1</b>	Check the quality and limitations of the model						
<b>5</b>	Evaluate and consider wider policy implications						

# **Define and understand public benefit and user need**

When starting a public sector data project, you must have a clear articulation of its purpose.

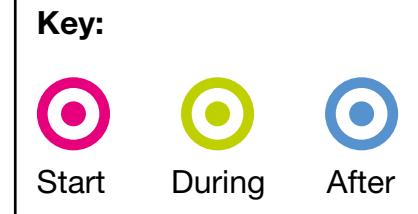
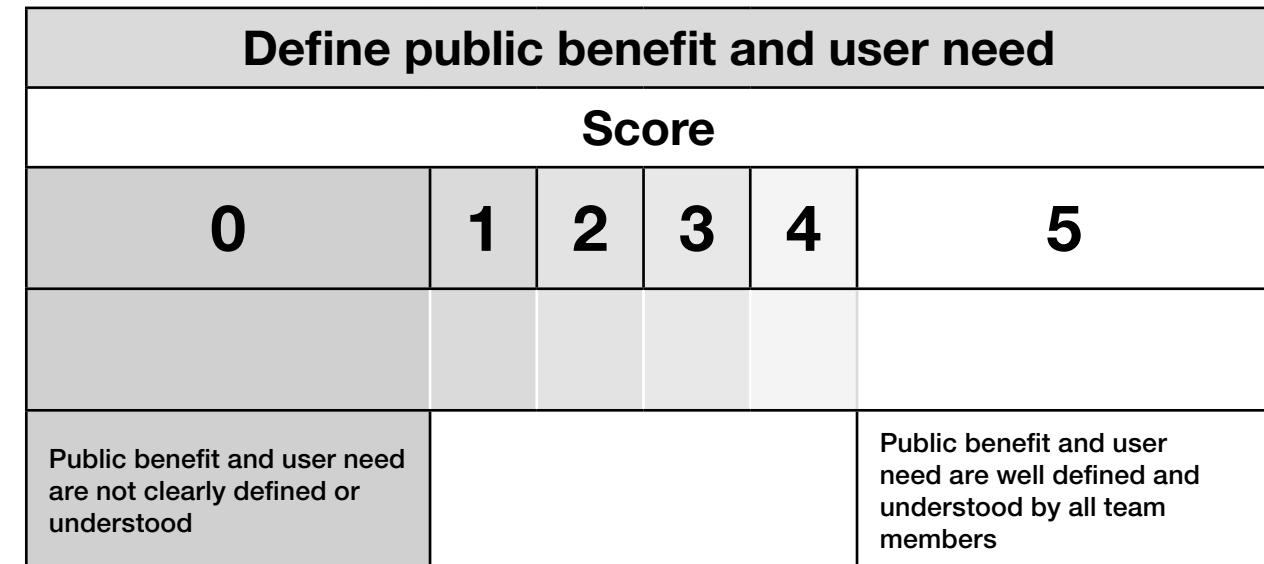
This includes having clarity on what public benefit the project is trying to achieve and what are the needs of the people who will be using the service or will be most directly affected by it.

1

# Define and understand public benefit and user need

## When starting a public sector data project...

Using the circles move them to where your project is on the spectrum. Do this at the start of a project, during a project and after a project has completed. This will help you track, communicate and share your progress.





# Define and understand public benefit and user need

## Public Benefit

1.1

### Understand the wider public benefit

- What are direct benefits for individuals in this project? (e.g. saving time when applying for a government service)
- How does the project deliver positive social outcomes for the wider public?
- How can you measure and communicate the benefits of this project to the public?
- What are the groups that would be disadvantaged by the project/ that would not benefit from the project? What can you do about this?

1.2

### Understand unintended and or negative consequences of your project

- What would be the harm in not using data? What social outcomes might not be met?
- What are the potential risks or negative consequences of the project versus the risk in not proceeding with the project?
- Could the misuse of the data/algorithm or poor design of the project contribute to reinforcing social and ethical problems and inequalities?
- What kind of mechanisms can you put in place to prevent this from happening?
- What specific groups benefit from the project? What groups can be denied opportunities/face negative consequences because of the project?

1.3

## Human rights considerations

1.4

### Justify the benefit for the tax-payers and appropriate use of public resources in your project

- How can you demonstrate the value for money of your project?
- Is there effective governance and decision-making oversight to ensure success of the project?
- Do you have evidence to demonstrate all of the above?



- How does the design and implementation of the project/ algorithm respect human rights and democratic values?
- How does the project/algorithm work towards advancing human capabilities, advancing inclusion of underrepresented populations, reducing economic, social, gender, racial, and other inequalities?
- What are the environmental implications of the project? How could they be mitigated?

1.5

### Make your user need and public benefit transparent

- Where can you publish information on how the project delivers positive social outcomes for the public?
- How have you shared your understanding of the user need with the user?

# User need

1.7

## Ensure there is a clear articulation of the problem before you start the project.

Describe the user need in your project:

- As a ...
- I need/want/expect to... [what does the user want to do?]
- So that... [why does the user want to do this?]

Example 1: the [Register to vote service's](#) user need:

- As a UK resident
- I want to get my details on the online electoral register
- So that I can vote

Example 2: user need when building a platform for fire safety checks:

- As a data analyst working in a fire and rescue service
- I need to identify homes which are likely to not have a smoke alarm fitted

- So that I can advise how to prioritise fire safety checks

For various user needs, using data can:

- help you identify themes in large volumes of text
- predict what will happen
- automatically categorise stuff
- spot something unusual
- show you how things are connected to each other
- spot patterns in large volumes of data
- spot geographic patterns in services or data

1.6

## Understand the user need

User needs' are the needs that a user has of a service, and which that service must satisfy for the user to get the right outcome for them. For more information about the user need, consult the [GDS Service Manual](#).

Example:

- running and improving services
- building new services
- trialling new processes for internal operations
- testing existing and new policies

1.8

## Check if everyone in your team understands the user need and how using data can help.

Does everyone in your team understand the user need?

- Often projects involving data analysis are requested by non-practitioners
  - people with an ill-defined problem they would like to understand better. Reframing their request as a user need will help you understand what they're asking for and why, or expose what you don't know yet.

1.9

## Repeatedly revisit your user need throughout the project.

Consider these questions as the project evolves:

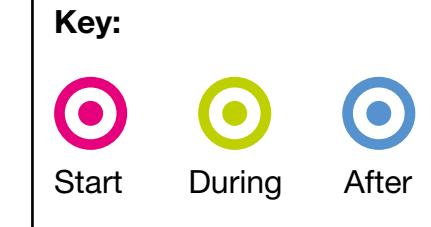
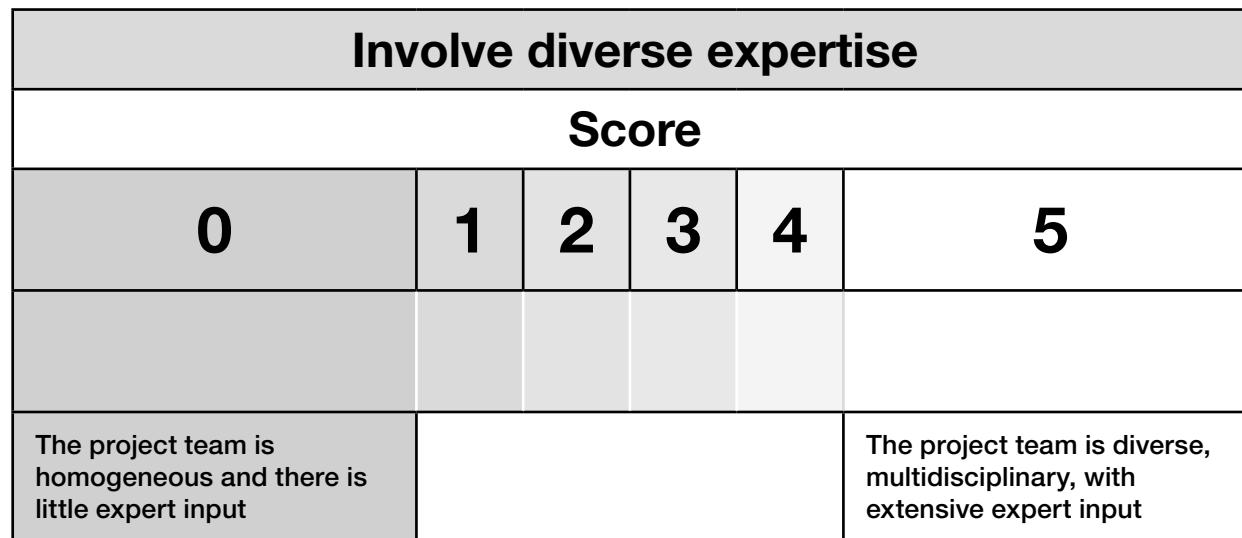
- What is the overall problem you're trying to solve?
- Who are the users of this data process or analysis?
- What needs do they have?

# Involve diverse expertise

Working in diverse, multidisciplinary teams with wide ranging skill sets contributes to the success of any data or tech project. If you do not have the sufficient skills or experience, you should involve others from your team or wider network with the right expertise.

2

# Involve diverse expertise





# 2 Involve diverse expertise

2.1

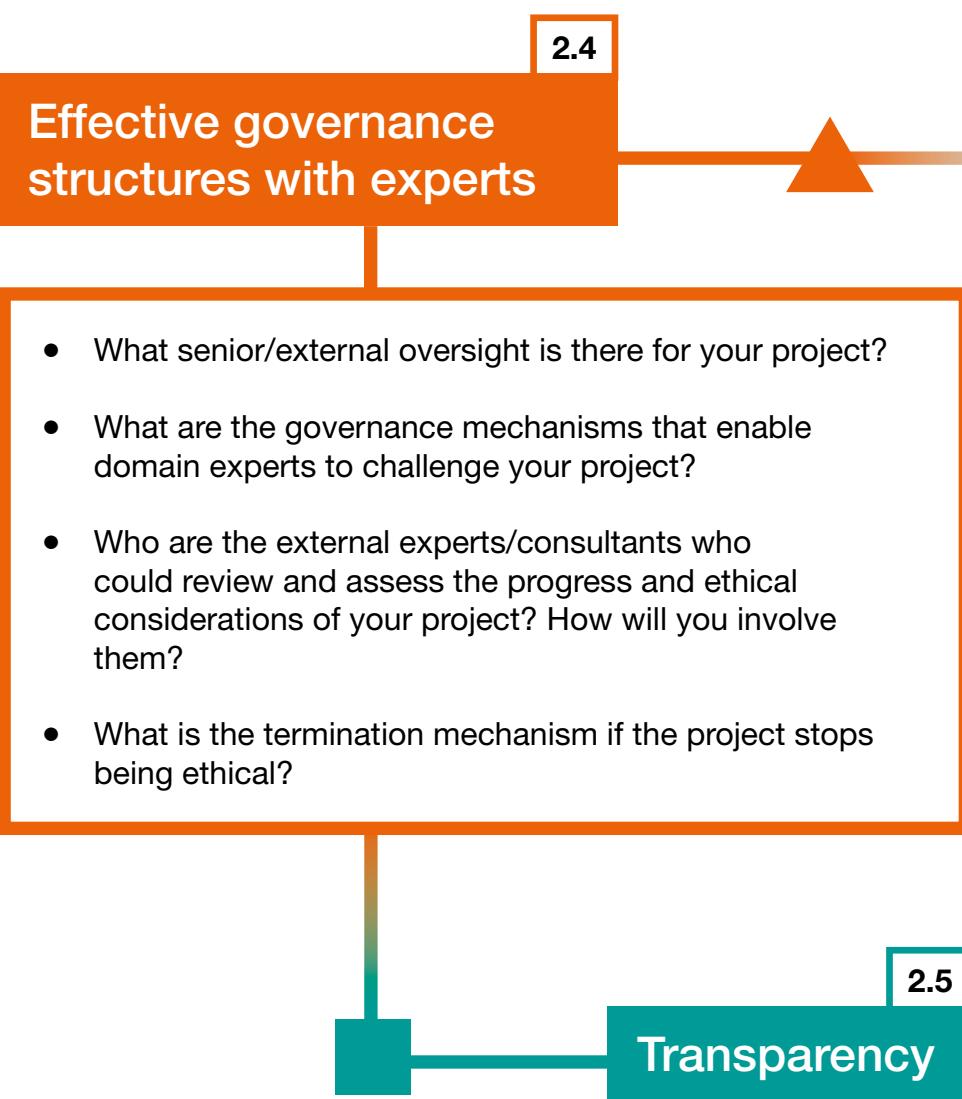
## Get the right expertise

- Beyond data scientists, who are the relevant policy experts and practitioners in your team?
- What other disciplines and subject matter experts need to be involved? What steps have been taken to involve them?
- What should their roles be? Have you defined who does what in the project?
- Ask your team and experts if you have the right data for your research questions. Get their help in matching the dataset to the problem.

2.2

## Ensure diversity within your team

- How have you ensured diversity in your team? Having a diverse team helps prevent biases and encourages more creativity and diversity of thought.
- Avoid forming homogenous teams, embrace diversity of lived experiences of people from different backgrounds. If you find yourself in a homogenous team, challenge it.



## 2.3

### Involve external stakeholders

- How have you engaged external domain experts in your project (e.g. academics, ethicists, researchers)?
- Have you consulted the relevant civil society organisations?
- What is the impact that external engagement/consultations have had on the project?
- Have you considered consulting the target audience or the users of your project? This could be done through a range of deliberative processes and consultations.

# Comply with the law

You must have an understanding of the relevant laws and codes of practice that relate to the use of data. When in doubt, you must consult relevant experts.

3

# Comply with the law

Comply with the law					
Score					
0	1	2	3	4	5
There is little clarity on legal requirements for the project					Relevant legal requirements have been met, compulsory assessments completed, legal experts have been consulted

Key:



Start

During

After

# 3

# Comply with the law

3.1

## Get legal advice

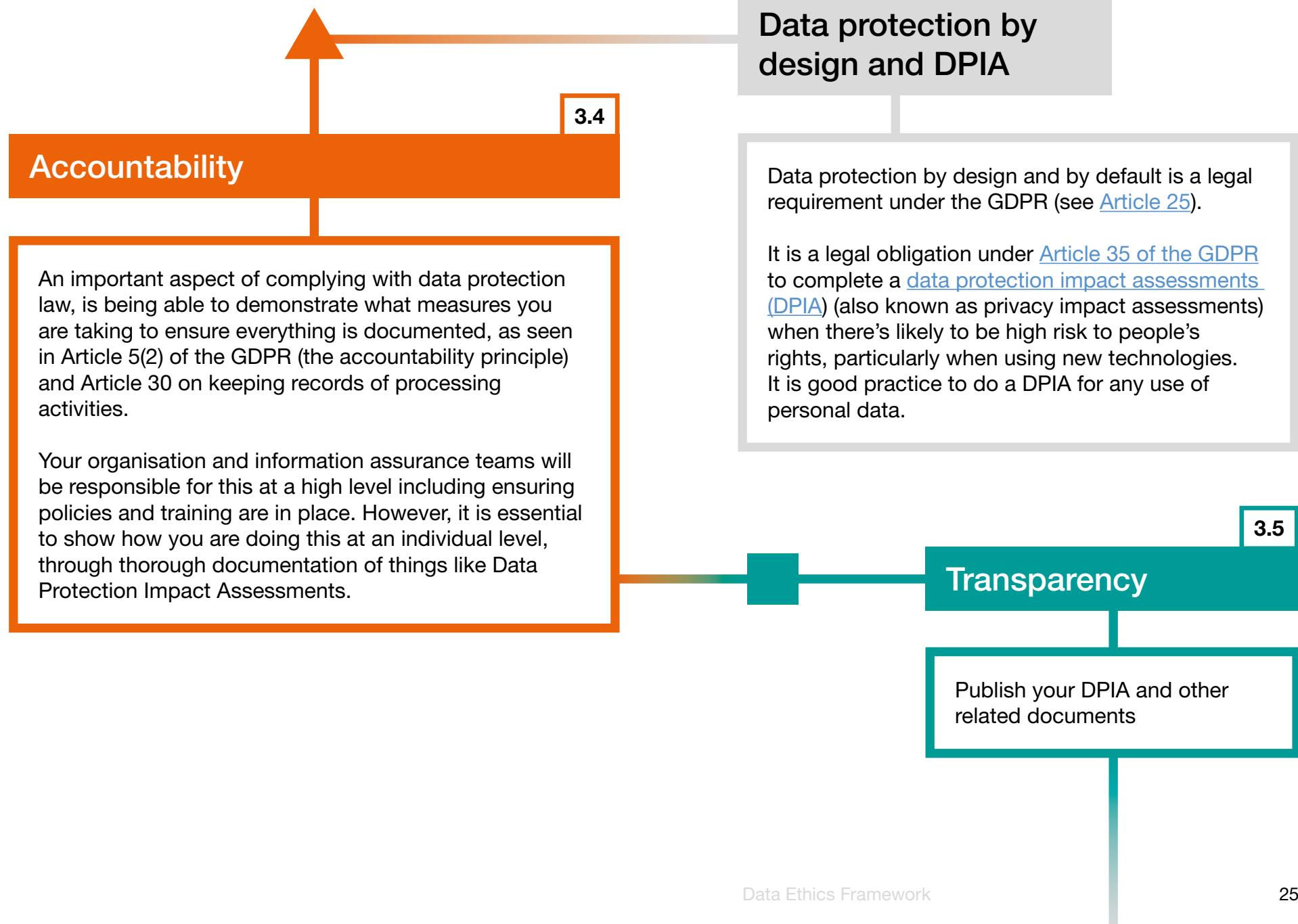
- Have you spoken to a legal adviser within your organisation?
- Have you spoken to your information assurance team?
- Have you consulted your organisation's Data Protection Officer when doing a DPIA?
- What legal advice have you received?

3.2

**It is your duty and obligation to obey the law in any data projects. You must ensure the project's compliance with GDPR and DPA 2018**

If you are using personal data, you must comply with the principles of the [EU General Data Protection Regulation \(GDPR\)](#) and [Data Protection Act 2018 \(DPA 2018\)](#) which implements aspects of the GDPR and transposes the [Law Enforcement Directive](#) into UK law. It also provides separate processing regimes for activities which fall outside the scope of EU law.

Personal data is defined in [Section 3\(2\) DPA 2018](#) (a wider explanation is detailed in [Article 4 of the GDPR](#)).



**3.6**

## Ensure the project's compliance with the Equality Act 2010

Data analysis or automated decision making must not result in outcomes that lead to discrimination as defined in the [Equality Act 2010](#).

- How can you demonstrate that your project meets the Public Sector Equality Duty?
- What was the result of the Equality Impact Assessment of the project?

**3.7**

## Ensure effective governance of your data

Organisations have a responsibility to keep both [personal data](#) and non-personal data secure.

- How have you ensured that the project is compliant with data governance policies within your organisation?

**3.8**

## Ensure your project's compliance with any additional regulations

Consider [additional relevant legislation](#) and [codes of practice](#).

# **Review the quality and limitations of the data**

Insights from new technology are only as good as the data and practices used to create them. You must ensure that the data for the project is accurate, representative, proportionally used, of good quality, and that you are able to explain its limitations.

**4**

# Review the quality and limitations of the data

Review the quality and limitations of the data					
Score					
0	1	2	3	4	5
Data for the project is of bad quality, unsuitable, unreliable, not-representative					
The model is not reproducible and is likely to produce invalid outputs					Data used in the project is representative, proportionally used, accurate, and of good quality The model is reproducible and able to produce valid outputs

Key:

  
Start  
During  
After



# Review the quality and limitations of the data

4.1

## Data source

- What data source(s) is being used?
- Are individuals and/or organisations providing the data aware of how it will be used? If the user is repurposing data for analysis without individual consent, how have you ensured that the new purpose is compatible with the original reason for collection (Article 6 (4) GDPR)?
- Are all metadata and field names clearly understood?
  - Do you understand how the data for the project is generated? Remember that depending on where the data came from, the field may not represent what the field name/metadata indicates.
- What processes do you have in place to ensure and maintain data integrity?
- What are the caveats? How will the caveats be taken into account for any future policy or service which uses this work as an evidence base?
- Would using synthetic data be appropriate for the project? Synthetic data is entirely fabricated or abstracted from real data through various processes, e.g. anonymisation or record switching. It is often created with specific features to test or train an algorithm.

4.3

## Bias in data

- How has the data being used to train a model been assessed for potential bias? You should consider:
  - Whether the data might (accurately) reflect biased historical practice that you do not want to replicate in the model (historical bias)
  - The data might be a biased misrepresentation of historical practice, e.g. because only certain categories of data were properly recorded in a format accessible to the project (selection bias)
- If using data about people, is it possible that your model or analysis may be identifying proxy variables for protected characteristics which could lead to a discriminatory outcome? Such proxy variables can potentially be a cause of indirect discrimination; you should consider whether the use of these variables is appropriate in the context of your service (i.e. is there a reasonable causal link between the proxy variable and the outcome you're trying to measure?; do you assess this to be a proportionate means to achieve a legitimate aim in accordance with the Equality Act 2010?)
- What measures have you taken to mitigate bias?

4.2

## Determining proportionality

You must use the minimum data necessary to achieve your desired outcome ([Article 5\(1\)\(c\)](#) of the GDPR). Personal data should be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

- How can you meet the project aim using the minimum personal data possible?
- Is there a way to achieve the same aim with less identifiable data, e.g. pseudonymised data?
- If using [personal data identifying individuals](#), what measures are in place to control access?
- Is the input data suitable and necessary to achieve the aim?

- Would the proposed use of data be deemed inappropriate by those who provided the data (individuals and/or organisations)?
- Would the proposed use of data for secondary purposes make it less likely that people would want to give that data again for the primary purpose it was collected for?
- How can you explain why you need to use this data to members of the public?
- Does this use of data interfere with the rights of individuals? If yes, is there a less intrusive way of achieving the objective?

4.4

## Data anonymisation

If you plan to anonymise or pseudonymise personal data before linking or analysis, make sure you follow the ICO's [Anonymisation: managing data protection risk code of practice](#) and document your methods. You can find more technical advice in the UK Anonymisation Networks [anonymisation guidance](#).

If the assumption is that data in the project is anonymised, consider the following:

- How can you demonstrate that the data has been de-identified to the greatest degree possible?
- Can the data be matched to any other datasets that will make individuals easily identifiable? What measures have you taken to mitigate this? Have you considered any [determined intruder testing](#)?

4.5

## Robust practices

- If necessary, how can you (or external scrutiny) validate that the algorithm is achieving the correct output decision when new data is added?
- How can you demonstrate that you have designed the project for reproducibility?
  - Could another analyst repeat your procedure based on your documentation? Have they tried?
  - Have you followed the 3 requirements for reproducibility? (Applying the same logic (code, model or process), to the same data, in the same environment).
  - How have you ensured that high quality documentation will be kept?
  - [Have you considered Reproducible Analytical Pipelines? \(RAP\)](#)
- How confident are you that the algorithm is robust, and that any assumptions are met?
- What is the quality of the model outputs, and how does this stack up against the project objectives?

4.6

## Make your data open and shareable whenever possible

4.7

### Share your models - developed data science tools should be made available for scrutiny wherever possible.

- Can you openly publish your methodology, metadata about your model, and/or the model itself e.g. on Github?

There are 2 main types of algorithms used in data science.

The first is the algorithmic methodology used to train a model. It's often more useful and clear to share a document describing the analytical process than the code.

The second is the trained model itself (the result of applying the methodology to the dataset). Releasing this model allows others to scrutinise and test it, and may highlight issues that you can fix as part of your continual improvement.

When sharing models it's important that it does not endanger either the:

- privacy of those whose data was used to train it
- integrity of the task being undertaken.

If data is non-sensitive, non-personal, and if Data Sharing Agreements with the supplier allow it, you should make the data open and assign it a digital object identifier (DOI). For example, scientists share data when publishing a paper on [Figshare](#) and [Datadryad](#). This gives others access to the data and the code, so the analysis can be reproduced. You can also publish data on [Find open data](#) and the [UK Data Archive](#).

4.8

## How to ensure transparency of sensitive models

- How are you planning to inform the public about the model?
- Even if the model cannot be released publicly, you may be able to release metadata about the model on a continual basis, like its performance on certain datasets.
- If your data science application is very sensitive, you could arrange for selected external bodies, approved by your organisation, to examine the model itself in a controlled context to provide feedback. This could be expertise from another government department, academia or public body.

4.9

## Explainability

Explainability is the extent to which the workings in a machine learning algorithm can be explained in human terms. It means expanding on the transparency of what variables are used to provide information on how the algorithm came to give an output, and how changing the inputs can change the output.

- Explain what your project does and how it was designed in plain language to a non-expert audience.
- Describe the process and the aim of your algorithm, as well as what variables are used for what outcomes without using technical terms.
- Make this explanation publicly available (e.g. on GitHub, blogs, or gov.uk).

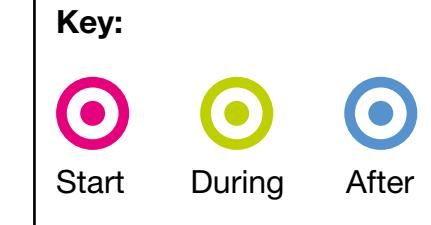
# Evaluate and consider wider policy implications

It is essential that there is a plan to continuously evaluate if insights from data are used responsibly. This means that both development and implementation teams understand how findings and data models should be used and monitored with a robust evaluation plan and effective accountability mechanisms.

5

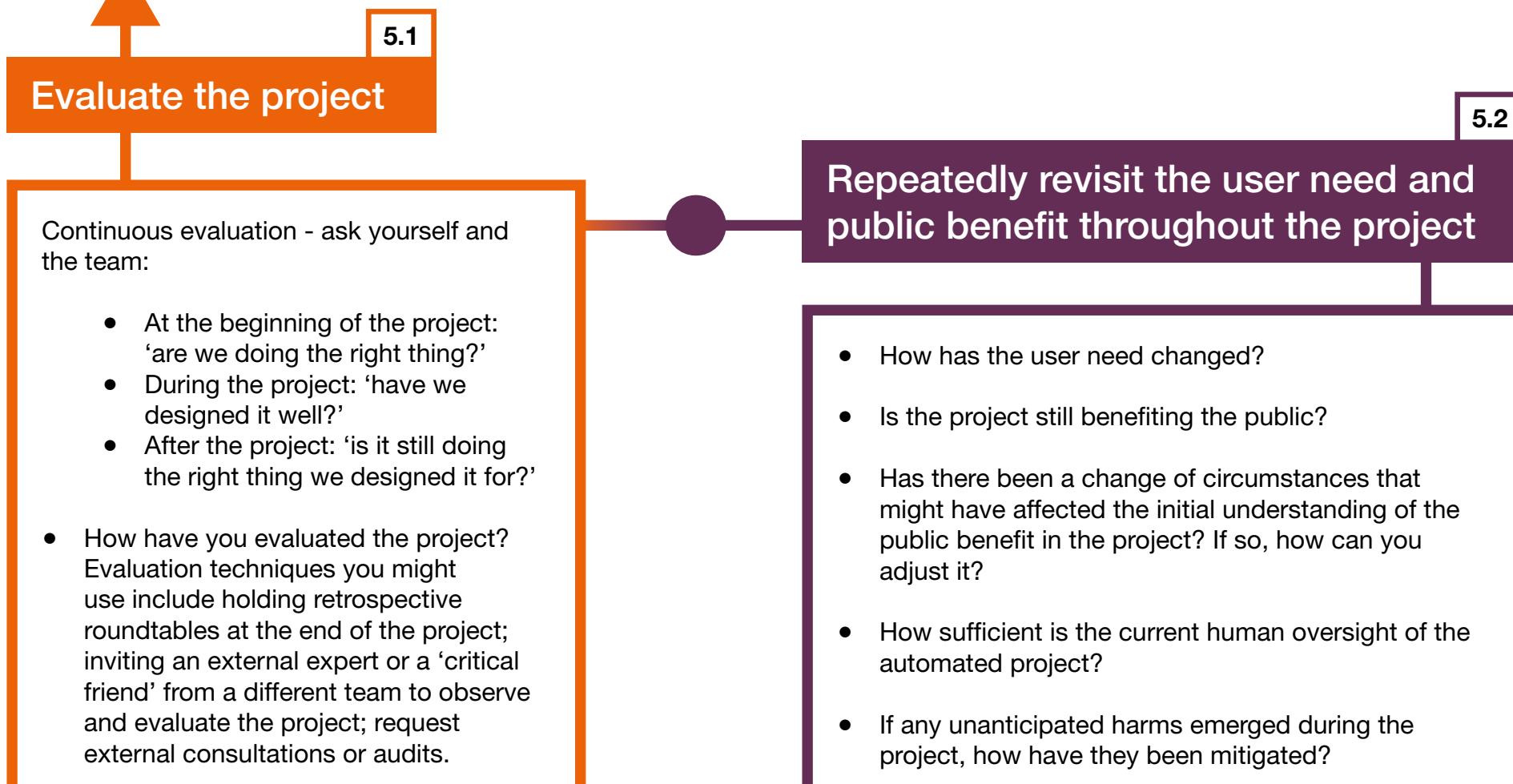
# Evaluate and consider wider policy implications

Evaluate and consider wider policy implications					
Score					
0	1	2	3	4	5
There are no long-term evaluation and maintenance structures in place				Continuous evaluation and long-term maintenance structures are in place	





# Evaluate and consider wider policy implications



5.4

## Ensure there are skills, training, maintenance for longevity of the project

- How have you ensured that the users have the appropriate support and training to maintain the new technology?
- What is the longevity of the project?
- How have you checked if the users understand the software they need to maintain the project? Have they been appropriately trained?

5.3

## Check how your project influences policy

- How accurately are the insights from the project used in the practical policy context?
- How have you ensured if policymakers and secondary users of the tool fully understand its purpose and structure?

5.5

## Accountability structures

- What are the governance structures in place to ensure a safe and sustainable implementation of the project?
- How often will you update the board/governing bodies?

5.6

## Public scrutiny

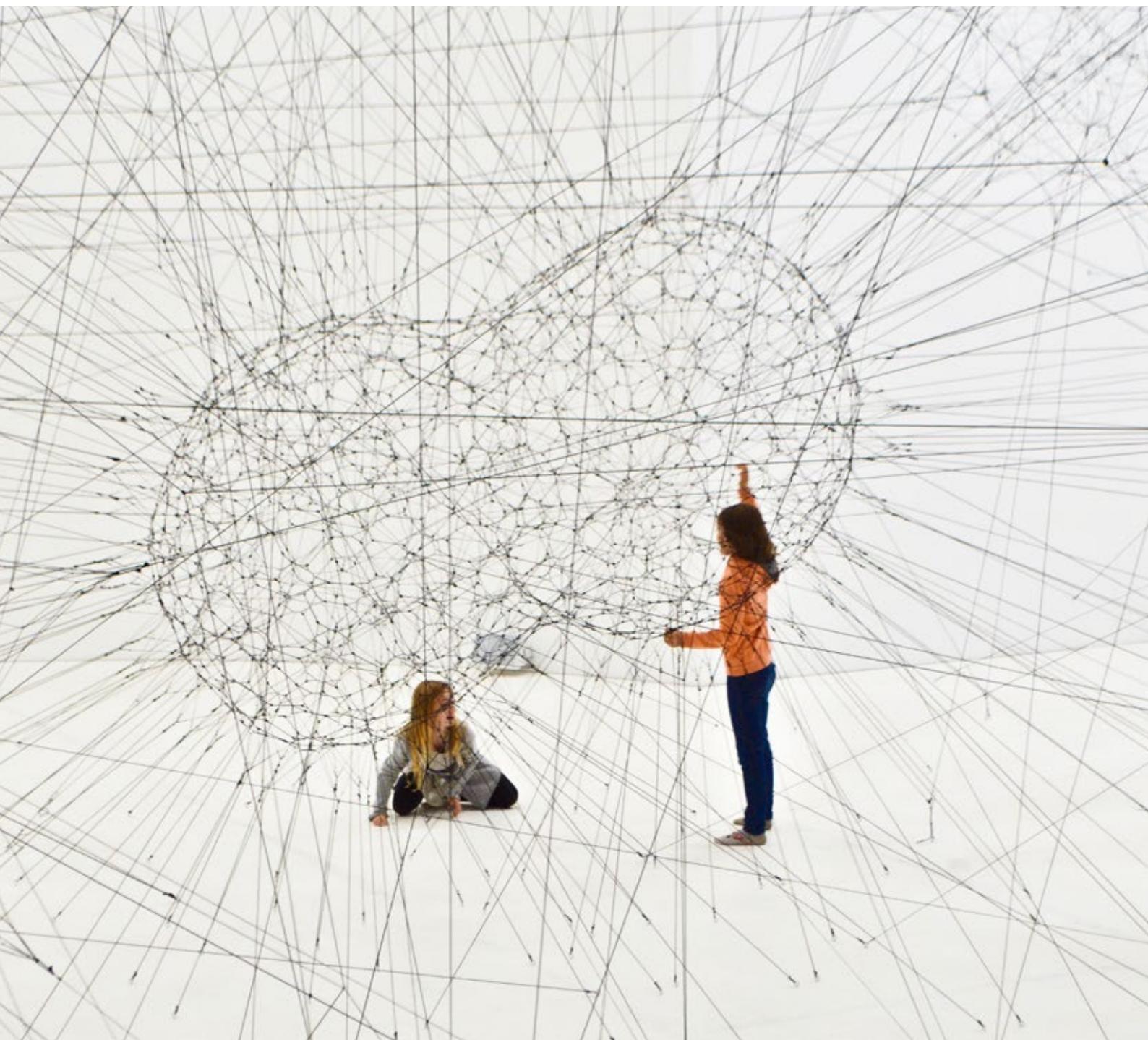
5.7

## Share your learnings

- How have you documented and shared the progress and case studies from your project with peers and stakeholders?

- Do members of the public/ end users of the project have the ability to raise concerns and place complaints about the project? If yes, how? If not, why?
- What channels have you established for public engagement and scrutiny throughout the duration of the project?

**Next steps:** if you have scored a 3 or less in any of the principles, this could indicate the need for additional checks and potential changes to make your project more ethical. Please explain the reason for the score and consult the outcome with your team leader/ organisational ethics board/data ethics lead to advise on the specific next steps to improve the ethical standards of your project.



# DataEthics.eu 2020

White Paper on Data Ethics in Public Procurement of AI based services and solutions.

Authors:

Gry Hasselbalch  
Birgitte Kofod Olsen  
Pernille Tranberg

Published by DataEthics.eu

[Info@dataethics.eu](mailto:Info@dataethics.eu)

CVR 38465724

Denmark

Graphic design

PAWs FABRIK

Front page photo

Clarisse Croset, Unsplash



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

ISBN 978-87-972168-0-4

## About DataEthics.eu

This white paper has been drafted by DataEthics.eu.

DataEthics is a politically independent non-profit ThinkDoTank founded in Denmark with global reach. We develop independent research, analysis, knowledge exchanges and tools, and raise awareness about data ethics to support a future with human agency at the centre.

We work to ensure the human interest in a fair world of data.

Data systems constantly evolve. From the data of the printing press to the data of the AI agent. Data will either empower or disempower human beings. We trace the data in human society, not the technology.

# DATAETHICS

See more: [www.dataethics.eu](http://www.dataethics.eu)

This white paper was supported by  
[www.luminategroup.com](http://www.luminategroup.com)

**Luminate**  
Building stronger societies

# **Introduction**

This white paper is an independent contribution in support of a general movement among European Union institutions and member states to guide and make strategic use of public procurement of AI, the ultimate goal being to reach shared objectives regarding the adoption and development of trustworthy AI in Europe.

Building on the recommendation of the EU High Level Group on AI to help the public sector foster responsible AI innovation, this white paper supports the development of data ethics and trustworthy AI components as part of the actions on public procurement proposed in the European Commission's 2020 AI white paper and data strategy.

The present white paper covers one pivotal general area of AI adoption in Europe in which a standardized general European approach is of key importance to the adoption of AI that truly respects fundamental rights and values in its governance structures, management systems, and technical, legal and social components.

For a quick read, we recommend the Executive Summary and Recommendations sections.

# Contents

<b>1. Background and Purpose</b>	9
1.1. Why We Need a Public Procurement Framework for Trustworthy AI	9
1.2. Procurement of public infrastructure	10
1.3. Harnessing the risks	10
1.4. Democratic processes at stake	10
1.5. Rights and AI opportunities are complementary	11
1.6. Integrating data ethics in AI	11
1.7. Background	12
<b>2. Scope and Concepts</b>	13
2.1. Scope	13
2.2. Defining Artificial Intelligence (AI)	13
2.3. A new strategic priority in procurement	14
2.4. Concepts and approaches	15
<b>3. Relevant Legal Frameworks</b>	16
3.1. Fundamental rights and freedoms	16
3.2. Equality and societal goals	18
3.3. Strategic procurement	19
<b>4. Data Ethics in AI Procurement</b>	21
4.1. A common framework	21
4.2. Data Ethics Principles	21
4.3. Compliance	23
4.4. Accountability	24
4.5. Technical Robustness	25
4.6. Sustainability	25
<b>5. Due Diligence in AI Procurement</b>	27
5.1. An adapted due diligence model	27
5.2. The due diligence process in AI procurement	27
<b>6. Preliminary Assessment and Screening</b>	30
6.1. General risk and impact assessments	30
6.2. Stakeholder dialogue	31
<b>7. Preliminary Screening of Potential Suppliers</b>	32
7.1. AI Design	32
7.2. Development and operations (DevOps)	33
7.3. Testing	35
<b>8. Contracting</b>	38
8.1. Exclusion Criteria	38
8.2. Selection Criteria	38
8.3. Technical specifications	41
8.4. Award criteria	43
<b>9. Contract Performance Conditions</b>	46
<b>10. Contract Implementation</b>	49
<b>11. Recommendations</b>	51
<b>Sources</b>	53

# Executive Summary

*Technologies are not neutral, neither are choices in the public procurement of AI. The AI systems we deploy today are the systems we will live with tomorrow.*

Artificial intelligence is increasingly shaping the opportunities of European citizens and transforming their relationships with governments and public authorities. Today we have the opportunity to responsibly define the way AI will be implemented in the future.

This white paper provides a detailed map for public procurers to choose **AI-based services and solutions that put data ethics, democracy and fundamental rights first**. It is an initial step towards creating a standardized framework for the questions and considerations public procurement processes should include to adopt trustworthy AI and reward the European development of it.

## **WHY - the quest for a human-centric approach to AI**

In late 2019, Ursula Von der Leyen pledged that within her first 100 days in office, she would propose legislation for a coordinated European approach on the human and ethical implications of AI. This white paper suggests integrating data ethics with the strategic approach to public procurement in the EU, making it supplementary to the green and social components that can be implemented as part of the overall legal framework on public procurement.

The social and ethical implications of the absence of a framework based on European values and norms for the adoption of AI in the public sector are already emerging. Automated systems to socially score families, automated systems with poorly-written code erroneously assigning public positions, black box big data analysis systems shared by different state institutions to track citizens etc.

If Europe is to truly innovate in terms of trustworthy AI systems, public institutions must lead the way. If we do not sufficiently address the trustworthy aspects of AI in public procurement, the cost for European societies, individuals and democracy will be high, with changing society forever and citizen's rights potentially being negatively impacted by the systems we choose to implement.

Public procurement of AI technology is not just a matter of choosing between more or less efficient technical tools. It is also a prioritization of interests and values embedded in their design. Trustworthy human-centric AI is an alternative type of innovation that can be developed and thrive in Europe, and the public sector could spur that development with data ethics and principles being hard-wired into public procurement.

## **WHAT - public procurement as leverage for trustworthy AI**

This white paper includes 'data ethics' as a horizontal theme that cuts across the components of trustworthy AI. Data ethics is the responsible and sustainable use of data. It is about doing the right thing for people and society. Data processes should be designed as sustainable solutions benefitting the interests of individual human beings first and foremost. Data ethics is about efforts to create transparency and foreseeability in regard to the social and ethical implications of data processing, and it is about real accountability in governance and management structures. Its goal is actively developing privacy-by-design and privacy-enhancing products and infrastructures and stressing the need always to handle someone

else's personal information in the same way as you wish your own data, or your children's data, were handled.

## **WHERE - digitized public sector institutions**

Strategic, guided public procurement processes could be implemented in public sectors such as:

- Justice and law enforcement
- E-government
- Government to Business (G2B)
- E-democracy
- Education
- Healthcare treatments and services
- Social security services
- Employment.

The process should involve all actors throughout the public procurement process from management, subject-matter experts, designers of systems, data scientists and engineers to civil servants, policy officials and governmental representatives.

## **HOW - a risk-based and systematic approach to public procurement**

Trustworthy AI is possible to achieve by establishing an AI procurement framework that includes data ethics components and applying them within the context of existing legal obligations as well as demands for accountability, technical robustness and sustainability.

This white paper suggests a risk-based approach in public procurement that is aligned with both formalized and applied due diligence processes. It recommends a due diligence process consisting of five phases:

### **1. Preliminary risk assessment**

Addressing any adverse impact on human beings or groups of people, their rights and freedoms, on democratic institutions and processes, and on society and the environment.

### **2. Preliminary screening of potential suppliers**

This involves screening the market for potential suppliers that possess the necessary skills, competences and organizational structures to fulfil requirements for data ethics components in AI services and solutions.

Data ethical requirements relating to AI should be considered, defined and implemented from the very beginning of the design process. For example:

- When AI systems interact with users directly (e.g. chatbots, virtual assistants) or indirectly (e.g. automated decision-making), they must reveal that they are not human
- AI systems must be traceable, explainable and include stakeholders
- AI systems must avoid bias, be made according to universal design and include procedures for reviews
- Technical robustness should be documented, as should explainability, fair communication and audits.

### 3. **Contracting**

General **exclusion criteria** should apply when assessing economic operators who have submitted tenders to provide AI-based services and solutions, including past participation in criminal organizations, corruption, fraud, child labour and human trafficking.

**Selection criteria** should cover relevant specialist technical competences and diverse, multidisciplinary teams that understand the interdependent disciplines that AI covers. In some circumstances, location within EU/EEA should be prioritized. Also, tenderers must guarantee their sub-suppliers comply with the same data ethics standards.

All the **technical specifications** for procurement of AI-based services or solutions should include requirements regarding the methodologies and processes foreseen in the development of the AI-based system or solution.

**Award criteria** Tenders should be assessed according to a set of economic and quality criteria and a best price-to-quality ratio. The quality criteria should reflect the technical specifications regarding applied standards and management systems for information security, data ethics, environmental aspects, privacy, universal design, etc.

4. **Contract performance conditions** To reach the overall goal of sustainability and respect for fundamental rights, and the specific goal of data ethics in AI-based services and solutions, the contracting authority should include clauses in the contract performance conditions on these issues, along with possible sanctions and documentation requirements.

5. **Contract implementation** A governance structure that includes top management in decision-making processes should support the project structure and identify roles and responsibilities in relation to all phases and levels of the AI project. The supplier should meet the requirements set out in public contracts under five headings: data ethics, legal compliance, accountability, technical robustness and sustainability. To do so, it should set up an organization with insight and overview of the contractual obligations and corresponding work processes.

Further, this white paper recommends (see Recommendations);

- An EU directive on public procurement of AI-based services and solutions for the public sector
- A guiding document on public procurement of AI-based services and solutions for the public sector
- Inclusion of data ethics as a strategic policy priority in the EU Public Procurement Strategy
- A training toolkit on data ethics in procurement of trustworthy AI-based services and solutions
- A handbook on data ethics in procurement of AI-based services and solutions
- An online help desk for public and private sector tenderers.

# 1. Background and Purpose

## 1.1. Why We Need a Public Procurement Framework for Trustworthy AI

AI-systems are increasingly becoming part of the socio-technical infrastructures of European societies. Based on complex data collection and analytics, they are shaping the opportunities of European citizens and transforming their relationships with governments and public authorities.

- Yet, we do not have standardized public procurement strategies with ethics and social impact assessments available for their adoption.

We are at an early stage of AI adoption in European societies, and therefore we have the opportunity to shape its direction in a responsible manner. In this context, national governments, public authorities and EU institutions are crucial when paving the way towards trustworthy AI consolidation and making strategic use of public procurement.

Today, European innovation is far from thriving in relation to trustworthy AI. Computer science and engineering students are not trained in the social and ethical implications of AI, most developers are not aware of it, scientists often work in disciplinary silos, businesses struggle to implement and comply with legal requirements without really being creative and innovative with trustworthy AI components. Perhaps most importantly, we do not demand trustworthy AI during public procurement processes.

We are still at an early experimental stage of AI's adoption in society, one in which the technologies and methods required to guarantee trustworthy AI remain scarce and widely debated. This white paper is the first step towards creating a standardized framework for the questions and considerations that public procurement processes should include to adopt trustworthy AI and simultaneously reward European development of it. An EU-wide approach is therefore essential to the creation of a level playing field across the EU for public procurement processes, thereby avoiding market fragmentation.

- Public procurers do not have the awareness, tools and frameworks needed to help them responsibly choose and adopt AI systems.

With digitalization, the public sector is moving towards greater efficiency. Governments and public authorities in Europe are increasingly using AI to streamline and rationalize services and interactions with citizens. AI programs are already in place to assess the risk of violence among adolescents, detect tax evasion, assess student learning patterns, profile unemployed people, find 'irregularities' in citizens' data to detect, for example, social benefit fraud, recognize motion patterns with intelligent video surveillance, and to automate the processing of traffic offences.<sup>1</sup>

In ordinary times, ethical choices are difficult. In moments of crisis and emergency, they are even harder. Urgent, rapid decisions are needed to employ the opportunities AI offers to help solve the big problems we are facing. But without proper guidance and methods for assessing social and ethical implications, the cost to European societies and democracy will be high. European societies will transform, and people's rights will be affected by the systems we choose to implement.

---

<sup>1</sup> Examples from the report "Automating Society Taking Stock of Automated Decision-Making in the EU" (2019), AlgorithmWatch in cooperation with Bertelsmann Stiftung. See more examples here: [https://algorithmwatch.org/wp-content/uploads/2019/01/Automating\\_Society\\_Report\\_2019.pdf](https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf)

- The AI systems we deploy today are the systems we will live with tomorrow.

This white paper taps into a general movement in EU member states to guide and make strategic use of the public procurement of AI to reach shared objectives regarding the general adoption of AI in Europe. It covers one pivotal, overarching area of AI use in Europe, one in which a standardized approach across Europe is key to the implementation of AI that truly respects and safeguards fundamental rights and values in its technical, legal and social components. It will also serve as a guideline for regions around the globe inspired by the European approach to AI.

## **1.2. Procurement of public infrastructure**

Public infrastructure like roads, streets and train tracks points us in specific directions. Public infrastructure enhanced by AI systems also directs our movements, informing or making decisions that shape our opportunities and limits in public and private life. Decisions made by public authorities will increasingly be extended by AI systems with, for example, decisions about individuals' entitlement to welfare services or assessments of educators' and children's performance in the school system. Thus, AI systems will impact people's lives not only in the present, but also in the future.

- Ensuring the stability of roads and train tracks is an infrastructural responsibility, as are the social and ethical implications of AI decision-making systems.

## **1.3. Harnessing risks**

The social and ethical implications of the absence of a guiding framework for the adoption of AI in the public sector based on European values and norms are already emerging in various ways. For example, in Denmark the suggestion to create an automated, data-driven risk assessment system to identify child neglect among vulnerable families, including the assignment of social credit scores to all families, has caused heavy criticism from human rights experts. Nevertheless, this model was later introduced as part of the so-called 'ghetto plan', in which special measures (e.g. harsher punishments for crimes) would be applied in designated 'ghetto' areas in Denmark. In the Netherlands, a group of civil rights initiatives have sparked legal proceedings against the use of a big data analysis system shared by different state institutions to detect citizens who have unlawfully collected public funds. In Spain, an automated system was designed to evaluate how to manage the mobility of teachers as part of educational reform. However, the program's code was so badly written that more and more teachers found themselves assigned to destinations they didn't state in their preferences,<sup>2</sup> which caused general doubt among the population about the reform.

## **1.4. Democratic processes at stake**

Fundamental relationships between the public institutions, governments and citizens are transforming rapidly. Democracy is built on carefully crafted relationships between citizens and their governments that are embedded in the very fabric of society. As the digital data linking of these relationships, the design and adoption of AI is of crucial importance to the shape of European democracy and society.

---

<sup>2</sup> AlgorithmWatch, 2019.

Recent examples of the democratic downsides of digitalization around the globe - from voter manipulation to mass surveillance - compel European societies to revisit fundamental rights and values to ensure that AI is woven into critical infrastructure responsibly and with respect for the continent's ethical frameworks.

- Technologies are not neutral; neither are choices in the public procurement of AI.

The choice to publicly procure one AI program over another is never a mere decision between more or less efficient technical tools. It is also a prioritization of interests and values embedded in their design.

### **1.5. Rights and AI opportunities are complementary**

To innovate is to meet new requirements in new ways, and trustworthy AI is the most innovative investment for Europe.

- Citizens are requesting it. Democracy is requiring it.

Building AI without infringing on fundamental rights and democracy while limiting its risks are not mutually exclusive objectives. Trustworthy, human-centric AI is an alternative type of innovation that can be developed and thrive in Europe. But it does not exist in a bubble. Trustworthy AI needs a technical design and organisational business culture in which people are, for example, in control of and have insight into their data, where processes and systems can be explained and audited, programmers and designers are aware of the social implications of their work, and goals are set for social good.

### **1.6. Integrating data ethics in AI**

AI is data in a specific form. It is developed from data, evolve on the basis of data and act on data in digital environments. Not only is data the building block of AI, it is also a core component of socio-technical infrastructure in contemporary societies.

Data ethics is the responsible and sustainable use of data. It is about doing the right thing for people and society. Data processes should be designed as sustainable solutions benefitting, first and foremost, the interests of individual human beings.

Data ethics is a step beyond mere compliance with personal data protection laws. All data processing therefore respects - as a minimum - the requirements set out in the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights, and those in secondary EU legislation, including the General Data Protection Regulation (GDPR).

Data ethics is also about our understanding and interpretation of legislation. It refers and adheres to the principles and values upon which fundamental rights and personal data protection laws are based. As such, data ethics is about efforts to create transparency and foreseeability with regard to the social and ethical implications of data processing. And it is about real accountability in data management. It pursues a goal of actively developing privacy-by-design and privacy-enhancing products and infrastructures and stresses the constant need to handle someone else's personal information in the same way as you wish your own data, or your children's data, were handled.

## **1.7. Background**

With the EU strategy on AI adopted in 2018, a general governance process was initiated in the EU member states.<sup>3</sup> Informed by the High-Level Expert Group on AI's ethics guidelines and policy and investment recommendations for Trustworthy AI, an approach to AI-based on European values and fundamental rights has now been put forward as Europe's answer to AI adoption in the global arena.

In 2019, the new president of the European Commission, Ursula Von der Leyen, pledged that, within her first 100 days in office, she would propose legislation for a coordinated European approach on the human and ethical implications of AI<sup>4</sup> and accordingly in February 2020, a white paper on AI and a European data strategy was published.<sup>5</sup>

This white paper on public procurement of AI takes its point of departure in the recommendation of the European High-Level Expert Group on AI to support the European public sector in responsible public procurement of AI: 'The public sector can make strategic use of public procurement to foster responsible innovation, as well as steering it towards tackling societal challenges and the development of trustworthy AI solutions.' (Policy & Investment Recommendations, June 2019).

Building on this recommendation, the white paper aims to support the development of the components of **trustworthy AI with an emphasis on data ethics** in Europe to contribute to the following actions proposed in the European Commission's AI white paper and data strategy:

1. To adopt AI programs that will support public procurement of AI systems, and help to transform public procurement processes themselves (white paper)
2. To elaborate a data initiative for public procurement data covering both the EU dimension and the national ones complemented by a procurement data governance framework (data strategy)
3. To facilitate the development of common European standards and requirements for public procurement of data processing services (data strategy)
4. To facilitate the set-up of a cloud services marketplace for EU users from the private and public sector facilitated by the Commission, which will put users (in particular the public sector and SMEs) in the position to select cloud processing, software and platform service offerings that comply with a number of requirements in areas like data protection, security, data portability, energy efficiency and market practice (data strategy).

---

<sup>3</sup> For the EU strategy, see: <https://ec.europa.eu/digital-single-market/en/artificial-intelligence>

<sup>4</sup> [A Union that strives for more My agenda for Europe, part 3 A Europe Fit for the Digital Age, 2019.](#)

<sup>5</sup> White Paper on Artificial Intelligence - A European Approach to Excellence and Trust. Brussels, 19.2.2020, COM(2020) 65. A European strategy for data. Brussels, 19.2.2020 COM(2020) 66 final. A European strategy for data. Brussels, 19.2.2020 COM(2020) 66 final.

## 2. Scope and Concepts

### 2.1. Scope

The public sector plays an important role in ensuring that AI-based services and solutions are developed, deployed and maintained across the public sector, aligned and compliant with EU standards and regulation, and in pursuit of an ethical application of AI to the benefit of the European citizens.

The aim of this whitepaper is therefore to facilitate a data ethics framework for the procurement of trustworthy AI in the public sector in the EU member states.

By paving the way for AI procurement that is in-line with data ethics, the public sector will have tremendous leverage in relation to innovation and services within AI business and research communities that respect fundamental rights and values and pursue trustworthy solutions.

Thus, the target group of this whitepaper is EU politicians, legislative bodies and institutions, and other relevant stakeholders who are part of the process of developing strategies and regulation within the area of public procurement.

### 2.2. Defining Artificial Intelligence (AI)

To this day, there has been no commonly accepted, broadly shared definition of the term Artificial Intelligence (AI). Coined in the 1950s, the term has had many meanings and applications, depending on context and application. Recently, AI has re-emerged in public discourse and policymaking as a more generic term to describe a data intensive development of digitalisation.

This white paper builds on the definition of AI presented by the EU High-Level Expert Group on AI:

**“** Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal (...) High level Expert Group on AI, Ethics Guidelines, p. 36.

In addition, we use the term ‘trustworthy AI’ as defined by the EU High-Level Group on AI to refer to AI solutions that embed legal and ethical principles and requirements in their design and implementation.<sup>6</sup> There are three layers to this concept, meaning that it should be:

- Lawful, complying with all applicable laws and regulations;
- Ethical, ensuring adherence to ethical principles and values; and
- Robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.<sup>7</sup>

<sup>6</sup> EU High Level Expert Group on AI, Ethics Guidelines, p. 5

<sup>7</sup> In this white paper, when using the term robust in specific, we are only referring to technical robustness. We consider social robustness a horizontal component together with data ethics.

Extending these definitions of AI, we would like to emphasize AI as complex socio-technical data systems and processes with different levels of human involvement. We argue that it is this same human involvement in AI data design, organisation, adoption and use that can be influenced and shaped with public procurement. We therefore have included ‘data ethics’ as a horizontal theme that cuts across the components of trustworthy AI.

DataEthics.eu has developed a set of data ethics principles<sup>8</sup> that we use as a general framework to assess the data ethics of trustworthy AI:

- The human being at the centre
- Individual data control
- Transparency and explainability
- Accountability
- Equality.

### **2.3. A new strategic priority in procurement**

To facilitate and support this process, this white paper suggests a guiding framework for integrating data ethics considerations and impact assessments into public procurement of AI-based products and services. Its ultimate purpose is to contribute to a wider uptake of *trustworthy AI* by European member states.

The goal is to ensure that future public procurement of AI is aligned with EU Directives on Public Procurement<sup>9</sup> and the EU Commission’s public procurement strategy and supplementary guidelines.<sup>10</sup> Therefore, it addresses two strategic priorities:

- Strategic public procurement: adding ‘data ethics’ as a new strategic criterion to existing social, environmental and innovative criteria.
- A guiding framework: how to best integrate ethical and legal considerations into public procurement procedures.

We envisage that strategic and guided public procurement of trustworthy AI could be implemented in public administration such as

- Justice and law enforcement
- E-government
- Government to Business (G2B)
- E-democracy
- Education
- Healthcare treatments and services
- Social security services
- Employment.

---

<sup>8</sup> <https://dataethics.eu/data-ethics-principles/>

<sup>9</sup> EU Directive 2014/24/EU on Public Procurement, EU Directive 2014/25/EU on Procurement by entities operating in the water, energy, transport and postal services sectors and EU Directive 2014/23/EU on the award of concession contracts.

<sup>10</sup> EU Public Procurement Strategy (2017), Green Public Procurement initiative (2027) and Innovation Guidelines (2017).

This should involve all actors throughout the public procurement process, from system designers, data scientists and civil servants to policy officials and governmental representatives.

## **2.4. Concepts and approaches**

This whitepaper addresses public procurement of AI-based services and solutions with a fundamental rights approach. Hence, it builds upon the legal framework within EU law on fundamental rights and freedoms as defined in the EU Charter of Fundamental Rights. As it is focused on data ethics, this whitepaper emphasizes the inherent right to dignity, the right to privacy and protection of personal data, non-discrimination and equal opportunities. The use of AI by public administration may impact other fundamental rights such as the freedom of movement, freedom of expression and information, the right to vote, to a fair trial and effective remedies. The whitepaper should thus be seen as only addressing a non-exhaustive list of fundamental rights.

Legal regulation operationalising the fundamental rights to respect for privacy and data protection is found in the EU General Data Protection Regulation (GDPR) and the EU Regulation on the free flow of non-personal data (FDDR). The whitepaper adds a supplementary soft law layer to the legal obligations consisting of the EU High-Level Group on AI's ethics guidelines and policy, and investment recommendations for trustworthy AI<sup>11</sup> and the European Commission Whitepaper on Artificial Intelligence.<sup>12</sup> The applied legal framework is described in part 3.1.

Future application of AI-based services and solutions to provide access to public goods and services affects the principles of equal treatment and non-discrimination. As these rights are pivotal for ensuring the overall purpose of the EU, fundamental rights and EU directives pertaining to the area of equality and social goods are included in the white paper. See more in part 3.2.

The suggested framework is modelled upon the existing EU legal framework and strategy for public procurement, cf. part 3.3. to make it feasible to integrate the framework in existing EU structures.

The underlying concept for integrating risk and impact assessments is based on generic models for due diligence as recommended by the EU in relation to non-financial reporting,<sup>13</sup> in the UN Guiding Principles on Business and Human Rights<sup>14</sup> and the OECD Guidelines for Multinational Enterprises.<sup>15</sup> In setting up the framework for this white paper, applied procurement practices at large corporations have been inspirational. The aim of aligning the suggested framework with existing standards and business practices is to make it recognizable for providers of AI-based services and solutions to the public sector.

---

<sup>11</sup> High Level Expert Group on AI - Ethics Guidelines for Trustworthy AI, (April 2019) and HLEG AI - policy and investment recommendation for trustworthy AI (June 2019).

<sup>12</sup> WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust Brussels, 19.2.2020, COM (2020) 65 final.

<sup>13</sup> Directive 2014/95/EU as regards disclosure of non-financial and diversity information by certain large undertakings and groups of 22 October 2014.

<sup>14</sup> UN Guiding Principles on Business and Human Rights, 2011.

<sup>15</sup> The guidelines were adopted in 2011 and supplemented with OECD Due Diligence Guidance for Responsible Business Conduct in 2018.

### 3. Relevant Legal Frameworks

For this white paper, we apply EU legal frameworks establishing fundamental principles and values in democratic societies. They encompass fundamental rights and freedoms, especially the right to respect for dignity and privacy, data protection, the protection of vulnerable groups, and social cohesion as they are imperative to thriving and growth in the digital age.

To ensure that the proposed recommendations are feasible, the legal framework and policies for public procurement in the EU have been used as a guide when establishing a proposed structure that is ambitious while also recognizable and realistic.

#### 3.1. Fundamental rights and freedoms

##### Dignity

The applied concept of giving priority to the interests of human beings in relation to data ethics is embedded in article 1 of the EU Charter of Fundamental Rights and Freedoms, stating that human dignity is inviolable and must be respected and protected.<sup>16</sup>

This principle, as it relates to AI, has been applied in accordance with the Council of Europe's Convention on Human Rights and Biomedicine.<sup>17</sup> Article 2 of the Convention stipulates the primacy of the human being and states that 'the interests and welfare of the human being shall prevail over the sole interest of society or science'.

This primacy principle reflects the need to respect human beings both as individuals and as members of the human species, and operationalizes the basic right to dignity of the human being as established in the UN's Universal Declaration of Human Rights, the European Convention on Human Rights and Freedoms<sup>18</sup> and the EU Charter. The preamble of the biomedicine convention thus mentions that the misuse of biology and medicine may lead to acts endangering human dignity. It seems fair to anticipate a similar risk in relation to the misuse of AI-based services used, for example, in automated decision making and individual profiling in the public sector, including individualized risk profiles for tax purposes, job performance in relation professional opportunities, or behaviour or fitness in relation to healthcare treatments.

##### Privacy and data protection

The EU Charter obliges EU member states to ensure the right to respect for the private life of every individual in article 7 and the right to personal data protection in article 8. The Charter introduced the right to data protection as a separate fundamental right and thereby strengthened the protection of personal data deriving from article 8 of the European Convention on Human Rights and Freedoms.

Secondary legislation in the EU imposes legal obligations on member states to protect

<sup>16</sup> EU Charter for Fundamental Rights, 2000/C 364/01.

<sup>17</sup> Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine, CETS 164, 4 April 1997.

<sup>18</sup> Council of Europe Convention on Human Rights and Fundamental Freedoms, CETS No. 194, 3 September 1953.

personal data. As a result, the GDPR applies to all data processing activities in the public and private sectors, and will be the standard for the development, application and maintenance of services and solutions. With its risk-based approach to data protection and specific principles for processing (article 5), it defines the guarantees for safe and legitimate use of AI, explains the roles and responsibilities for data controllers and data processors (article 24 og 28), including in relation to appropriate security measures reflecting the risk picture (article 32).

An important obligation is put on the data controller in relation to ensuring the implementation of data protection by design and by default, both when determining the means of personal data processing and at the time of the processing itself (article 25).

Moreover, the explicit right of data subjects not to be subject to a decision based solely on automatic processing and profiling if the decision has legal effects concerning the person or similarly significantly affects the person, is of paramount importance in relation to AI. As a result, it introduces a new standard for risk analysis.

The EU directive covering data protection in law enforcement<sup>19</sup> safeguards citizens' fundamental right to data protection whenever personal data is used by criminal law enforcement authorities for law enforcement purposes. It thus plays an important role in the manner in which law enforcement authorities obtain AI-based services and solutions to support decision-making processes as part of the prevention, investigation, detection or prosecution of criminal offences or for the implementation of criminal penalties. In these situations, the directive ensures that protection standards for the personal data of victims, witnesses, and suspects are duly protected and will facilitate cross-border cooperation in the fight against crime and terrorism.

The EU regulation on the free flow of non-personal data (FFDR) is highly relevant in relation to AI as AI-based services and solutions draw on large scale data sets.<sup>20</sup> The FFDR regulates mixed data sets of personal and non-personal data. The latter is defined as:

- Data which initially did not relate to an identified or identifiable natural person. E.g. data on maintenance needs for industrial machines or trading data on the financial sector.
- Data which was originally personal data but was later anonymized, e.g. aggregated datasets used for big data analytics or anonymized data used for reporting purposes.

With regards to the second group of non-personal data, the FFDR underscores that anonymization is different than pseudonymization. Pseudonymized data is still considered personal data, as the information can be linked to an individual by combining it with additional data.

The assessment of whether data has been properly anonymized is not a one-size-fits-all operation as it depends on the circumstances of the specific case. As the Commission mentions in the guidelines, if non-personal data can be attributed to a person in any way (and

<sup>19</sup> EU Directive on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, 2016/680 of the European parliament and of the Council, 27 April 2016.

<sup>20</sup> EU Regulation on a framework for the free flow of non-personal data in the European Union 2018/1807 of the European parliament and of the Council, 14 November 2018.

hence, make the person identifiable), the data must be considered personal and within the scope of the GDPR.<sup>21</sup>

In a mixed dataset, if the non-personal and personal data are inextricably linked, the GDPR applies to the whole set of mixed data, regardless of the amount of personal data contained in the dataset. In other words, if separating the set of personal data from the non-personal one is impossible, economically inefficient (e.g. if by separating the datasets, its value is compromised), or not technically feasible, the set of data shall be considered inextricably linked and the data protection rights and obligations stemming from the GDPR apply to the whole dataset.

### **3.2. Equality and societal goals**

#### **Equal treatment and non-discrimination**

Several examples show that AI can potentially further stereotypes and bias in relation to ethnicity, skin colour, gender and age, highlighting the need for data sources and elements to be thoroughly analysed in AI development processes. Furthermore, it necessitates comprehensive maintenance and auditing of the output and outcome of applied AI-based systems and solutions.

In that light, the foundational principles of equal treatment and non-discrimination stated in the EU Treaty and the EU Charter are of outmost importance in relation to AI procurement.

In the EU Charter, a full section is dedicated to fundamental rights ensuring equality before the law (article 20), non-discrimination (article 21), respect for diversity with regard to culture, religion and linguistics (article 22), equality between men and women (article 23), rights of the child (article 24) and of elderly persons (article 25), and the integration of those with disabilities in order for them to benefit from measures designed to ensure their independence, social and occupational integration, and participation in the life of the community (article 26).

To a large extent, the fundamental rights on equal treatment and non-discrimination have been operationalized through EU directives on discrimination at work on grounds of religion or belief, disability, age or sexual orientation,<sup>22</sup> discrimination on grounds of race and ethnic origin,<sup>23</sup> on equal treatment of men and women in employment and occupation,<sup>24</sup> and on access to and supply of goods and services.<sup>25</sup> The standards and requirements established in the directives are to be interpreted and applied in a digital age and are highly relevant when designing and developing sustainable AI-based systems and solutions.

---

<sup>21</sup> Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union, Communication from the commission to the European parliament and the council, COM(2019) 250 final, Brussels, 29.5.2019.

<sup>22</sup> Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation

<sup>23</sup> Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin

<sup>24</sup> Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation

<sup>25</sup> Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services

In the recently-adopted EU Directive on accessibility requirements for products and services, new obligations have been imposed on member states to ensure a more inclusive society and facilitate independence for people with disabilities.<sup>26</sup>

Also, the concept of universal design that improves the usability of products, environments, programs and services for all people, including those with disabilities, through design was introduced by the UN Convention on the Rights of Persons with Disabilities.<sup>27</sup>

### **Common societal goals**

EU legislation on public procurement also contains wording relating to the overall goals of society. Thus, the EU directive on public procurement is framed by guiding principles on equal treatment and non-discrimination. Authorities in member states tasked with procuring works, supplies and services are thus expected to comply with these principles, as they derive from the free movement of goods, freedom of establishment and the freedom to provide services.<sup>28</sup>

The directive furthermore reflects a social purpose in its preamble by listing one of its purposes as helping procurers make better use of public procurement in support of common societal goals.<sup>29</sup>

### **3.3. Strategic procurement**

#### **High quality and efficiency**

The legal framework for public procurement is outlined in the provisions of the EU Treaty on the Functioning of the European Union and in the EU Procurement Directives, adopted in 2014.<sup>30</sup> The 2014 Directives were designed to create an open, competitive and well-regulated procurement market. Not only does this guarantee EU citizens high-quality public services, but it also increases efficiency in spending and improves the transparency of procurement processes.

In 2017, as part of the strategy to make the Single Market stronger, the EU Commission published Making Public Procurement Work in and for Europe, a public procurement strategy that defines an overall policy framework and identifies priorities to improve procurement within the EU.<sup>31</sup>

The document establishes six strategic priorities:

- Ensuring wider uptake of strategic public procurement
- Professionalizing public buyers
- Improving access to procurement markets
- Increasing transparency, integrity and better data.

---

<sup>26</sup> EU Directive (EU) 2019/882 on the accessibility requirements for products and services.

<sup>27</sup> Signed on 30 March 2007 and in force from 3 May 2008.

<sup>28</sup> EU Directive 2014/24/EU of 26 February 2014, preambular provision 1

<sup>29</sup> EU Directive 2014/24/EU of 26 February 2014, preambular provision 2.

<sup>30</sup> Directive 2014/24/EU on public procurement, Directive 2014/25/EU on procurement by entities operating in the water, energy, transport and postal services sectors, Directive 2014/23/EU on the award of concession contracts

<sup>31</sup> EU Public Procurement Strategy (2017).

- Boosting the digital transformation of procurement
- Cooperating to procure together.

### **Innovative, green and social procurement**

The priority area aiming at achieving a wider uptake of strategic public procurement is particularly relevant for this white paper. This strategic approach is defined in the document as encompassing innovative, green, and social components.

In this perspective, public procurement should be a tool used to boost innovation in the EU. To move public authorities in this direction, the Commission has published a notice titled ‘Guidance on Innovation Procurement’, establishing the main elements of a relevant policy framework.<sup>32</sup>

Besides innovation, strategic procurement should aim to increase the uptake of green and environmentally friendly solutions. Green Public Procurement (GPP) is a voluntary tool that EU member states and public administrations can decide to adopt.<sup>33</sup> GPP is defined as the process whereby public authorities seek to procure goods, services and works with reduced environmental impact throughout their life cycle when compared to standard goods, services and works with the same primary function.

The initiatives implemented by the Commission to facilitate green purchasing encompass:

- Identification of GPP criteria for specific products and services
- A Buying Green Handbook
- A website dedicated to GPP.

With regard to the social aspect of public procurement, European public authorities are urged to buy in a socially responsible manner and purchase ‘fair trade’ products and services that promote sustainable development. To facilitate social procurement purchasing, the Commission has published a specific guide titled ‘Buying social’. Not only does it define a strategy for buying social, but it also sets forth the requirements and specifics for contracting with suppliers and service providers.<sup>34</sup>

This white paper suggests integrating data ethics into the strategic approach to public procurement in the EU as a supplementary aspect to the communicated innovative, green and social components, implemented as part of the overall legal framework on public procurement.

---

<sup>32</sup> Innovation Guidelines (2017).

<sup>33</sup> Green Public Procurement initiative (2027) and Communication (COM (2008) 400) “Public procurement for a better environment”.

<sup>34</sup> Buying Social A Guide to Taking Account of Social Considerations in Public Procurement, European Commission, October 2010.

## 4. Data Ethics in AI Procurement

### 4.1. A common framework

To improve the procurement processes of data ethical AI-based services and solutions for the public sector in EU member states, a common framework should be applied, ensuring that an appropriate due diligence process is performed and that relevant requirements are defined and applied.

Due diligence and requirement specifications are instrumental for assessing the quality and security of the solutions offered, and also for revealing the risks inherent to solutions pertaining to fundamental rights, especially privacy, data protection, and non-discrimination rights.

Moreover, a generic set of requirements and specifications may be used when designing and qualifying development processes, i.e. where public procurement is aimed at purchasing new AI-based solutions and systems that are not yet available on the market.

This white paper suggests an AI procurement framework that is based on a set of data ethics principles that are to be applied within the context of legal obligations and demands for accountability, technical robustness and sustainability.



Figure 1 An AI procurement framework

### 4.2. Data Ethics Principles

This white paper recommends a trustworthy AI framework for public procurement that horizontally incorporates a set of data ethics principles reflecting individual agency and individual data control as core elements for maintaining human dignity in a digitized age.

The rationale is that every individual should be in control of his or her personal data and be empowered by access to data. A person's self-determination should be prioritized in all data processes and each person should be actively involved in the recording of his/her data. Individuals should have primary control over the usage of their data, the context in which said data is processed and how it is activated.<sup>35</sup>

<sup>35</sup> See DataEthics.eu data ethics principles.

These data ethics principles also aim to ensure an effective European data sharing infrastructure that makes it possible to explore and use data for common societal goals and to boost sustainable growth. However, this objective cannot be achieved without ethical and social risks if the technical, organizational and market conditions are not developed in Europe and supported within a European institutional framework.

The applied principles are aligned with the requirements for trustworthy AI suggested in the High-Level Expert Group on AI's ethics guidelines, the EU White Paper on AI, and the considerations laid out in section 3 above. Hence, they cover:

### **Human primacy**

Human interests always prevail over institutional and commercial interests. People are not computer processes or pieces of software, but unique, with empathy, self-determination, unpredictability, intuition and creativity. AI-based services and solutions should support human autonomy and enable them to prosper in democratic, sustainable and robust societies.

### **Universal design**

Everyone should benefit from products, environments, programs and services based on AI, regardless of their abilities. AI design should not exclude or hinder people from thriving in digitized societies on equal terms but allow for all people to use AI-based products and services, irrespective of their age, gender, abilities or characteristics. Universal design should be used to empower people with disabilities in particular and enable equitable access and active participation.

### **Transparency**

AI-based data processing activities and automated decisions, including profiling, must make sense for individuals. They should not undermine human autonomy but provide individuals with the knowledge and tools required to comprehend and interact with AI-based services and solutions, and to support them in making informed choices and decisions. The purpose and outcomes of data processing must therefore be clearly understood by each individual in terms of comprehending risks and any social, ethical and societal consequences.

### **Traceability**

AI-based services and solutions should be developed and applied in a way that makes it possible to verify and document data sources, data sets and categories, the algorithm used and its variables, and the processes that lead to automated decision making. Therefore, errors and adverse impact on individuals and society can be identified and prevented.

### **Explainability**

The technical processes and related human decisions in AI-based services and solutions should always be explainable. If the AI system or solution produces results or decisions based on correlations, a requirement for causality should be provided for by ensuring human intervention in the decision-making process. The demand for explainability is of vital importance both when the AI-based system or solution has significant impact on people's lives and when it influences and shapes organizational decision-making processes, e.g. in relation

to procurement, design choices and when defining the purpose of deploying the system/solution.

### **Fair communication**

AI-based services and solutions should identify themselves as non-human by informing human users that they are interacting with an AI system/solution. Such information should include notice of the level of accuracy, the risk of adverse impact, and limitations to AI practitioners and end-users. Individuals should also be informed of how they can opt out of AI interactions and demand that human interaction is provided.

#### **Data sharing spaces based on personal data control**

AI is data-hungry. It evolves, learns, predicts and decides on data. As such, many ethical implications of AI regard data ethics by their very nature. One overarching ethical challenge is the use of people's data for AI innovation. To ensure a new paradigm in which data asymmetries between institutions, businesses and individuals are limited, a European data infrastructure must be based on individual empowerment and personal data control not only with imposed requirements, but also naturally evolving with the help of motivation and reward systems.

A current movement in technology and business development is addressing this type of individually-controlled data sharing infrastructure with the creation of personal data management systems and services. These are interoperable services that enable individuals to share data and either donate or activate their data for personal benefits (such as personalized finance or medicine) while being in control of the use of their data. Examples can be found in the [Mydata.org](#) network of entrepreneurs, activists, academics, corporations, public agencies and developers, and in technological initiatives such as [Solid](#).

### **4.3. Compliance**

AI-based services and solutions of relevance from a public procurement perspective must comply with the legal obligations of all applicable laws and regulations. This is the component of trustworthy AI that the AI HLEG has referred to as 'lawful AI'. This includes, but is not limited to, legislation on privacy and data protection, especially the GDPR, and supplementary national legislation as well as sector-specific regulations.

Legal regulation on state surveillance and profiling as part of crime investigation and prevention, such as control and monitoring schemes in relation to tax and social benefit fraud, is also highly relevant to AI procurement.

Access to the labour market or to health and social services illustrates areas where AI-based services and solutions are introduced to facilitate job matching, healthcare visits or education by screening and profiling users or applicants. The procurement and application of such AI-based services and solutions calls for thorough compliance-oriented analyses of anti-discrimination laws and public administration law.

Similarly, legislation that furthers freedom of movement, freedom of association and assembly, and participation in politics and cultural life must be complied with when procuring AI-based systems and solutions for smart city programmes, among other things.

The application of conformity assessments for high-risk AI uses as stated in the EU's AI white paper (February 2020) are particularly important when procuring AI services and products from outside the EU, in which case it is pivotal to ensure that European legal and ethical standards are still met.

Other areas of legal concern in relation to AI cover the right to vote and to run for election, the right to good administration and access to public documents. The right to file complaints on decisions made by public authorities or processes failing to comply with public administration law is another example that must be taken into consideration when procuring AI-based services and solutions.

Linked to this area is the need for compliance analyses that unveil potential negative impact on the fundamental principles of democracy, justice and the rule of law, e.g. through the application of AI-based services and solutions undermining democratic processes, the plurality of values regarding individual life choices and the basic principle of equality.

A risk-based approach to the procurement of AI is crucial. Areas in which the application of AI carries particular high risks and potential for adverse consequences requires particular attention, such as the tracking and profiling of citizens, citizen scoring, biometric recognition (e.g. facial recognition), predictive policing, and the monitoring and assessment of children.

#### **Compliance with ethical standards**

An additional 'ethical standards compliance' layer may be considered, taking internationally established ethical standards in the research and scientific fields into account. These are already formally applied in the form of ethics audits and assessments within the European funding programme Horizon2020. A few relevant examples are:

[The Helsinki declaration](#)

[The Oviedo Convention \(the 'bioethics convention'\)](#)

See also: [Horizon2020 ethics guidance framework](#)

#### **4.4. Accountability**

Closely linked to the demand for legal compliance is the fulfilment of accountability principles. Along with the GDPR and the introduction of a risk-based approach to data protection and the obligation to set up organisational and technical measures, accountability is an important means to achieve and document compliance.

Data protection management systems and defined governance structures, including roles and responsibilities, internal policies, procedures, guidelines, training, communication, monitoring and control mechanisms, have – in practice – become a prerequisite for ensuring the maintenance and review of compliant data-processing activities.

In relation to AI, an accountability approach ensures a reflective, reasonable and systematic use and protection of personal data by forming an integral part of all aspects of data processing, and efforts can be made to reduce the risks for the individual and to mitigate social and ethical implications. It also contributes to the procurement process by highlighting the need to

evaluate whether subcontractors have incorporated accountability measures throughout their organization.

#### **4.5. Technical Robustness**

Technical robustness in AI-based services and solutions is imperative for ensuring an effective level of information security and cyber security.

Technical robustness encompasses adequate protection of the integrity and confidentiality of personal data and provides a protective layer against unauthorized access to data; alteration, loss or destruction of data; and unauthorized disclosure of data.

Resilience to hostile attacks such as hacking, malware and spyware, espionage and destruction of infrastructure should form part of effective security systems and measures to protect AI-based services and solutions.

Other technical measures should include backup plans, e.g. for switching from an algorithm-based procedure to human interaction, and mechanisms that ensure that unintended consequences and errors are minimized. Technical measures ensuring correct, accurate decisions, predictions, and recommendations are other safeguards that will help make AI-based services and solutions robust as they help prevent harm to human beings and society.

Similarly, measures that ensure reliable and reproducible results contribute to properly functioning systems and enhance transparency and explainability.

#### **4.6. Sustainability**

Preventing harm and ensuring fairness are core aspects of sustainable personal data processing. They address both environmental and social concerns with the goal of benefitting all human beings, including future generations. The two concerns may be intertwined, as is the case in the UN Sustainable Development Goals, by tackling a pressing social need in an environmentally friendly way, for example.

Environmental well-being could also be pursued by integrating the examination of resource and energy consumption in the development and deployment phase of the AI-based system or solution.

##### **Environmental Impact of Data Processing**

Algorithmic decision-making has immense impact on the environment due to the extensive data processing and analysis it requires. As an example, the consumption of energy to run a Bitcoin network is equivalent to the energy needs of Cambridge University for 360 years. <https://www.cbeci.org/comparisons>

A goal of social well-being would encourage greater focus on the risks generated by simulating sociality, relationships and emotional attachment when interacting with the chatbots, robots or virtual reality avatars of an AI-based system or solution.<sup>36</sup> Thereby, it potentially impacts

<sup>36</sup> Examples are listed in the EU High Level Group on AI's, ethics guidelines p. 19.

the physical and mental well-being of humans and may alter socio-cultural practices. Other societal risks are related to its impact on democratic institutions and AI-based decision-making processes.

Assessments of the sustainability of data processing in AI-based services and solutions should thus encompass, as a minimum, the identification of social and environmental risks and adequate measures to minimize intentional or unintentional harm in the short, medium and long term. As society changes, the AI model must be seen as a dynamic tool that requires continuous control with applied variables. Variables that appeared relevant and essential at the time of development may lead to error synchronization, because of changes in behavioural patterns, perceptions and other contextual factors, and produce harmful results. The assessment of sustainability should thus be integrated in regular and systematic application and monitoring processes.

#### **Data ethical trade**

During public procurement processes, public institutions should be able to purchase goods which make a special contribution to sustainable development that is ethical in terms of its data. The AI white paper issued by the EU suggests the establishment of a 'voluntary labelling scheme' for low-risk AI applications that would allow users to easily recognize AI-enabled products and services that are in compliance with certain objective, standardized EU-wide benchmarks, going beyond the normally applicable legal obligations. The criteria of such labelling schemes could be incorporated as considerations in, for example tender, specifications (similar to how 'ethical trade' is handled in the ['Buying Social' guide](#)).

## 5. Due Diligence in AI Procurement

### 5.1. An adapted due diligence model

As part of a strengthened focus on good governance and sustainability across sectors, concepts and models for due diligence processes have emerged that effectively help organizations to gain insight into their potential negative impact on the environment, human beings, society and the economy. They can be grouped into three main strands:

One recommendation for due diligence processes is embedded in the EU directive regarding the disclosure of non-financial and diversity information and its supplementary guidelines.<sup>37</sup>. It reflects an acknowledged and widely-used method developed by the UN in relation to the responsibility of corporations to respect human rights,<sup>38</sup> and a method presented by the OECD with regard to overall sustainability in the business sector, including human rights, labour rights, the environment, consumer rights, bribery and taxation. The OECD guidelines have led to standards being set for due diligence processes in general and in specific sectors, including institutional investors.<sup>39</sup> Most recently, a tool was launched to guide the planning of due diligence processes in connection with the UN Sustainable Development Goals.<sup>40</sup>

In parallel, large corporations have established formalized procurement processes that include third-party risk management programmes. Such programmes are designed to diligently identify and handle risks as part of the procuring, contracting and implementation phases.

As a third strand, public sector procurement has been aligned to the principles and requirements established in EU directives on procurement.

This whitepaper is inspired by all three strands and uses an adapted model as a basis for the following sections, sharing its logic and concepts with methodologies developed in international fora and in the EU. This model may easily be aligned with the formalized procurement structures of contracting public sector entities and integrated with existing practical structures pertaining to tenderers in the business sector.

### 5.2. The due diligence process in AI procurement

The due diligence process in relation to the public procurement of AI-based systems and solutions improves transparency in the planned AI model and its impact. It helps ensure that all relevant risks and impacts in relation to data ethics are identified and handled, and that the development, deployment and maintenance of the AI model accommodates the needs and requirements of a trustworthy AI-based system or solution. As such, it must span the whole lifecycle of the AI-based system or solution.

---

<sup>37</sup> Directive 2014/95/EU as regards disclosure of non-financial and diversity information by certain large undertakings and groups of 22 October 2014 and Communication from the Commission - Guidelines on non-financial reporting (methodology for reporting non-financial information) 2017/C 215/01.

<sup>38</sup> UN Guiding Principles on Business and Human Rights, 2011.

<sup>39</sup> OECD Due Diligence Guidance for Responsible Business Conduct, 2018.

<sup>40</sup> See the SDG Compass - The guide for business action on the SDGs.

The due diligence process consists of six phases:

1. Preliminary risk assessment
2. Preliminary screening of potential suppliers
3. Contracting
4. Contract performance conditions
5. Contract implementation
6. Transition

The process thus encompasses an initial phase in which a preliminary risk assessment is completed with regard to the scope/matter the AI-based system or solution is designed to handle (e.g. automated or assisted decision-making for triage in healthcare services).

These assessments should be performed prior to screening potential suppliers or tenderers and before defining requirements and specifications. In this way, the due diligence process informs and facilitates a sharper focus on the demands and specific requirements of a planned AI-based system or solution. The results from this phase may also be used in the process of drafting the tender.

In phase 2, potential suppliers are evaluated, and their performance capabilities are assessed against the expected level of skills, competences, methodologies, governance and procedures for development, training and validating data, testing environments, maintenance and audit functions. This may include information from relevant stakeholders, e.g. academia, groups of impacted citizens, or local committees.

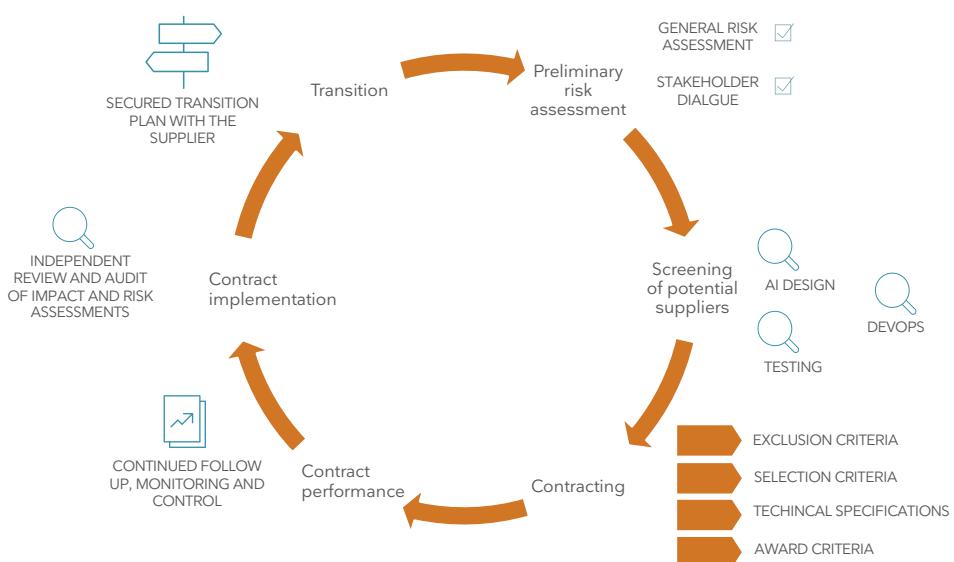


Figure 2 Due diligence process in AI procurement

In the contracting phase, a set of criteria should be adopted that guides the selection of relevant tenderers, allowing for the exclusion of others. Technical specifications help assess, in depth, whether the tenderers can accommodate the integration of data ethics in their development, deployment and maintenance operations. Well-defined award criteria will assist in the final selection of a supplier.

The fourth phase of the due diligence process focuses on the conditions for contract performance. Contract clauses on data ethics should include relevant management and control systems or guidelines to be followed by the contracting party to ensure - as a minimum - that requirements in relation to data ethics, legal compliance, accountability, technical robustness and sustainability are met.

During the fifth phase, the contract should be implemented with a focus on processes and tools that ensure that the contracting party can identify, evaluate and report on progress, including on any risks or incidents, to the public contractor.

In the final phase, the contract is concluded, and the AI-based system or solution securely transmitted to the public contractor. This phase is not covered by the white paper.

The content and requirements set up within the first five phases are described in the following sections.

## 6. Preliminary Assessment and Screening

### 6.1. General risk and impact assessments

This white paper suggests that risk assessments be carried out as part of the first phase of the AI procurement due diligence process. The aim is to provide an initial assessment of the risk and potential impact of a planned AI project.

Impact and risk assessments should uncover potential negative outcomes, risks and threats linked to the design, development, deployment and maintenance of a given AI-based system or solution. It should thus cover the expected results from an AI-based decision-making process or interaction.

Preliminary assessments could encompass five sub-categories:

- **Data ethics impact assessment**

A data ethics assessment should address the potential adverse impact on individuals or groups of individuals and democratic societies, and the technical and organizational measures put in place to reduce and monitor such impact. Part 5 and 6 of this white paper contain components that could be transformed into a data ethics impact assessment tool.

- **Legal compliance assessment**

Data protection impact assessments (DPIA) should be performed in accordance with the GDPR to unveil risks to fundamental rights and freedoms. If the planned AI-based service and solution is to be deployed in areas governed by national legislation, e.g. education, the workplace, employment, healthcare or social services, patient rights etc., a compliance assessment of relevant legislation should be performed.

- **Accountability impact assessment**

An accountability assessment is necessary if the expected risks and impacts of the AI-based service or solution are linked to governance structures and management systems. The assessment should focus on organizational measures and cover adequacy and effectiveness in relation to designing, developing, deploying and maintaining the planned AI project.

- **Security risk assessment**

An assessment of security measures and the security level should focus on risks and negative impact stemming from a lack of technical robustness, including the integrity and confidentiality of the data to be processed by the AI-based service or solution, accessibility to data, cyber security measures, business continuity plans and incident response plans.

- **Social and environmental impact assessment**

If a planned AI-based service or solution is expected to impact democratic institutions and processes, or have excessive impact on the environment, an assessment should be performed to identify the extent, severity and likelihood of such impact.

Whether the assessments should be performed independently or in combination depends on the reach of the AI-based service or solution itself.<sup>41</sup> If it will impact multiple arenas of

---

<sup>41</sup> See an example of extending the GDPR DPIA to cover good governance and good process in Swee Leng Harris,

human activity or behaviour, and entails extensive impact on society and the environment, the assessment should include all of the categories above. The same is true if sensitive information about citizens is collected and processed in IT systems and networks in which security measures do not provide appropriate and effective protection. If the effect on citizens is insignificant or minor, a preliminary assessment could concentrate on governance and accountability measures in combination with a data ethics impact assessment.

### **Assessment Tools and Methodologies**

There is a need to develop standardized risk assessment methodologies and tools that cover the areas mentioned in this section. Several assessment tools for AI and algorithmic impacts have already been developed by various organizations. For example:

'Algorithmic Impact Assessments' by AI NOW  
<https://ainowinstitute.org/aiareport2018.pdf>

'Artificial Intelligence Impact Assessment' by ECP  
<https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf>

'Algorithmic Impact Assessment' by the Canadian government  
<https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html>

'The Assessment List' by the EU HLEG on AI  
<https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/2>

In addition, guidelines and methodologies exist for assessing the impact of research and science which may be used as a basis for developing such tools. For example, the European Commission has published a comprehensive guide for the assessment of research projects in relation to ethics and data protection:  
[\(2018\)](https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-data-protection_en.pdf)

## **6.2. Stakeholder dialogue**

As part of the due diligence process, consultations with stakeholders and rights holders should be performed in order to expose potential negative impact on or harm to people or groups of people due to their age, gender, ethnicity, sexual orientation, disability or other characteristics. Stakeholder dialogue is especially relevant where gaps in information exist.

The inclusion of other stakeholders such as suppliers, industrial federations, authorities, academia and non-governmental organizations (e.g. those for consumers, citizens or patients) in active dialogue prior to the procurement process would provide useful insight into risk prediction and risk handling, supply chain mechanisms, and marked maturity in relation to planned AI-based services and solutions.

## 7. Preliminary Screening of Potential Suppliers

The following should be considered as part of the second phase of the due diligence process. The purpose is to screen the market for potential suppliers that possess the necessary skills and competences and will implement the proper procedures relating to data ethics in AI products and services.

### 7.1. AI Design

The development of an AI system should focus on delivering a product that is both technically robust and ethical. To ensure this, those involved in designing the AI-based system or solution should follow methodology composed of specific key activities.

Data ethics requirements for AI should be considered, defined and implemented from the very start of its development, meaning in the early phases of designing and establishing its architecture. Implementing ethical considerations from day one will make it possible to develop an efficient yet cost-effective solution.

The preliminary screening of suppliers of AI-based services or solutions should focus on the implementation of organizational and technical measures that provide insight into their general methodologies, governance and structures.

The fundamental concept in this phase is transparency, which should include the following requirements and criteria:

#### *Self-identification of an AI system*

When AI systems interact with users directly (e.g. chatbots, virtual assistants) or indirectly (e.g. automated decision-making), they must reveal that they are not human. This ensures that individuals can make informed decisions as to whether they wish to use the provided service or product. Moreover, users must always be given the possibility to request the intervention of a human being.

#### *Traceability*

While designing AI solutions, developers must present a clear record of the systems, algorithm training methods, models, etc. used, as well as the decisions made. It should therefore be possible to keep track of the decisions made by the AI algorithm itself. Traceability facilitates the identification of mistakes and makes their correction easier.

#### *Explainability*

Explainability refers to the decisions made by the AI system. The choices made should be comprehensible to individuals interacting with the AI. However, it should also be possible for programmers and coders to explain the combination and weighing of data during the training of the algorithm and in the final application of the data. Hence, an explainable AI can describe the choices made and the outcomes of those decisions. Explainability - as with traceability - should be kept in mind from the beginning of the AI system design process. This is possible, for instance, by selecting the simplest and most interpretable model for the AI system, transparency of the business model selected, etc.

#### *Inclusion of stakeholders*

Formalized processes for dialogue with individuals, groups of individuals or their representatives whose life and fundamental rights are affected by the AI can provide insight about adverse impact and risks. Furthermore, inclusion of other relevant stakeholders can inform procurers of the market's readiness to engage in the sustainable development of AI-

based services and solutions.

RECOMMENDATION: Preliminary screening - AI design

Requirement	Criteria
<b>Self-identification</b>	<ul style="list-style-type: none"><li>▪ Notification of non-human interaction</li><li>▪ Documentation for easy-to-understand and accessible notifications</li></ul>
<b>Traceability</b>	<ul style="list-style-type: none"><li>▪ Adoption of procedures for documenting the development and maintenance of AI-based services and solutions</li><li>▪ Demonstration of records</li></ul>
<b>Explainability</b>	<ul style="list-style-type: none"><li>▪ Explanation of logic and mathematics behind algorithms, choices and deselections, outputs and outcomes</li></ul>
<b>Inclusion of stakeholders</b>	<ul style="list-style-type: none"><li>▪ Rights holders impacted by planned AI</li><li>▪ Researchers and market representatives</li></ul>

## 7.2. Development and operations (DevOps)

The procurement of AI-based solutions and systems for the public sector should ensure that data ethics have been integrated in both development and planned operations with potential suppliers. Hence, a preliminary screening should focus on business decision makers, and the involvement of relevant experts and data scientists.

On the technical side, the tasks of programmers and coders should be transparent, and they should account for their analyses and be able to explain choices and decisions relating to data sources, variables in the algorithm, or operation and maintenance flows in the programming of the system. This should also include a reflection on and explanation of the consequences of leaving out certain sources, variables or data flows. Ethical considerations should be integrated in each phase of development, including when writing, analysing and reviewing the code that forms the basis of that AI solution.

This presupposes that the DevOps team is trained in data protection legislation and information security requirements. Moreover, team members should possess knowledge about data ethics as it relates to their jobs, including that they learn how to identify a code's vulnerabilities and risks as they relate to the fundamental rights and freedoms of individuals, and have access to tools or methods to eliminate or minimize them.

Data ethics within coding should be ensured by requesting that suppliers fulfil the following criteria:

*Avoid bias*

Trustworthy AI systems should not have a discriminatory effect, directly or indirectly. Identified bias should be considered as soon as it is discovered and removed if deemed discriminatory. Potential suppliers should demonstrate how procedures and checklists support this process.

To expedite the detection of bias and the risk of discrimination, it can be beneficial to ensure diversity in DevOps teams. This may also be achieved by including potentially-impacted stakeholders in relevant phases. Furthermore, it is important to appoint a person who is responsible for overseeing the process and who will examine the purpose of each development phase, its limits, and check that the set requirements are met by the AI.

*Universal design*

In designing AI-based services and solutions, the concept of universal design should be considered and applied to the widest extent possible. Enabling accessibility and participation through suitable user interfaces that allows minority groups such as people with disabilities, cultural or linguistic minorities, children or elderly people to benefit from AI-based solutions would thus further inclusion and diversity in society.

To educate developers about specific needs and preferences among minority groups, dialogue with stakeholders, representatives from relevant minority groups and NGOs, academia, civil servants and other practitioners, should form part of the development process. Similarly, stakeholders could be included in the testing of developed systems and solutions.

*Programming and code review*

To achieve high ethical standards in an AI, a code review phase is essential. This consists of a manual revision of the code performed by fellow programmers with the aim of detecting and correcting mistakes. Reviewing code can ensure that there is no leakage or excessive processing of personal data. The reviewer should have the goal of identifying where data is stored and ensure that an adequate level of protection is in place.

RECOMMENDATION: Preliminary screening – development and operations

Requirement	Criteria
<b>Avoid bias</b>	<ul style="list-style-type: none"><li>▪ Training programme for DevOps teams</li><li>▪ Methodologies for identifying bias with a discriminatory effect</li><li>▪ Procedures for detecting bias and corrective measures</li></ul>
<b>Universal design</b>	<ul style="list-style-type: none"><li>▪ Inclusion of stakeholders</li><li>▪ Testing accessibility</li></ul>
<b>Programming and code review</b>	<ul style="list-style-type: none"><li>▪ Procedures for internal and external reviews</li></ul>

### **7.3. Testing**

In this phase, the AI system or solution needs to be assessed to ensure that it meets the stated requirements and all checks and balances are in place before it's launched. Moreover, programmers need to test the AI to detect any technical vulnerabilities.

#### *Technical robustness*

In the testing phase, security measures need to be assessed to evaluate if the correct level of protection to individuals has been achieved. Therefore, the resilience of the AI system to potential attacks and data leaks need to be tested.. This should be done through the several types of scans, including but not limited to security, penetration, fuzz and dynamic testing.

It should be documented that the applied security measures are adequate and appropriate, and that the resilience of the AI-based service or solution has been tested and meets the necessary level of security.

#### *Traceability*

All relevant tests of the AI-based system or solution should be documented. This should include descriptions and a track record of methods used to test and validate the program and the applied algorithm.

Testing and validation should include training data and the scenario tested on such data, and the data used as part of its real-world application.

#### *Explainability*

The output of an AI-based system or solution should be understood by human experts. This may include the application of explainable artificial intelligence (xAI) methodology. Explanations should cover the correlation of data, including misfit risk. If possible, the explanation should also encompass the underlying rationale for causality between the input

data processed by machine learning and the expected decision or proposal put forward by the AI-based system or solution.

In this light, systems based on ‘black box’ concepts should not be seen as fulfilling data ethics requirements.

#### *Fair communication*

The supplier should make use of open source methodologies, programming, coding and testing tools. Testing procedures and results should be made available and communicated in a transparent manner.

#### *Maintenance*

Monitoring mechanisms should form an integral part of maintenance procedures for AI-based services and solutions regarding security, accuracy, transparency, reliability, correlation and causality. A monitoring routine could consist of three steps: 1) testing, assessing and evaluating, 2) updating and 3) auditing to be performed annually or biannually depending on the risk picture related to the processing activities and purpose of the AI-based system or solution.

#### *Audit*

Apart from an audit of the quality and technical robustness of the applied AI system, the potential supplier should demonstrate that output and impact assessments are audited in relation to sustainability, covering the relative environmental footprint, social inequality, exclusion or marginalization of certain groups, and other ethical implications.

The audit should be external and could be based on:

- The auditors’ observations and investigations
- Self-assessment by the tenderers
- Documentation, provided by the tenderer, of established measures, the inclusion of stakeholders and impact assessments performed.

RECOMMENDATION: Preliminary screening - testing

Requirement	Criteria
<b>Technical Robustness</b>	<ul style="list-style-type: none"> <li>▪ Documentation of appropriate security measures</li> <li>▪ Resilience tests</li> </ul>
<b>Traceability</b>	<ul style="list-style-type: none"> <li>▪ Documentation of testing and validating the algorithm</li> <li>▪ Validation of training data</li> <li>▪ Procedure for testing and validation</li> </ul>
<b>Explainability</b>	<ul style="list-style-type: none"> <li>▪ Explanation of correlation</li> <li>▪ Explanation of the underlying rationale for causality</li> <li>▪ Applied standards for xAI</li> </ul>
<b>Fair communication</b>	<ul style="list-style-type: none"> <li>▪ Open source</li> <li>▪ Transparent testing procedures</li> </ul>
<b>Maintenance</b>	<p>Testing and evaluation of:</p> <ul style="list-style-type: none"> <li>▪ Security</li> <li>▪ Transparency</li> <li>▪ Quality (accuracy, reliability)</li> <li>▪ Causality</li> </ul>
<b>Audit</b>	<p>Checking for:</p> <ul style="list-style-type: none"> <li>▪ Discriminatory bias and appropriate mitigating measures</li> <li>▪ Stakeholder inclusion</li> <li>▪ Programming and code review</li> <li>▪ Fair communication</li> <li>▪ Maintenance management</li> <li>▪ Environmental impact</li> <li>▪ Social impact</li> <li>▪ Ethical impact</li> </ul>

## 8. Contracting

The existing public procurement framework contains concepts, procedures and requirements that may be used as leverage to ensure the integration of data ethics throughout the procurement process.

The following sections illustrate how data ethics principles may be integrated in all phases of the procurement process, adding supplementary elements to already-existing procedures.

### 8.1. Exclusion Criteria

General exclusion criteria should be applied when assessing economic operators submitting tenders for AI-based services and solutions, including participation in criminal organizations, corruption, fraud, and child labour and human trafficking violations.

Supplementary data ethics exclusion criteria should encompass violations established by national, regional or international courts, and EU or Council of Europe institutions, such as:

- Profiling or assigning a score to citizens or consumers in violation of fundamental rights
- Large-scale identification and tracking of individuals without a specific purpose
- Development and deployment of lethal autonomous weapon systems.

RECOMMENDATION: Procurement process - exclusion criteria

Exclusion criteria	Content
<b>Unlawful exploitation of personal data</b>	<ul style="list-style-type: none"><li>▪ Profiling or social scoring</li><li>▪ Identification and tracking</li><li>▪ Lethal autonomous weapon systems</li></ul>
<b>Hostile use of personal data</b>	<ul style="list-style-type: none"><li>▪ Abusing the fundamental right to dignity</li></ul>

### 8.2. Selection Criteria

To be selected, the tenderer should meet a set of requirements linked to their technical and professional abilities, including human resources and technical capacities and experience. The selection criteria may also include requirements regarding technical facilities, the use of sub-suppliers and verification mechanisms applied to ensure the quality, compliance and security of the sub-suppliers, and references to previously-fulfilled contracts.

#### Diverse, multidisciplinary teams

A vital component for ensuring that data ethics considerations become an integral part of the procurement process is to assess whether the tenderer has or will have a diverse, multidisciplinary team that understands the interdependent disciplines that AI covers.

A variety of skills and experience is needed to cover all aspects of the AI life cycle. As part of preliminary research and in the design phase, the tenderer should show that it has access to specialized knowledge on how to identify relevant data sources and data elements, while also ensuring compliance with data protection regulation.

When initiating the design of the AI system, the tenderer should make use of team members with expertise in AI systems, data analytics and engineering, model development (e.g. deep learning), model/information visualisation and experience with relevant and reliable AI methods such as text analysis, sentiment analysis, content categorization, process analysis, or augmented government. Also, an agile process to ensure compliance and risk testing during the development, coding and testing phases is imperative. This will typically require specific knowledge and skills relating to fundamental rights, data protection legislation, information security and cyber security.

Deploying and implementing AI-systems in practice requires technical skills (e.g. user interfaces), as well as organizational skills within governance, change management, monitoring and audit programmes, training and awareness-raising, just to name a few.

### **Technical facilities**

Companies who submit a tender should meet specific requirements regarding access to their relevant specialist/technical facilities. The development environment should include state-of-the art facilities, with methodologies, software, programs, devices and processes for the development team, and relevant tools for integrating data ethics throughout the development process. Similar facilities and tools should be available during testing processes.

The tenderer should also document the installation and maintenance of adequate and robust physical and IT security measures within their facilities and on applied technical equipment, including hardware, software, servers, cloud solutions and networks, and during data training (ML), and the transmission and storage of data. An effective access management system should be in place along with mechanisms set up to track unauthorized access to data or the disclosure, alteration or disappearance of data.

In certain circumstances, tenderers who will locate their technical facilities within the EU/EEA should be given priority. This could be relevant in relation to the processing of sensitive data in combination with critical infrastructure, e.g. structures serving a vital societal interest in relation to energy, health, security, food, transportation and the economy.<sup>42</sup>

### **Sub-suppliers**

If the tenderer plans to use sub-suppliers, the tenderer should specify which part of the contract will be performed by each sub-supplier, and specify their geographical location within or outside the EU/EEA, technical facilities, diversity and skills, efficiency, experience and reliability in relation to developing and/or applying AI/ML methodologies and tools.

The tenderer should guarantee that sub-suppliers have not been involved in any abuses of fundamental rights and freedoms or complicit in violations covered by the exclusion criteria.

---

<sup>42</sup> EU Directive COUNCIL DIRECTIVE 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection.

This presupposes that risk assessments have been carried out by the tenderer in relation to data ethics, fundamental rights, accountability, security and sustainability (social and environment) and can be documented.

Tenderers should specify how sub-suppliers will be administered during the processes covered by the procurement contract and what monitoring/auditing mechanisms will be set up in order to monitor and review the quality of the work performed by sub-suppliers. Such mechanisms may include self-assessments, observations, testing and external audits.

### **References**

Any tenderer bidding on the development and delivery of an AI-based system or solution should demonstrate a sufficient level of experience with the application of AI/ML in a similar or comparable sector or thematic area. That level of experience should be demonstrated by suitable references from past contracts.

General requirements for public procurement, such as the establishment of any conflicts of interest, may be supplemented with considerations regarding ethical impacts or risks linked to fulfilled contracts if they are part of thorough, independent evaluations of projects or external audits.

RECOMMENDATION: Procurement process - selection criteria

Selection criteria	Content
<b>A diverse, multidisciplinary team</b>	<ul style="list-style-type: none"> <li>▪ AI system and data engineering</li> <li>▪ Data science</li> <li>▪ Coding</li> <li>▪ Testing</li> <li>▪ Law</li> <li>▪ IT and cyber security</li> <li>▪ Change management</li> <li>▪ Monitoring mechanisms</li> <li>▪ Process facilitation</li> <li>▪ Training and awareness raising</li> </ul>
<b>Technical facilities</b>	<ul style="list-style-type: none"> <li>▪ Development environment</li> <li>▪ Test environment</li> <li>▪ Physical and IT security measures</li> <li>▪ Location within or outside the EU/EEA</li> </ul>
<b>Sub-suppliers</b>	<ul style="list-style-type: none"> <li>▪ Location within or outside the EU/EEA</li> <li>▪ Risk assessments</li> <li>▪ Control mechanisms, including audits</li> </ul>
<b>References</b>	<ul style="list-style-type: none"> <li>▪ Past AI contracts</li> <li>▪ Independent evaluations and audits of projects</li> </ul>

### 8.3. Technical specifications

The technical specifications of all tenders for AI-based services or solutions should include requirements regarding the methodologies and processes planned for the development of that AI-based system or solution. This should address all aspects of the contract and cover any stage during the design, development, training and deployment of the service or system.

In that light, public procurement processes should be in place to consider and assess if the tenderer of AI-based services and solutions meets the following supplementary requirements.

#### Applied standards

If standardized management systems are applied to quality, information security, cyber security, cloud solutions, privacy/PII, data ethics, the environment, social responsibility and AI, the tenderer should identify and describe applied standards, e.g. ISO standards, OECD guidelines, IEEE recommended practices or their equivalents.

The tenderer should demonstrate that certifications or official recognition/approvals have been obtained or granted from international, regional or national standardization or accreditation institutions or their equivalents.

### **Fundamental rights compliance**

Due to the inherent high risk that AI-based systems and solutions pose in causing direct or indirect discrimination, unequal treatment, inequality, exclusion and stigmatization, the tenderer should - as a minimum requirement - demonstrate compliance with EU anti-discrimination legislation.

If the AI-based services or solutions entail negative impact on other fundamental rights, especially the right to privacy, free movement and freedom of expression, the tenderer must demonstrate compliance with such rights.

Also, the tenderer should demonstrate how, once identified, risks to fundamental rights will be handled, i.e. eliminated, minimized or prevented in the design phase and when developing, training and deploying the AI-based service or solution, and through which methodologies, tools and measures (e.g. by design strategies and design patterns that ensure privacy and data protection).

### **Universal design**

The tenderer should explain and demonstrate how accessibility for persons with disabilities or how 'design for all' are integrated in the design, development and testing of the proposed AI-based service or solution. This includes explaining the choices and outcomes of the service/solution that could impact its accessibility, and any potential mitigating measures.

### **Data ethics management**

Equivalent to the requirement to demonstrate applied standards, a requirement to demonstrate data ethics management should form part of the technical specifications, as it may reveal that the tenderer has (or lacks) the technical capability to actually fulfil the contract. Such a management system should encompass a data ethics policy, the governance structure, and a description of the roles and responsibilities for each function or task covered by the policy, due diligence processes, monitoring and complaint mechanisms.

### **Environmental management**

In public procurement contracts for AI-based services and solutions, the technical specs should include a requirement regarding the application of environmental management measures or frameworks.

Thus, the tenderer should demonstrate it has measures or structures in place that integrate procedures and processes for staff training, and monitoring, summarizing, and reporting on specialized environmental information relevant to the AI-based service/solution. The application of environmental criteria could encompass, for example, the Ecolabel, which is relevant in relation to hardware, office/lab furniture and equipment, the carbon footprint of the AI development project and the resulting solution's maintenance.

RECOMMENDATION: Procurement process - technical specifications

Technical specifications	Content
<b>Applied standards</b>	<ul style="list-style-type: none"> <li>▪ ISO standards</li> <li>▪ OECD</li> <li>▪ IEEE</li> <li>▪ EU HLEG on AI</li> <li>▪ National or regional sector-specific guidelines</li> </ul>
<b>Fundamental rights compliance</b>	<p>Demonstration of compliance:</p> <ul style="list-style-type: none"> <li>▪ Anti-discrimination legislation</li> <li>▪ Gender equality</li> <li>▪ Privacy and data protection</li> <li>▪ Freedom of movement</li> <li>▪ Risk handling</li> </ul>
<b>Universal design</b>	<ul style="list-style-type: none"> <li>▪ Demonstration of means to ensure accessibility for disabled persons</li> </ul>
<b>Data ethics management</b>	<p>Demonstration of:</p> <ul style="list-style-type: none"> <li>▪ A data ethics policy</li> <li>▪ Governance structure</li> <li>▪ Monitoring mechanisms</li> <li>▪ A complaint mechanism</li> </ul>
<b>Environmental management</b>	<p>Demonstration of:</p> <ul style="list-style-type: none"> <li>▪ Environmental measures or guidelines</li> <li>▪ Environmental criteria</li> </ul>

#### 8.4. Award criteria

Tenders should be assessed according to a set of economic and quality criteria set up by the contracting authorities. They should be transparent and ensure effective, fair competition among the tenderers.

The contracting authorities should also ensure that it is possible to effectively verify the information provided by tenderers and facilitate the process of identifying the most economically advantageous tender, i.e. the tender that demonstrates the best price-performance ratio.

By awarding the contract, the contracting authorities should ensure that objective criteria are applied in compliance with the principles of transparency, non-discrimination and equal treatment, and that they are aimed towards an overall goal of sustainability.

## **Quality**

The quality criteria of a request for tenders should reflect the technical specifications regarding applied standards and management systems for categories such as information security, data ethics, the environment, privacy and universal design.

Quality criteria could also include the ability of the AI model to adapt to change, e.g. through monitoring the model management and key figures.

Also, they could include the tenderer's technical merits within AI/ML, and the aesthetic, functional and innovative characteristics of the proposed service or solution.

Other quality criteria relevant to the subject matter of the contract could include the handling or integration of data ethics, environmental and social aspects, as revealed in the impact assessments performed as part of the due diligence process (see section 5).

It may be necessary to operate with a non-exhaustive list of potential quality criteria to ensure a comparative assessment of the performance offered by each tender.

As with general awarding processes in public procurement, the assessment of the best price-performance ratio should define the economic and qualitative criteria linked to the subject matter of the contract and reflect the specific characteristics of the AI-based service or solution.

## **Organization**

Designing, developing, testing and deploying AI-based services and solutions typically requires staff members who possess specific qualifications and experience. In public procurement contracts in particular, the tenderers' merits and experience within accessibility and universal design, data ethics, environmental and social impact assessments, fundamental rights compliance analyses and risk assessments, including privacy, data protection, gender equality, non-discrimination and equal treatment, are pivotal to the quality of the outputs and outcomes generated by the service or system.

Public-sector AI-based services and solutions to automatize or support decision making in cases or processes involving citizens will affect the lives of men, women and those with other gender identities. To avoid discriminatory gender bias in designing such services or solutions and choosing and coding the variables in the algorithm, it may be relevant to add the demonstration of gender equality in design and development teams to the list of award criteria.

Similarly, any potential adverse impact on citizens due to their age, race, ethnicity, religion or belief, disability, or sexual orientation should be eliminated or mitigated when developing AI-based services and solutions through the tenderers' demonstration of diversity in all DevOps teams.

### **Best price-performance ratio**

In order to assess the best price-performance ratio, the applied qualitative criteria should be accompanied by a cost criterion, e.g. the price or cost defined on the basis of a cost-effectiveness approach (such as life-cycle cost analysis).

RECOMMENDATION: Procurement process - award criteria

Award criteria	Content
<b>Quality</b>	<ul style="list-style-type: none"><li>▪ Technical merits within AI/ML</li><li>▪ Experience with model management systems</li><li>▪ Functionality, aesthetics and innovative characteristics</li><li>▪ Integration of data ethics, environmental and social aspects</li></ul>
<b>Organization</b>	<ul style="list-style-type: none"><li>▪ Data ethics, environmental and social impact assessments skills</li><li>▪ Fundamental rights compliance and risk assessment skills</li><li>▪ Non-discrimination and equal treatment skills</li><li>▪ Gender balance in teams</li><li>▪ Diversity in teams</li></ul>
<b>Best price-performance ratio</b>	<ul style="list-style-type: none"><li>▪ Price or cost of each qualitative criterion</li></ul>

## 9. Contract Performance Conditions

To fulfil the overall goal of sustainability and respect for fundamental rights, and the specific goal of data ethics in AI-based services and solutions, the contracting authority should include clauses in the contract performance conditions on these issues, together with possible penalties and documentation requirements.

### Contract clause

The data ethics contract clause should include a requirement regarding relevant management systems or structures with the contracting party to ensure - as a minimum - data ethics, legal compliance, accountability, technical robustness and sustainability.

Specific requirements regarding relevant information security standards and specific legal frameworks may be added.

As part of fulfilling such requirements, the contract party should be under an obligation to ensure appropriate technical and organizational measures with the contracting party.

For example:

- Due diligence processes
- Compliance analyses
- Data ethics policies and procedures
- Data protection
- Information and cybersecurity
- Procedures for integrating data ethics principles and data protection requirements throughout the project's life cycle, i.e. in the design phase and during development, testing, launching, deployment and maintenance
- Any relevant requirements regarding the team's diversity and gender equality could be included.

If the supplier uses sub-suppliers to carry out specific tasks under the public contract, an obligation should be imposed on the supplier to provide sufficient guarantees to establish appropriate technical and organizational measures with the sub-suppliers that correspond to those imposed on the supplier.

### Penalties and termination

Contract performance conditions should include penalties for non-compliance. They should not be imposed immediately but seen as a process initiated by the contracting authority with the purpose of finding constructive solutions to problems and challenges in relation to data ethics, compliance, accountability, technical robustness and sustainability.

Such a process should be based on dialogue and engagement with the goal of finding agreed-upon solutions. It may include a wide range of dialogue and mediation tools and approaches and may also include other stakeholders. The process could cover initiatives that mitigate the negative impact on fundamental rights, preventive mechanisms, those that improve the integration of data ethics in the system's design, other forms of reparations, or - as a last resort - termination of the contract and a claim for damages.

## **Documentation**

The supplier should demonstrate accountability across all phases and all levels of the AI-based service/solution and its development. Documentation requirements should encompass:

- Governance structure
- Management systems or frameworks
- Applied standards
- Methodologies
- Analyses and assessments
- Procedures and processes
- Risk-management initiatives
- Monitoring/auditing mechanisms
- Training and testing
- Evaluations.

Other activities or incidents related to contractual requirements on data ethics, compliance, accountability, technical robustness and sustainability should also form part of the contract.

The public authority should have access to documentation upon request and no later than 30 days after a written notification is delivered to the supplier.

## **Inspection and monitoring mechanisms**

The contract should include a clause regarding monitoring and auditing mechanisms in relation to appropriate fulfilment and due diligence of contractual requirements relating to data ethics, compliance, accountability, technical robustness and sustainability.

Inspections should be carried out on a regular and systematic basis and may be based on a self-assessment by the supplier, on-site observation by the commissioning public authority, or by an external and independent evaluation or audit.

The contract may also specify monitoring mechanisms to be applied to sub-suppliers.

RECOMMENDATION: Procurement process - contract performance conditions

Contract performance conditions	Content
<b>Contract clauses</b>	<ul style="list-style-type: none"> <li>▪ Management systems</li> <li>▪ Technical and organization measures</li> <li>▪ Data ethics policy</li> <li>▪ Sub-suppliers</li> </ul>
<b>Penalties</b>	<ul style="list-style-type: none"> <li>▪ Dialogue and engagement</li> <li>▪ Mitigation of negative impact</li> <li>▪ Preventive mechanisms</li> <li>▪ Termination</li> <li>▪ Damages</li> </ul>
<b>Documentation</b>	<ul style="list-style-type: none"> <li>▪ Life cycle documentation</li> <li>▪ Access to documentation</li> </ul>
<b>Inspection and monitoring mechanisms</b>	<ul style="list-style-type: none"> <li>▪ Regular and systematic inspections</li> <li>▪ Verification mechanism regarding sub-suppliers</li> </ul>

## 10. Contract Implementation

When implementing contracts for AI-based services and solutions, it is necessary to focus on the thematic areas covered by data ethics in public procurement as laid out in this white paper.

To meet the requirements established by a public contract under the categories of data ethics, legal compliance, accountability, technical robustness and sustainability, the supplier will have to set up an organization to monitor and provide insight into its contractual obligations and corresponding processes.

Similarly, the contracting public authority should be engaged in and have access to processes and tools that enables it to identify, evaluate and report on progress, documentation and potential risks with the supplier for not fulfilling the contract.

### Organizational measures

The effective implementation of a public contract with a supplier presupposes clearly-defined work processes within a managerial structure, such as a project management unit or a programme management office, depending on the size of the project.

A governance structure that includes top management in decision-making processes should support the AI project and identify roles and responsibilities in relation to all its phases and levels.

The resulting organization should be designed to oversee development processes, administrative processes, evaluations and inspections, and reporting schemes internally and with the public authority. It may also involve training employees in areas covered by the contract, e.g. impact assessments on data ethics and fundamental rights.

Apart from providing support to all employees and teams within the project, these organizational measures should ensure that the contract's requirements are fulfilled, and that anomalies, delays or other relevant impediments are reported to the public authority.

Depending on the project, the public authority may play an important part in the development and deployment process, e.g. by participating in steering committees or acting as the project manager/owner. In that case, all organizational measures should reflect the collaboration or partnership between the supplier and the public authority.

### Follow-up mechanisms

The public authority procuring an AI-based service or solution should have follow-up mechanisms in place to effectively monitor and check on progress and results, and the fulfilment of the contractual requirements.

Follow-up measures may include initiation meetings, regular status meetings and continuous dialogue with the supplier. Certain parts of the initial phases of the AI project may require the inclusion of stakeholders, e.g. groups of citizens, NGOs, consumer organizations, academia, and other public authorities (data protection authorities, national human rights institutions, etc.) to elaborate on potential adverse effects on minority groups.

Also, the contracting public authority may have an interest in testing and evaluating applied methodologies, model management, coding reviews, or may have insight into training data etc., or into the documentation of certification updates and other matters.

Monitoring mechanisms may also include self-assessments performed by the supplier in relation to some or all phases of the project. Inspections could also be performed by third parties or as part of external and independent audits.

As part of contract management, the contracting public authority must often approve the use of new sub-suppliers. This requires formalized processes on both sides.

RECOMMENDATION: Procurement process - contract implementation

Implementation of contracts	Content
<b>Organizational measures</b>	<ul style="list-style-type: none"><li>▪ Formal procedures</li><li>▪ Defined work processes</li><li>▪ Monitoring mechanisms</li><li>▪ Reporting structure</li><li>▪ Contract management system</li><li>▪ Training</li></ul>
<b>Follow-up</b>	<ul style="list-style-type: none"><li>▪ Start-up and status meetings</li><li>▪ Self-evaluation</li><li>▪ Continuous dialogue</li><li>▪ Model management</li><li>▪ Certification updates</li><li>▪ Evaluation and audits</li><li>▪ Approval of sub-suppliers</li></ul>

## 11. Recommendations

### **Inclusion of data ethics in the procurement process**

This white paper recommends the inclusion of data ethics requirements in the public procurement of AI-based services and solutions.

To this end, a procurement model should include a **due diligence process**. This makes it possible for the public procurer to gain qualified insight into the planned AI model, its risks and impacts. It would also provide the public procurer with important information about potential suppliers in the market, their skills, methodologies and management systems.

Due diligence helps ensure that all relevant risks and impacts relating to data ethics are identified and handled, and that the development, deployment and maintenance of AI accommodates the requirements of a trustworthy AI-based system or solution. As such, a properly diligent procurement process spans the whole lifecycle of the AI-based system or solution.

Recommendations are made in relation to five phases of the procurement process:

1. Preliminary risk and impact assessments
2. Preliminary screening of potential suppliers
3. Contracting
4. Contract performance conditions
5. Contract implementation.

All recommendations address how to include data ethics in the development and operations of AI-based services and solutions, identifying the relevant processes, tools and methodologies to integrate data ethics requirements and demonstrate their fulfilment.

### **The need for a common standard**

This whitepaper demonstrates the need for a common **standard for data ethics within the EU for the public procurement of AI-based services and solutions**. Such a step would further a human-centric, sustainable usage of AI in the public sector and provide leverage to a business community that is eager to sell AI-based services and solutions but lacks insight into fundamental rights, IT security and data ethics.

This whitepaper suggests that the following initiatives are integrated and promoted within the European Union's Public Procurement Strategy.

These recommendations encompass the integration of data ethics principles in regulation, political priorities, training and awareness-raising initiatives.

- **An EU directive on the public procurement of AI-based services and solutions in the public sector**

A new directive or amendments to existing EU regulations should legally require member states to consider data ethics when procuring trustworthy AI for public sector purposes. The requirement could be integrated into all phases of the procurement process. Such integration should be guided by a due diligence approach, as illustrated by this white paper.

- **Guidelines on the public procurement of AI-based services and solutions in the public sector**

As an alternative or supplement to an EU directive, official EU guidelines should be adopted. The main purpose should be to guide the establishment of procurement and due diligence processes in public sector. It should illustrate how the inclusion of data ethics in the aforementioned processes may lead to procurement of trustworthy AI, including data ethics.

Also, these guidelines should inform private sector suppliers about requirements and expectations regarding their governance and management schemes in relation to design, development, deployment and maintenance of AI-based services and solutions.

- **Inclusion of data ethics as a strategic policy priority in the Public Procurement Strategy of the EU**

The EU Procurement strategy reflects the need to consider the environmental and social impact of new products and services as an integrated part of the procurement process. In light of the rapid development of AI and the expected influence on the decision-making processes of public administrations, it seems both timely and necessary to include data ethics as a new strategic priority when it comes to policy.

- **A training toolkit on data ethics in the procurement of trustworthy AI-based services and solutions**

To promote the principles of data ethics and improve knowledge and awareness of data ethics and due diligence in public procurement processes, a training toolkit should be developed.

It should address the public sector as the contracting party in the procurement process and the private sector in its role as the bidder on tenders issued by public entities.

- **A data ethics handbook for the procurement of AI-based services and solutions**

Complementary to the training toolkit, a handbook should be distributed. It should explain data ethics principles and how to integrate them when designing, developing, deploying and maintaining trustworthy AI-based services and solutions.

- **An online help desk**

If implemented as suggested, contracting public authorities and private sector tenderers may need help understanding procedures, compliance requests, technical specifications and organizational measures such as processes and the demonstration of the fulfilment of requirements. An online help desk would meet that need and be instrumental for greater, more active engagement in data ethics and trustworthy AI in the EU.

## **Sources**

Reports & frameworks:

*AI Auditing Framework*, ICO, 2019

(including <https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/>)

*Opinion of the Data Ethics Commission*, 2019

[https://datenethikkommision.de/wp-content/uploads/191023\\_DEK\\_Kurzfassung\\_en\\_bf.pdf](https://datenethikkommision.de/wp-content/uploads/191023_DEK_Kurzfassung_en_bf.pdf)

*Automating Society: Taking Stock of Automated Decision-Making in the EU*, AlgorithmWatch in cooperation with Bertelsmann Stiftung, 2019.

*Data Flows - Future Scenarios: In-Depth Analysis for the ITRE Committee*, Simon Forge & Collin Blackman, 2017.

*Communication: A European strategy for data*, European Commission, 2020.

*Ethics Guidelines for Trustworthy AI*, European Commission's High-Level Group on AI, 2019.

*Menneskerettigheder og offentlige indkøb*, Institut for Menneskerettigheder, 2019

*OECD Guidelines for Multinational Enterprises*, OECD 2011

*OECD Due Diligence Guidance for Responsible Business Conduct*, 2018

*Policy and investment recommendations for trustworthy Artificial Intelligence*, 26 June 2019.  
European Commission's High-Level Group on AI.

*UN Guiding Principles for Business and Human Rights*, 2011

*White paper on artificial intelligence - A European approach to excellence and trust*, European Commission, 2020

*World Economic Forum, Guidelines for AI Procurement - Whitepaper*, September 2019

Books:

*Håndbog i dataansvarlighed*, Djøf 2020. Birgitte Kofod Olsen.

*Data Ethics - The New Competitive Advantage*, 2016. Gry Hasselbalch & Pernille Tranberg.

# DATAETHICS

[Dataethics.eu](http://Dataethics.eu)  
[info@dataethics.eu](mailto:info@dataethics.eu)

CVR 38465724  
ISBN 978-87-972168-0-4



# Emerging Technologies and Acquisition

*How Blockchain, RPA, Data Analytics, and AI are Enabling Federal Procurement Transformation*

**Emerging Technology Community of Interest / Acquisition Community of Interest**

Date Released: April 26, 2021

## Synopsis

One of the major drivers of procurement reforms in recent times has been the advent of technology in the federal acquisition process. From being an enabler to becoming a key driver of procurement's strategic transformation — technology is allowing procurement organizations to demonstrate true transformative effects across the federal government. With technology, procurement will no longer be just a purchasing or sourcing function, but instead transform into playing the role of an innovator, integrator, and collaborator, impacting agencies in myriad ways and driving overall performance for government IT management.

New and emerging technologies such as blockchain, robotic process automation (RPA), artificial intelligence (AI), and big data can help procurement find new avenues to transform, and in the process, drive greater spending visibility, increased compliance, enhanced accuracy, as well as achieve significant cost savings. As digital procurement and procurement transformation gain momentum, procurement organizations must identify and implement the right technology that will bring the desired value. This paper will look at the various technologies being implemented throughout procurement organizations across the federal government today.

*This page is intentionally blank*

### **American Council for Technology-Industry Advisory Council (ACT-IAC)**

The American Council for Technology-Industry Advisory Council (ACT-IAC) is a non-profit educational organization established to accelerate government mission outcomes through collaboration, leadership and education. ACT-IAC provides a unique, objective, and trusted forum where government and industry executives are working together to improve public services and agency operations through the use of technology. ACT-IAC contributes to better communication between government and industry, collaborative and innovative problem solving, and a more professional and qualified workforce.

The information, conclusions, and recommendations contained in this publication were produced by volunteers from government and industry who share the ACT-IAC vision of a more effective and innovative government. ACT-IAC volunteers represent a wide diversity of organizations (public and private) and functions. These volunteers use the ACT-IAC collaborative process, refined over forty years of experience, to produce outcomes that are consensus-based. To maintain the objectivity and integrity of its collaborative process, ACT-IAC welcomes the participation of all public and private organizations committed to improving the delivery of public services through the effective and efficient use of technology. For additional information, visit the ACT-IAC website at [www.actiac.org](http://www.actiac.org).

### **Emerging Technology Community of Interest and Acquisition Community of Interest**

This paper is a joint effort between the ACT-IAC Emerging Technology and Acquisition Communities of Interest (COI). The **Emerging Technology COI** mission is to provide an energetic, collaborative consortium comprised of leading practitioners in data science, technology, and research, engaged with industry, academia, and public officials and executives focused on emerging and leading technologies, which transform public sector capabilities. The **Acquisition COI** mission is to connect Government (procurement and program) and industry to new ideas and approaches to improve Federal acquisition in a forum that enables collaboration, encourages exploration and innovation, and generates results that have impact and endure.

### **Disclaimer**

This document has been prepared to contribute to a more effective, efficient, and innovative government. The information contained in this report is the result of a collaborative process in which several individuals participated. This document does not – nor is it intended to – endorse or recommend any specific technology, product, or vendor. Moreover, the views expressed in this document do not necessarily represent the official views of the individuals and organizations that participated in its development. Every effort has been made to present accurate and reliable information in this report. However, neither ACT-IAC nor its contributors assume any responsibility for consequences resulting from the use of the information herein.

### **Copyright**

©American Council for Technology, 2021. This document may not be quoted, reproduced and/or distributed unless credit is given to the American Council for Technology-Industry Advisory Council.

*This page is intentionally blank*

## Table of Contents

Introduction and Executive Summary .....	7
Approach Methodology .....	9
Identifying Emerging Technologies.....	11
Solutions Identified .....	12
Pilot IRS DATA Act, Internal Revenue Service.....	12
Current State.....	12
Description of the Portfolio .....	13
Constraints.....	13
Transformation Enablers .....	14
Stakeholders .....	14
Technology Selection .....	14
Accomplishments to Date.....	15
Artificial Intelligence for Past Performance Prototypes and Piloting Initiative, Department of Homeland Security, Procurement Innovation Lab .....	16
Description of the System or Portfolio .....	16
Future State .....	17
Constraints.....	18
Transformation Enablers .....	19
Stakeholders .....	19
Business Process Model.....	20
HHS and the BUYSMARTER Initiative.....	22
Current State.....	22
Future State .....	23
Constraints.....	27
Transformation Enablers .....	27
Stakeholders .....	27
Business Process Model.....	27

Organizational Readiness.....	27
Technology Selection .....	27
Lessons Learned .....	27
Conclusion.....	28
People produce permanent processes .....	28
A starting point for today's practitioners .....	29
Next steps for a team new to these technologies.....	30
Authors.....	31
References .....	32

## Table of Figures

<b>Figure 1:</b> Notional Acquisition Life Cycle .....	7
<b>Figure 2:</b> CPARS AI Initiative Project Timelines .....	18
<b>Figure 3:</b> BUYSMARTER Before and After State .....	22
<b>Figure 4:</b> BUYSMARTER Business Philosophy .....	23
<b>Figure 5:</b> BUYSMARTER UI/UX .....	23
<b>Figure 6:</b> BUYSMARTER UI Example 1.....	24
<b>Figure 7:</b> BUYSMARTER UI Example 2.....	25
<b>Figure 8:</b> BUYSMARTER UI Example 3.....	25
<b>Figure 9:</b> BUYSMARTER UI Example 4.....	26
<b>Figure 10:</b> BUYSMARTER UI Example 5.....	26

## Introduction and Executive Summary

If the acquisition workforce had an “official” mission statement that could guide their efforts, it would most likely be summed up in these two paragraphs from the Federal Acquisition Regulation (FAR):<sup>1</sup>

The vision for the Federal Acquisition System is to deliver on a timely basis the best value product or service to the customer, while maintaining the public’s trust and fulfilling public policy objectives. Participants in the acquisition process should work together as a team and should be empowered to make decisions within their area of responsibility.

Along with this, the FAR outlines:<sup>2</sup>

Procurement policies and procedures that are used by members of the Acquisition Team. If a policy or procedure, or a particular strategy or practice, is in the best interest of the Government and is not specifically addressed in the FAR, nor prohibited by law (statute or case law), Executive order or other regulation, Government members of the Team should not assume it is prohibited. Rather, absence of direction should be interpreted as permitting the Team to innovate and use sound business judgment that is otherwise consistent with law and within the limits of their authority. Contracting officers should take the lead in encouraging business process innovations and ensuring that business decisions are sound.



*Figure 1: Notional Acquisition Life Cycle*

Not only is the acquisition workforce entrusted with obtaining the American taxpayer the best possible value in the procurements they deliver, the individuals who assist those members of the acquisition workforce - the participants in the acquisition process - play an essential role in achieving that mission.

Similarly, if the acquisition workforce had a mandate, it would be to use any and all means available to them within the four corners of the law in order to innovate and implement those processes and procedures which allow for quick adoption of emerging technologies, as well as adaption within and throughout the public sector. In fact, for those teams with the budget and

staff constraints that many smaller programs and projects have, using the freedom to pursue those ideas is necessary for success.

To that end, the use of emerging technologies to transform federal procurement has accelerated under the direction of the Office of Federal Procurement Policy (OFPP). OFPP is a component of the Office of Management and Budget (OMB) whose mission is to provide overall direction for government-wide procurement policies, regulations, and procedures and to promote economy, efficiency, and effectiveness in acquisition processes.<sup>3</sup> Joanie Newhart, associate administrator of acquisition workforce programs within OFPP, said at the FedScoop 2019 IT Modernization Summit, that her office and the federal Chief Acquisition Officer Council are “Focused on emerging technology and how can we use it in acquisition. We think it’s going to explode this year, so we want to get in front of it and use it wisely.”<sup>4</sup> This shift in using technology to transform a traditionally manual, paper-based process such as federal procurement had its roots in the previous administration’s President’s Management Agenda (PMA). That PMA outlined the strategies needed to address the critical challenges where the federal government as a whole still operates in the past.<sup>5</sup>

Further, the PMA outlined various Cross-Agency Priorities—or CAP—goals for each of central areas of transforming government.<sup>6</sup> As stated on the website for the CAP Goals:<sup>7</sup>

As a subset of Presidential priorities, CAP Goals are used to implement the President’s Management Agenda and are complemented by other cross-agency coordination and goal-setting efforts. CAP Goals are updated or revised every four years with each Presidential Administration’s term.

Furthermore, the PMA outlined:

CAP Goals fall into four categories: key drivers of transformation, cross-cutting priority areas, functional priority areas and mission priority areas.

This mission and this mandate are part of the reason why the United States of America serves as the global leader in emerging and innovative technologies. In fact, the FAR provides public servants with the platform to partner with the engine of the American economy - those businesses, both large and small, that develop these advancements. This freedom to improve and iterate upon existing procurement processes is an essential component of keeping our position as global leaders. It is also essential to helping ensure the government promotes those technologies and tools that provide the greatest return on the investment being made with taxpayer dollars.

Further, these technology initiatives were also directly aligned with the CAP Goal: “Shifting from Low-Value to High-Value Work”<sup>8</sup> and “Frictionless Acquisition”, respectively.<sup>9</sup> These two CAP Goals combined created the vision and the future of federal procurement through shifts in technology, practices, and culture through the continued action plans and focused leadership being executed to realize these goals. However, the goal statement of the “Frictionless Acquisition” CAP Goal encapsulated the future state of federal procurement where emerging technologies can have a major impact:

*"The Federal Government will deliver commercial items at the same speed as the marketplace & manage customers' delivery expectations for acquisitions of non-commercial items by breaking down barriers to entry using modern business practices and technologies."<sup>10</sup>*

In order to effectively make use of capabilities in emerging technologies for use in future procurement, it is imperative that those who participate in acquiring emerging technologies keep in mind who the ultimate end users of their procurements are. Whether it is the American citizen interacting with an agency or department, an employee benefiting from automation or other assistive technologies, or even industry partners, understanding who it is that will ultimately derive value from the work being performed is key.

Defining that value being derived by that ultimate end user is the first step an innovative acquisition process participant must take. This will lead to a number of goals being satisfied including but not limited to:

- Developing requirements around the true objectives of the project.
- Selecting evaluation factors that will ensure focus is placed on technical brilliance.
- Sharing the solicitation with communities and in locations that will encourage participation from those sources in industry who have innovative solutions, but not necessarily established firms with a federal sector business footprint (i.e., non-traditional firms).<sup>11</sup>

It is hoped that the concepts discussed in this white paper will help any participant in the procurement process, whether directly involved in the acquisition itself, or indirectly as support staff. Additionally, that the practical advice and frameworks presented here provide members of the acquisition workforce, whether new or experienced, a starting point for discussions within their agencies and departments on the art-of-the-possible. Finally, this white paper will also begin to connect those individuals who wish to learn more about what we have presented to one another, as a robust network of collaborators will be the quickest way to promote innovation that sticks.

In closing, having developed these during a period which saw the regular forces of daily operations and the irregular forces of a national emergency declaration, combine to impact the American public in profound ways. The ACT-IAC Emerging Technology and Acquisition Communities of Interest proudly recognize the members who volunteered the precious time they had remaining after supporting their agencies to this effort.

## Approach Methodology

Understanding the importance of utilizing emerging technologies, both within general efforts to support the mission of a program, as well as to support the procurement process itself, a project approach, broken into phases, was utilized. By outlining this process, the hope is that others can verify, validate, and even replicate the information utilized to derive these conclusions. With four different phases, during which a number of different individuals graciously assisted the core

group of authors, this research was performed with a degree of mindfulness as it related to the FAR and the limitations -- as well as the flexibilities -- it offers.

Seeing that all supplements build off this body of regulations, it serves as the foundation for any and all efforts that can be implemented to improve the efforts that make up the acquisition lifecycle and the procurement process in general. With that holistic view in mind, the viewpoints of those involved in developing requirements, acquisition planning, contracting, and post-award contract administration were required from the private sector and public sectors. Public sector individuals held positions that constitute those considered part of the acquisition workforce, such as Contracting Officers, Contracting Specialists, Contracting Officer Representatives, Project Managers, and Program Managers.

With the current state defined and key performance indicators (KPIs) ready to be measured, the next phase involved interviews with individuals from the various agencies represented in the use cases (Internal Revenue Service, Department of Homeland Security, and Health and Human Services) who held the roles identified as essential to understanding the issues being researched. These interviews were designed to be free-flowing conversations that included a set of predefined questions while permitting the interviewee to provide their observations on current workflows. They took place among four government settings and among enough of the private sector to provide a review of cross-industry leading practices.

The intention of these interviews was to identify key friction points across several domains, including processes, people, and policies. Through these discussions, a review of current and emerging technologies and capabilities was performed, identifying the factors that contribute to successful implementations. These factors identified were broadly categorized into those that should be considered for ongoing use, as well as those which had room for improvement.

Similarly, it was important to use these interviews to identify the actual key technologies, processes, and people that are implicated in any discussion related to the procurement and use of these advances. The technologies identified included blockchain, robotic process automation (RPA), data analytics, and artificial intelligence (AI). The processes implicated were those related to training of the workforce itself, application of advanced techniques, and continuous improvement of business workflows and processes themselves. The people involved are the end-users, dedicated trainers, technical subject matter experts (SMEs) involved in the implementation of these advances, and staff that can support programs with the integration of technologies throughout the enterprise.

With these discussions completed, Phase 3 involved the vetting of the technology, people, and process combinations that were suggested. The various impacts, both measured and theoretically possible, were analyzed to find those that suggested they would provide a positive impact to those KPIs. Using industry and government examples to provide the qualitative and quantitative data necessary for this analysis, several use cases and organizations were identified.

Once that analysis was completed, Phase 4 involved writing this paper. Much like the concept of Continuous Acquisition Improvement, discussed further below, the recommendations and conclusions of this white paper should be considered ripe and ready for the next iteration. This

document should be considered a living document, with nothing permanent or sacrosanct. What works today may not work tomorrow, and similarly, that perfect state tomorrow may need something different done today.

## Identifying Emerging Technologies

As determined through the phased approach of this project, the four key emerging technologies identified as having the potential to improve the government are **blockchain**, **RPA**, **data analytics**, and **AI**.

**Blockchain**, along with the concept of smart contracts and ledgers, have been used in several successful applications in industry and government alike. In the public sector, one of the more intriguing applications of blockchain technology to the procurement process is the work being performed by the Department of Health and Human Services (HHS) “Accelerate” program. As the first blockchain-based program in the federal government to get an authority to operate (ATO), HHS not only provides technical advances, but also procurement advances that have shaped several important conversations about what is possible.

**RPA**, which has been utilized by agencies such as the Internal Revenue Service (IRS), General Services Administration (GSA), Department of Labor (DOL), and the Army to create a number of different bots (e.g., Contractor Responsibility Determination, Section 508 Clause Review, etc.) has provided immediate results. Furthermore, the way in which these bots were developed, procured, and integrated into these agencies has provided several opportunities to implement new technologies. The sheer amount of data that can be collected using these bots can be used as an opportunity rather than an end goal.

Collecting that data, whether in order to satisfy requirements outlined in the Open, Public, Electronic and Necessary (OPEN) Government Data Act, the PMA, or any other number of statutes and policies, allows for several **data analytics** related advances. The collection, organization, and visualization of all this data to be used in decision-making processes is one of the more tangible forms of emerging technology. It also makes it one of the easier advances to understand when it comes to the importance of applying modern practices to the evaluation of the services and products associated with its use. This data is critical as it holds valuable insights and patterns that can help organizations assess potential risks, find new business opportunities, and most importantly, improve the overall efficiency and productivity of organizations.<sup>12</sup>

**AI and Machine Learning (ML)** allow data scientists to delve deep into massive volumes of data and uncover meaningful insights from it. Once the differences in modern data analytics versus previous iterations of the technology are understood, AI becomes possible. As one of the latest entrants into the emerging technology space, and one of the most compelling in terms of potential and importance, it is absolutely necessary for the public and private sector to work together to determine the optimal and most efficient path forward in terms of buying, using, and enhancing AI. The reports published by the National Security Commission on Artificial Intelligence (NSCAI) leave no doubt as to the need for the acquisition workforce to be prepared to use and

procure AI, and the hope is that the conclusions of this research will be one of the building blocks towards establishing that necessary competency.

## Solutions Identified

Industry best practices adapted for public sector use cases have shown the identified categories of impediments are not only surmountable, but that there are several ways in which teams can implement their optimal solutions. The ultimate result is utilizing agile techniques in the processes associated with emerging technologies and modernization at an agency. In acquisitions, this can be done by treating the procurement package as a product while incentivizing and encouraging positive behaviors and thinking when it comes to continuous improvement.

### ***Pilot IRS DATA Act, Internal Revenue Service***

One part of modernization efforts at the Internal Revenue Service (IRS) is focused on improving financial and non-financial procurement data. To improve the accuracy and transparency of data in the Federal Procurement Data System – Next Generation (FPDS-NG), as well as the compliance with the Digital Accountability and Transparency Act of 2014 (DATA Act), IRS launched its DATA ACT initiative. The solutions to the project were focused on reducing manual work, shifting personnel from low-value to high-value tasks, in conjunction with Office of Management and Budget memorandum M-18-23, “Shifting From Low-Value to High-Value Work”, and the PMA’s CAP Goal number six; “Shifting From Low-Value To High-Value Work”.<sup>13</sup>

### **Current State**

The Treasury Inspector General for Tax Administration (TIGTA) conducts audits of the IRS’s compliance with the DATA Act. The DATA Act requires federal agencies to report financial and award data in accordance with established data standards. The data is then published on [USAspending.gov](#), for use by decision-makers and the American taxpayer. The 57 DATA Act data elements<sup>14</sup> include Parent Award ID, Current Total Value of Award, Action Date, and Primary Place of Performance Congressional District. In its Fiscal Year 2019 report, TIGTA reported inaccuracies across the 57 DATA Act data elements, ranging from 5 to 52 percent.<sup>15</sup> In its planned corrective action responding to the audit, the IRS committed to develop a quality assurance review process to ensure DATA Act information is accurate.

The data that IRS certifies, and TIGTA audits for DATA Act compliance, is derived from the IRS data in FPDS-NG. Prior to Pilot IRS DATA Act, a senior level IRS employee would manually correct FPDS-NG as needed in response to TIGTA audit errors, inquiries from the Office of the Chief Financial Officer (OCFO), and errors identified within IRS Procurement. Manual corrections took roughly three minutes per data element. Applying the TIGTA audit accuracy error rate to a year’s worth of IRS contract actions, it would take nearly two full-time equivalent staff to correct data element errors.

## Description of the Portfolio

According to the Pilot IRS webpage on IRS.gov:<sup>16</sup>

*Pilot IRS is an iterative procurement technique focused on outcomes that allows the Internal Revenue Service (IRS) to test new technologies on faster timelines. If a solution fails to meet expectations, then it will not proceed to the next phase of funding. This methodology creates an agile approach to identify, test, and deploy solutions that support the mission, regardless of whether the solution, technology, or service currently resides within the IRS or the federal government.*

Like all federal agencies, the IRS is under several mandates to modernize its operations and its mission of administering the nation's tax code. However, the IRS lacks any research and development capability or special acquisition authorities such as other transaction authority (OTA) or a commercial solution opening program to rapidly procure emerging technologies for these modernization efforts. Further, as previously mentioned, the IRS does not have funds allocated for research and development.<sup>17</sup>

To solve this issue, the Office of the Chief Procurement Officer (OCPO) used FAR Parts 12 and 13 to create Pilot IRS; a phased, incremental funding procurement initiative that will let the agency test new technologies on faster timelines with reduced risk. This initiative was also in keeping with modular contracting best practices, as outlined in guidance provided by the OFPP.<sup>18</sup> Under Pilot IRS, the agency issues a proposal under what is referred to as a "Solution Challenge", specific problems that the IRS is trying to solve with emerging technologies. The IRS issued a request for proposals for the first solution challenge, DATA Act Improvements, on August 23, 2019.<sup>19</sup> The Pilot IRS DATA Act had three goals:

1. Improve the data which resides in FPDS-NG.
2. Limit the amount of manual work required by government personnel in improving the IRS data.
3. Achieve incremental improvement in IRS data in the near term – the aspirational goal was to improve data no later than 31 December 2019.

## Constraints

The Pilot IRS DATA Act team faced four general constraints: culture, budget, security, and information technology onboarding, especially the Enterprise Life Cycle (ELC). Culture and budget remain mostly within IRS Procurement's control and have therefore been more navigable. To manage concerns surrounding use of robotics process automation solutions, the team has worked to socialize the high benefits and low risks of the solution. Examples included:

- Held demonstrations of the solutions to showcase capabilities and controls.
- Controlled the pace of corrections and completed full quality checks of results.
- Notified contracting personnel before making any FPDS-NG change, to prevent any surprises or fear of data manipulation.

To date, IRS Procurement has funded the Pilot IRS DATA Act. As the project is a top priority for IRS Procurement, the program was able to secure the necessary funding. This may become more challenging for Phases 4.1 (scaling and deployment) and beyond, as budgets may be constrained by pandemic-related priority shifts.

Another major constraint identified outside of IRS Procurement was the IRS ELC process. The IRS requires information systems-related business changes to be approved through the ELC process. ELC outlines project paths and numerous required artifacts.

However, the IRS found that there was little, if any, ability to tailor the ELC process to fit the scope of the project at hand and the need for rapid prototyping and development. Pilot IRS DATA Act will likely run into delays, which could potentially be significant as the team navigates the process and seeks ELC approvals. Currently, there is not a streamlined approach to ELC for RPA projects, though ELC and the Office of the Chief Information Officer (OCIO) hopes to develop one in the near future.

### **Transformation Enablers**

The Deputy Chief Procurement Officer was a staunch supporter of innovative technologies and the main champion of Pilot IRS DATA Act. The willingness to take risks, while doing so incrementally and continually evaluating return on investment (ROI), is critical to this project's success. Pilot IRS DATA Act was solicited substantively different from how the government normally buys technology, resulting in non-traditional contractors that had not previously done business with the federal government submitting responses. Pilot IRS aggressively streamlined a cost-effective approach to testing and deploying technology solutions that had an immediate impact on its mission and how the IRS supports the American taxpayer.

### **Stakeholders**

Pilot IRS DATA Act impacts a wide array of stakeholders:

- IRS OCPO—Executives, Procurement Innovation Branch, all Contracting Officers and Contract Specialists, contract writing system team (Procurement for the Public Sector and Folders Management)
- IRS OCFO—Financial personnel collaborating with OCPO for DATA Act submissions
- IRS OCIO—Information technology personnel providing IRS laptops and necessary licenses to contractor personnel, as well as involved in the ELC process
- IRS ELC—includes multiple IRS business units such as Privacy, Government Liaison, and Disclosure and Security Risk Management Office

### **Technology Selection**

Pilot IRS DATA Act was solicited as Solution Challenge One under the Pilot IRS framework. The intent of the Pilot IRS DATA Act solicitation was to acquire innovative tools to meet the three goals previously mentioned and the IRS was open to any viable solution to achieve the goals. The solicitation was well received by industry, resulting in 38 proposals. Comparative evaluations in accordance with FAR 13.106-2(b)(3) were used to determine the best value based on: Technical

(capability of the solution to meet the primary goals), Past Performance, and Price. The requirement was solicited as an agile modular approach that would allow flexibility to pivot as needed and this has contributed to the program's success as additional within scope items have been identified throughout the phases.

Based on the written proposals, ten contractors were selected to participate in technical demonstrations of their solution, to include discussion on Phases 2-4, ROI, and funding levels. Five contractors were ultimately selected to receive contracts and funding for Phase 1 (Proof of Concept). The technical solution of each selected contractor was slightly different, allowing the IRS to make funding decisions in future phases based on the contractor(s) whose solution proved to be the most promising. Two of the five selected contractors received funding to continue into Phase 2, and both remaining contractors continue to support this requirement and improve their solutions in Phase 4.

These contractors developed and began to deploy solutions using RPA, natural language processing (NLP), and AI. The contractors' solutions review data in FPDS-NG and on contract documents, compare data elements between the two data sets, and correct FPDS-NG as requested. Recent modifications to the DATA ACT contracts, which began in Phase 4, have contractors expanding their solutions to interact with the contract writing system.

### **Accomplishments to Date**

Through the recently completed Phase 3, IRS Procurement realized significant Return on Investment (ROI). ROI, by contractor, was 78 percent and 116 percent over break-even, comparing the amount of labor hours and costs associated with manually performing these actions. Assuming the contractors were tasked with correcting a year's worth of DATA Act errors, the contractors' solutions could replace the work of nearly two full-time equivalents per year.

Specifically related to the DATA Act, IRS Procurement used one contractor's solution to correct data elements in real-time, which is especially critical as DATA Act certifications moved from quarterly to monthly. The Pilot IRS DATA Act team loaded the contractor's spreadsheet "corrections file" with 20 FPDS-NG dates signed that required correction. These errors were identified by the OCFO in preparing for the monthly DATA Act certification. The team is about to correct an additional 178 FPDS-NG dates signed, representing nearly two percent of annual volume in this high error rate data element. For this batch of corrections, the team used the solution of the second contractor: the contractor's AI solution extracted dates signed from contract award documents, compared the data with that in FPDS-NG, and identified discrepancies.

IRS Procurement was also able to quickly deploy one contractor's solution to review and correct COVID-19-related FPDS-NG data. The contractor's RPA bot scans FPDS-NG weekly for Department of the Treasury (USDT) contract actions related to COVID-19 and corrects the requested FPDS-NG data elements in less than one-quarter of a second per element. As of June 22, 2020, IRS Procurement corrected 76 COVID-19-related data elements with this solution.

## ***Artificial Intelligence for Past Performance Prototypes and Piloting Initiative, Department of Homeland Security, Procurement Innovation Lab***

Led by the Department of Homeland Security (DHS), OCPO, Procurement Innovation Lab (PIL), this project, focused on the Contractor Performance Assessment Reporting System (CPARS). CPARS is an evidence-based acquisition modernization project under the OFPP Acquisition Modernization Plan and is part of the “Frictionless Acquisition” CAP Goal of the President’s Management Agenda.

The CPARS AI effort focuses on improving the ability of the acquisition workforce at DHS to rapidly access relevant records from CPARS. The CPARS system serves as the central repository of contractor past performance assessment records. Contracting officers are required to consider contractor past performance when conducting source selection as part of the federal contracting process, and CPARS is the place they can access information to do so.

For several years, in reverse industry days (RIDs) and acquisition innovation roundtables (AIRs), as well as more informal meetings, industry raised concerns related to CPARS including issues related to data quality and stating that contracting officers were not using the system for acquisitions. The federal acquisition community, those on the front lines of contracting, shared similar concerns. CPARS has a wealth of data, with over 1 million records for over sixty-thousand vendors, but it is hard to rapidly access assessment reports that are relevant to a given source selection.

The difficulty leads to some less than ideal outcomes, primarily workarounds (like past performance questionnaires that operate outside of the CPARS system), but also a reduction in the quality of data in the system. This is part of a Catch-22 – the perceived value of the CPARS system is reduced when it is difficult for acquisition professionals to get data out of it, leading to diminished perceived value of their time spent inputting quality data into the system. Three out of four acquisition professionals identified the past performance evaluation process as an opportunity for intelligent automation to support the workforce. The challenge was then raised to industry: can emerging technology such as ML and AI help federal contracting professionals rapidly access relevant records from CPARS for a given source selection (set of vendors against a specific requirement, or solicitation) and provide them insight into the data that can aid them as they review and evaluate a vendors’ past performance?

### **Description of the System or Portfolio**

The type of solution solicited fell into the class of decision support systems, where ML solutions (falling under the general term of AI) are used to support the knowledge discovery process of a human. Such decision support systems are common within the business intelligence (BI) community where metrics and performance measures are summarized through interactive dashboards. Other terminology common to this domain is visual analytics, where researchers explore design principles and practices for creating novel tools and techniques to support analysis.

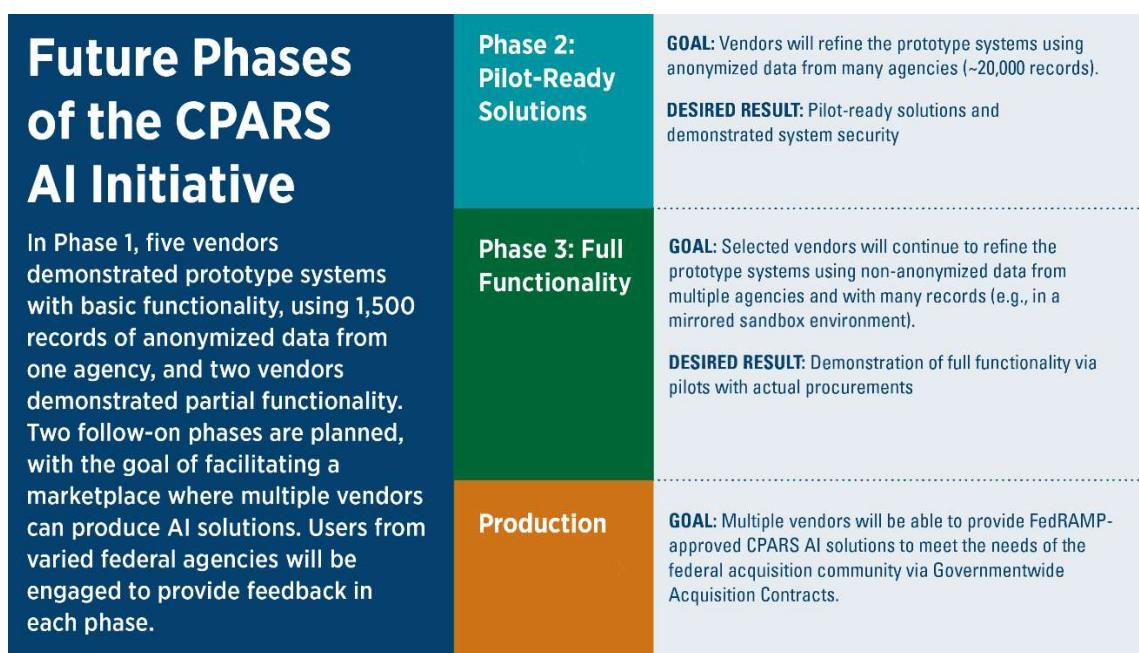
For the problem of using CPARS data during past performance analysis, this can be thought of as a classic information retrieval problem, which can be augmented through interactive visualizations. Given a series of records for a particular Dun and Bradstreet (DUNS) number, one can envision a tool that can graph average past performance histories related to the DUNS number, show outliers from the average, and support interactive filtering and drill-down into records. Underlying the summary and filtering can be machine learning techniques that can process text to help identify records that fall into the same business domain as the Statement of Work (SOW) being solicited. Text analysis (again falling under the umbrella of AI) can also be applied to the contracting officer notes to highlight sentiment within a textual description that may indicate mismatches to quality score rankings (for example, providing a score of “satisfactory” in a given performance metric, but having text descriptions that indicate unsatisfactory work).

While many deep learning solutions (another class of AI) rely on millions of records for training, this class of problem can leverage novel text analysis solutions (such as Bidirectional Encoder Representations or BERT), to support the information retrieval process. In this solution space, the goal is to order records based on relevancy, as well as highlight potential problematic records based on quality metrics and text descriptions related to past performance. By developing advanced search functions, aggregations and filters, the manual search process can be considerably improved, leading to efficiencies in procurement.

Such tools do not replace the human decision-making process. Instead, AI is used to augment the normal workflow done by the Contracting Officer. Historically, aggregation of scores, trend analysis, and text summary analysis were manual processes. By providing tools that can automatically aggregate scores, visuals that enable trend analysis, and highlights of potentially relevant text, decision support tools can reduce the cognitive burden on Contracting Officers and make them more efficient. As such, Contracting Officers will not rely on mechanistic comparisons, but they will utilize the decision support system to make their own informed decisions.

### **Future State**

Ultimately, the prototyping and piloting initiative for CPARS AI solutions intends to have multiple FedRAMP accredited cloud-based commercial Software as a Service (SaaS) solutions for federal agencies to procure and utilize to improve the work life of our federal acquisition workforce and to support more meaningful outcomes of source selection past performance evaluation.



*Figure 2: CPARS AI Initiative Project Timelines*

## Constraints

From the outset, DHS realized it did not want to develop its own IT systems/solutions. CPARS would remain the system of record for contractor past performance assessments, given the significant resources committed to the system and its modernization. Further, the intent was not to replace CPARS. Rather, DHS wanted to support the development of third-party solutions that connect to CPARS so that DHS can gain additional value from the data contained in the CPARS system.

DHS maintains rights to the data in the system but did not want to own the AI solutions that ingest the data and derive outputs from the data. DHS wanted to contract for the output from these AI solutions – in the form of CPARS AI Reports or Subscription Services with dashboards or similar visualization interfaces. If successful, the end state of this AI effort for CPARS for DHS will be a self-sustaining, multi-contractor commercial marketplace.

Currently, this marketplace does not exist. DHS views the government's role to be an incubator fueling the development of this new commercial marketplace, and DHS has an active stake in the success of these solutions. DHS also does not intend to contract for one solution. Multiple vendors in a production phase helps DHS keep costs down and improve quality through competition and diversity in solution offerings, which ensures DHS is not faced with vendor lock-in and helps incentivize industry to be attentive to model degradation and consistently correct model drift.

DHS is leading and building these solutions for the entire federal acquisition community. It took both the government and industry partners extended time to understand this, but these solutions will be available to every federal agency once they are production ready – likely through

Government-Wide Acquisition Contracts (GWACs). Finally, yet most importantly, DHS is steadfast in their commitment to develop user-driven solutions. Through a user-centered design approach, the DHS initiative will ensure their operational contracting community to be intimately involved in the development and design of these solutions.

### Transformation Enablers

DHS followed several goals in the development of the project:

1. Start with the end in mind – knowing the desired end state, recognizing constraints, and building backwards from this starting point enabled DHS to understand what was ahead and allowed them to scope, resource, and build the right coalition of stakeholders to support that journey.
2. Make time for the data – recognizing and planning for the time needed to prepare the data that digital transformation technology solutions such as these need for training and testing helped DHS ensure that vendors had what they needed. This required support from subject matter experts was critical to have those stakeholders on board from the start. It also required strong collaboration with industry who needed that data to be successful.
3. DHS supported the development of commercial solutions that did not currently exist, rather than asking industry to build solutions for DHS that the government would own and maintain.. First, it helped DHS identify the appropriate procurement authority – the Commercial Solutions Opening Pilot Program (CSOP) authority provided to DHS under the FY19 National Defense Authorization Act (NDAA).<sup>20</sup> It also helped DHS adopt the principle of minimal restriction on the commercial marketplace – DHS structured intellectual property rights to protect its rights to the CPARS data, not to own the rights to the commercial solutions the contractors were developing to meet DHS needs. This approach enabled the contractors to secure additional, private funding to support their ability to design and develop these commercial CPARS AI solutions, which the government would buy later – in the form of subscriptions or reports.
4. Building incrementally has been critical to ensuring the users and policy stakeholders understand the purpose and value of the CPARS AI solutions, as well as how the AI interacts with the CPARS data and supports user decision-making. Building incrementally has also ensured the government and industry are investing time in solutions that are meeting the government's objectives and helps ensure that each contractor has appropriate access to users, decision-makers, and feedback throughout the development of these commercial solutions. This model supports private sector investment in these solutions. This model also supports appropriate, necessary federal data governance and policy changes.

### Stakeholders

1. Industry – Contractors that do business with the federal government and have data in CPARS.

2. Industry – The contractors that are developing the commercial CPARS AI solutions.
3. OMB and OFPP.
4. Chief Acquisition Officer's Council (Agency Senior Procurement Executives serving as executive sponsors).
5. Procurement Council on E-Government.
6. The GSA's Integrated Award Environment (IAE).
7. Federal CPARS Users (Operational and Policy).
8. Federal Information Technology Subject Matter Experts (DHS OCIO partnership support through SME technical assistance for data preparation, technology assessment, and security accreditation).
9. Third-Party Partners (Technology Assessment & Validation).
10. Federal Innovation Stakeholders (including executive sponsors, champions).
11. Federal Information Technology Stakeholders (including executive sponsors, champions).
12. DHS Center of Academic Excellence – Center for Accelerating Operational Efficiencies (CAOE), Arizona State University (providing technology assessment and SME support).

### **Business Process Model**

In order for industry to address the challenge question, on behalf of the federal acquisition community, DHS issued a general solicitation using the CSOP authority derived from the Fiscal Year 2017 NDAA. They used that authority to obtain demonstrations of a proof of concept to determine the extent to which AI can assist Contracting Officers conducting past performance evaluations in making efficient and effective use of CPARS information. The government hoped, as a threshold, that the demonstrations would show that AI can help the Contracting Officer identify which records in CPARS contain the most relevant information to the source selection in question. The government also desired data-driven and evidence-based recommendations about opportunities to improve the data quality of the past performance information inputted by Contracting Officers into the CPARS, based on the provided test data set and informed by the development of the prototype.

The solicitation was posted in August 2019 and 40 proposals were received. Peer review identified nine that demonstrated technical merit and awards were negotiated with those nine contractors. The following constraints were incorporated into the initial CPARS AI prototype awards:

- The government does not require delivery of any intellectual property.
- The government will provide test data for use in the prototype design and development. This test data will be sanitized to safeguard the integrity of the CPARS system. The test data will be provided as a data table developed using a parsing script from PDF CPARS reports generated from the CPARS system. The data table will be comprised of approximately 1,000 records representing approximately eight contractors. Additionally, a sample solicitation will be provided as part of the test data for purposes of developing the proof of concept or viable prototype.
- To the extent that FAR 42.1503 or any other FAR provision limits use of information

from CPARS to supporting specific source selection decisions, from a policy perspective, OFPP does not view this AI initiative as an inappropriate or misuse of the data, as the purpose of the proof of concept is to explore how to maximize the value of the CPARS data specifically for purposes of source selection. Further, DHS will take appropriate safeguards to protect the data and intends to limit the use of the data to only the number of reports sufficient to enable DHS to obtain meaningful proof of concepts or viable prototypes.

- The government will provide regular, scheduled access to SMEs during the performance of resultant contracts, including operational contract specialists who use CPARS to perform past performance evaluations, for purposes of user feedback during the development of the proof of concept / working prototype.
- The government may engage academic researchers through a DHS Academic Center of Excellence partnership agreement for purposes of supporting the government's assessment of the validity and reliability of the proof-of-concept prototypes.
- The government prefers solutions that will ultimately be able to comply with Federal Information Security Management Act (FISMA) of 2002 and 2014 (as amended) and IAE design standards which may mean a solution capable of obtaining a Federal Risk and Authorization Management Program (FedRAMP) certification or ATO. Phase 2 contracts will require the final delivered CPARS AI solution to obtain the FedRAMP certification and DHS ATO. CPARS AI solutions unable to achieve those accreditations will not be eligible for piloting contracts (Phase 2).
- The government requires that the solution be "explainable" – meaning that the solution shall describe the methodology that was applied to the data to obtain the results and outcomes from the use of the AI methodology

Government stakeholders attended each of the demonstrations, including the final viable prototype demonstration, and these stakeholders included users, as well as decision-makers. Commercial companies were awarded \$50,000 in seed funding, provided test data, and consistent access to a product owner and users to iteratively develop the prototypes with scheduled demonstrations along the four-month period. In January 2020, the initial prototypes were delivered. Most of the vendors' prototypes effectively demonstrated the capabilities of these cognitive solutions to the use case and warranted further prototyping.

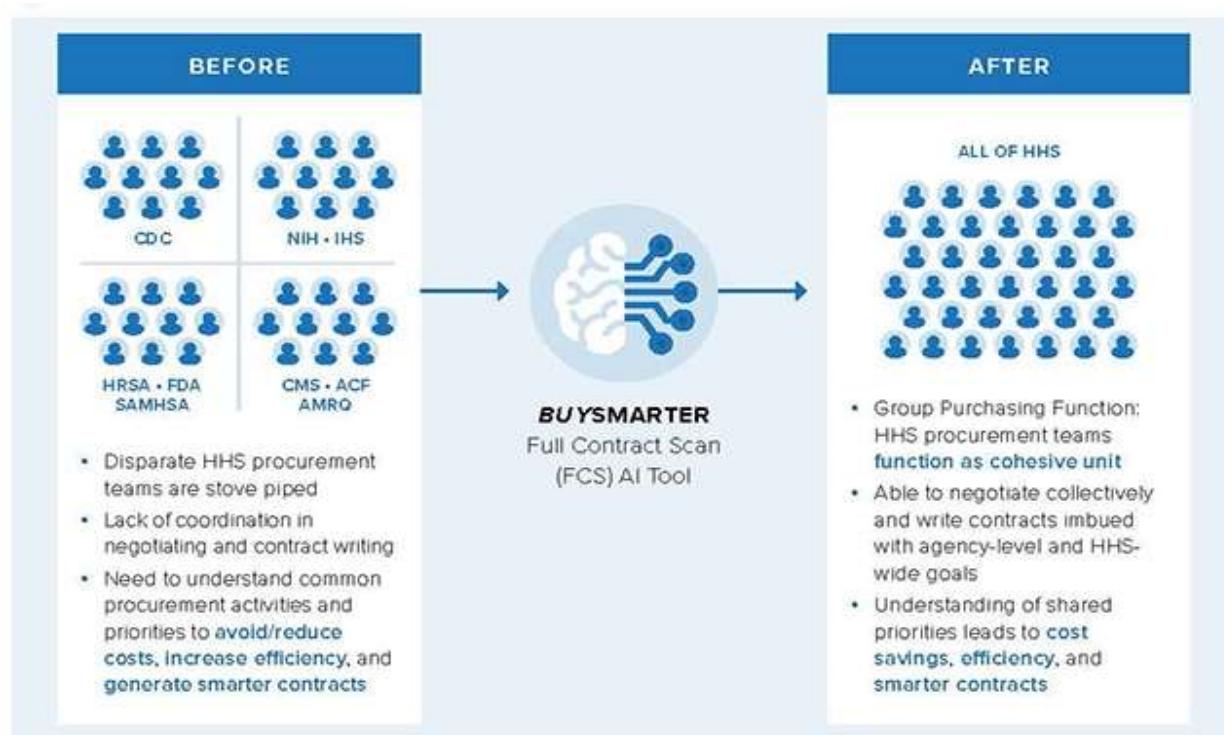
DHS continues the building out the next phase of prototyping by validating user stories, developing the baseline security control requirements for the cloud-based SaaS solutions, coordinating the cross-agency data preparation of a larger test data set, and lining up additional agencies to support this journey, including expanding the user group that engages with the vendors throughout the prototyping process. Ultimately, if the prototyping and piloting phases prove as intended, DHS will have multiple FedRAMP accredited cloud-based commercial SaaS solutions for federal agencies to procure and utilize to improve the work life of our federal acquisition workforce, and to support more meaningful outcomes of source selection past performance evaluation.

## **Department of Health and Human Services (HHS) and the BUYSMARTER Initiative**

The intent of **BUYSMARTER** is to maximize HHS' enterprise purchasing power to establish a cohesive acquisition structure across the department in order to drive better pricing and achieve better terms and conditions, while consolidating the total number of contracts. This technical solution leverages AI technology to assist in spend data and identifying opportunities that support enterprise acquisitions. A solution includes the Full Contract Scan (FCS) AI Tool. The purpose of BUYSMARTER is to enable the 26 operating and staff divisions within HHS to operate as a singular buying entity by bringing all HHS contracting data together and then deploy AI to empower the HHS acquisition community with powerful analytical, collaboration, and negotiation capabilities.

### **Current State**

HHS has five acquisition management systems across HHS's 11 operating divisions and 15 staff divisions (OpDivs/StaffDivs). If HHS wanted to buy as an enterprise, it would have to search through and read thousands of contracts across the OpDivs/StaffDivs to determine who buys the same things, at what prices, and with what terms and conditions. HHS did not have the ability to "scan" full contracts and compare "like to like" requirements and items to increase bargaining power. This is where AI was the logical solution to solve this age-old problem.



**Figure 3: BUYSMARTER Before and After State**

## Future State

With the FCS AI tool, HHS can quickly scan existing contracts, negotiate collectively, and write future contracts with agency wide goals. The *BUYSMARTER* AI FCS Microservice's purpose is to enable the 26 OpDivs and StaffDivs within HHS to operate as a singular buying entity by leveraging the full suite of HHS contracting data and AI capabilities to empower the HHS acquisition community with powerful analytical, collaboration, and negotiation capabilities.

In short, the *BUYSMARTER* philosophy is, fundamentally, a human experience where a group of individuals from different agencies acquiring together as one, as

“...this tool is designed to AI-enable that human experience. *BUYSMARTER* is a human experience of working together to achieve cost savings and improvements to mission capabilities.”<sup>21</sup>

*The key word is ‘Together’, as shown in the following graphic.*



*Figure 4: BUYSMARTER Business Philosophy*

*BUYSMARTER*'s approach empowers collaborative requirements gathering and all historical data that leverages AI technology to assist in understanding spend data and identifying opportunities that harness collective purchasing power while consolidating contract vehicles. The goal of the FCS AI tool is to give the Category Collaborative teams an entirely new capability to understand where the most spend is happening, where operating and divisions are contracting for the same things, and where the differences in prices paid are the most pronounced. Building the tool to meet the defined mission, the *BUYSMARTER* team had to leverage many types of AI tools, such as NLP, ML, Nearest Neighbor Inference Models, Predictive Search Engines, and dozens of AI open source code tools.

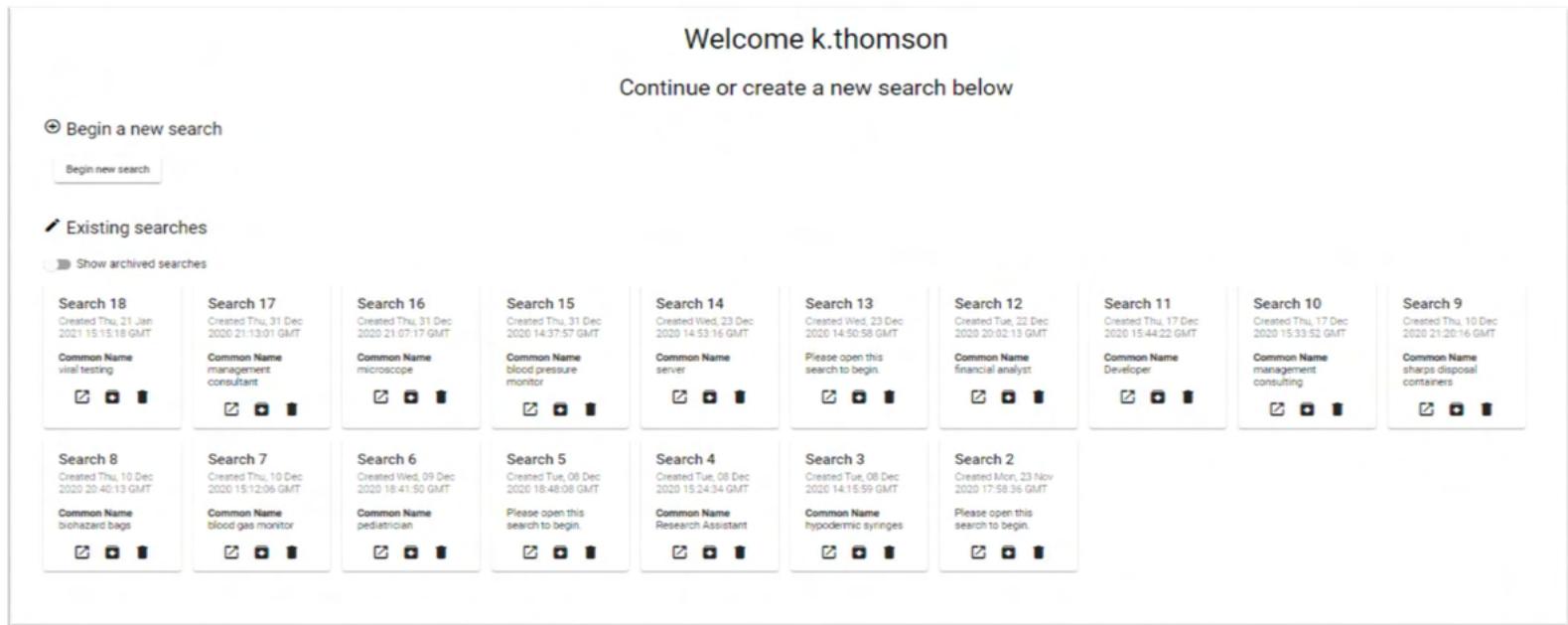
Having the ability to review the entire department's spend within a category, its sub-categories and down to the specific products/services will allow the agencies to make sound, logical decisions as to whether to pursue an enterprise acquisition for that product/service. The tool was designed to be very simple to use based on familiar tools that we use regularly:



*Figure 5: BUYSMARTER UI/UX*

- The search function looks like a regular **Google Search**
- The process follows a simple **TurboTax** style step-by-step process
- And the output looks and works just like **Excel**

The AI is designed to ‘look’ for information the same way you and I would – by making associations within and between the things we are searching for that make sense in what is known as an ‘inferential model’. Given the learning nature of AI, the tool will get better and smarter every time it is used. The tool currently processes 1.37 million contract attachments in under eight seconds. The navigation tool seen below allows the user to start a new search or to jump back into any of their historical searches.

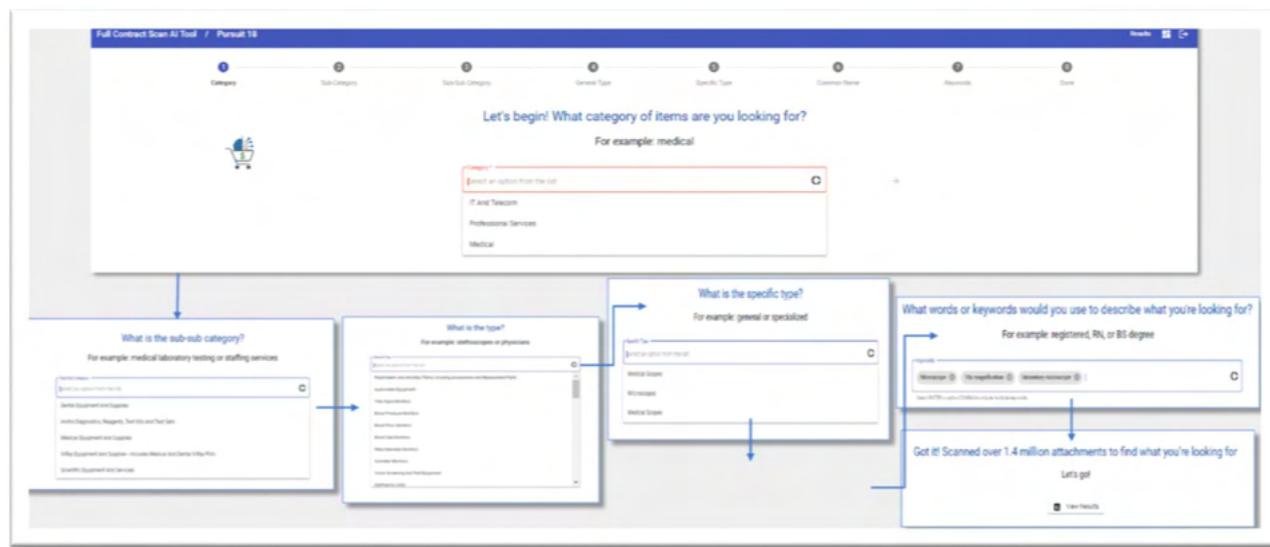


The screenshot displays a user interface for managing search history. At the top, a header reads "Welcome k.thomson" and "Continue or create a new search below". Below this, there are two sections: "Begin a new search" (with a "Begin new search" button) and "Existing searches" (with a "Show archived searches" link). The "Existing searches" section contains a grid of 18 search results, each represented by a card:

Search ID	Created Date	Common Name	Action Buttons
Search 18	Created Thu, 21 Jan 2021 15:15:18 GMT	Common Name viral testing	
Search 17	Created Thu, 31 Dec 2020 21:13:01 GMT	Common Name management consultant	
Search 16	Created Thu, 31 Dec 2020 21:07:17 GMT	Common Name microscope	
Search 15	Created Thu, 31 Dec 2020 14:37:57 GMT	Common Name blood pressure monitor	
Search 14	Created Wed, 23 Dec 2020 14:53:16 GMT	Common Name server	
Search 13	Created Wed, 23 Dec 2020 14:50:58 GMT	Please open this search to begin.	
Search 12	Created Tue, 22 Dec 2020 20:02:13 GMT	Common Name financial analyst	
Search 11	Created Thu, 17 Dec 2020 15:44:22 GMT	Common Name Developer	
Search 10	Created Thu, 17 Dec 2020 15:33:52 GMT	Common Name management consulting	
Search 9	Created Thu, 10 Dec 2020 21:20:16 GMT	Common Name sharps disposal containers	
Search 8	Created Thu, 10 Dec 2020 20:46:13 GMT	Common Name hazardous bags	
Search 7	Created Thu, 10 Dec 2020 15:12:06 GMT	Common Name blood gas monitor	
Search 6	Created Wed, 09 Dec 2020 18:41:50 GMT	Common Name pediatrician	
Search 5	Created Tue, 08 Dec 2020 18:48:08 GMT	Please open this search to begin.	
Search 4	Created Tue, 08 Dec 2020 15:24:34 GMT	Common Name Research Assistant	
Search 3	Created Tue, 08 Dec 2020 14:15:59 GMT	Common Name hypodermic syringes	
Search 2	Created Mon, 23 Nov 2020 17:58:36 GMT	Please open this search to begin.	

*Figure 6: BUYSMARTER UI Example*

The FCS process flow, highlighted in the following graphic, walks users through each step to ‘narrow the search funnel’ to direct the AI engine on where to look and what it should find.



**Figure 7: BUYSMARTER UI Example 2**

The search screens have been designed with a simple look and feel where the user is led through each step with plain English questions and examples. As the user moves through each step of the sequence, the tool begins to pre-populate what fits within that search parameter. The more search parameters the user inputs into each step will help the AI tool to ‘hone-in’ on exactly what is being sought. The Output tool, shown in this Figure 8, displays all the key variables that the teams will need to master throughout the acquisition process.

The results are in. Let's make some refinements. Give any number of results a thumbs up or down to update the recommendations									
Actions	Agency	Sub-Agency	Vendor	Title	Date Sent	Date Due	User Profile	Total Price	Quantity
	National Institutes of Health	NIH/NIM		Ultramicroscope declustered rapidly imaging large samples chemically cleaned many basic science instruments. ultramicroscope declustered rapidly imaging large samples chemically cleaned many basic science instruments. magnetic field surfaces performed near center six	08/07/18	08/07/18			0
	National Institutes of Health	NIH/NCI ACQUISITIONS		Radiometer microscope optics overcome problem live blood cells stained green/red counting gels	08/01/18	08/08/18			0
	National Institutes of Health	NIH/NIDR		confocal microscope	08/03/18	08/04/18			0
	National Institutes of Health	NIH/NCI ACQUISITIONS		and observer etl microscope components d-l-l shutters	08/03/18	08/13/18			0
	National Institutes of Health	NIH/NCATS		electron microscope equipment chilled water measure flow	12/03/17	09/16/18			0
	National Institutes of Health	NIH/NCI ACQUISITIONS		electron microscope lens building 35 norme option 1	08/01/18	08/09/17			0
	National Institutes of Health	NIH/NCATS		fluorescence microscope based grid ground state depolarization	08/03/18	04/28/19			0
	National Institutes of Health	NIH/NCI ACQUISITIONS		private pr office dr lu basic microscope room	02/28/18	09/25/18			0
	National Institutes of Health	NIH/NCI ACQUISITIONS		residual length inverted brightfield fluorescent microscope auto fluorescence	05/04/18	05/02/18			0.1
	National Institutes of Health	NIH/NCI ACQUISITIONS		specimen temperature control optimized individual microscope frame vci	12/29/17	07/31/17			0
	National Institutes of Health	NIH/NIDR		spectral microscope imaging specimens labeled multiple different types	03/01/18	03/17/18			0

**Figure 8: BUYSMARTER UI Example 3**

The number and type of variables is based on the category and type of product, equipment, or service being targeted. The actual contract language can be viewed, as well. Users can Sort, Hide, and Filter any column just like in Excel to target the data of interest. Each time the ‘Thumbs-Up/Thumb’s-Down’ flagging is complete, the AI tool re-runs and updates the results as it further ‘learns’.

Full Contract Scan AI Tool / Search 16 Results

The results are in. Let's make some refinements.

Give any number of results a thumbs up or down to update the recommendations

Actions	Agency	Sub-Agency	Vendor	Title	Date End	Date Start	Unit Price	Total Price	Quantity	Desire Terms/Acceptable Quality	Color	Length of Warranty/License/Merit
	National Institutes of Health	NIMH		illumination microscope dedicated rapidly imaging large samples chemically cleaned many basic science applications	06/07/19	06/07/19	0		0			
	National Institutes of Health	OD (BIO) DEDICATED	CA (FRC) ACQUISITIONS	microscope without stage dslr ambient studio as an magnetic field survey performed new center as	06/01/19	06/29/17						
	National Institutes of Health	NIDDK		bioluminescence optics overcome problems	06/01/19	06/08/14						
	National Institutes of Health	NIMH		confocal microscope	05/31/18	05/04/18						
	National Institutes of Health	NIHDK		bio observer d1 microscope components tif fil filters	06/26/19	06/13/16						
	National Institutes of Health	OD (BIO) DEDICATED	CA (FRC) ACQUISITIONS	electron microscope equipment cooled water measure cheer type	12/30/17	09/18/15						
	National Institutes of Health	OD (BIO) DEDICATED	CA (FRC) ACQUISITIONS	electron microscope lab building 35 room option 1	06/01/19	06/29/17						
	National Institutes of Health	NIHES		bioluminescence microscope based grid ground state objective electron	06/16/19	04/28/16						
	National Institutes of Health	OD (BIO) DEDICATED	CA (FRC) ACQUISITIONS	private office or lu better microscopy room	05/28/19	09/20/16						
	National Institutes of Health	NIHES		inverted upright inverted brightfield fluorescent microscope ebs laboratories	06/04/19	04/02/18						
	National Institutes of Health	OD (BIO) DEDICATED	CA (FRC) ACQUISITIONS	spontaneous temperature control optimized individual microscope frame cost	12/26/17	07/01/17						
	National Institutes of Health	NINDS		spatial microscopy imaging specimens handled multiple offsite days	03/15/19	03/13/15						
	National Institutes of Health	NIBIB, NIAAA, SA (FRC) ACQUISITIONS		background/procurement history avoidance procure phase	06/16/19	04/23/17						
	National Institutes of Health	OD (BIO) DEDICATED	CA (FRC) ACQUISITIONS	electron microscope lab building 35 m 6	06/31/18	06/29/17	0		0			

**Figure 9: BUYSMARTER UI Example 4**

Finally, the AI tool calculates the metrics that would be key for making enterprise buying decisions as shown below:

Metrics	
Name	Value
Total Spend	\$224,968.75
Count of Contracts	71
High Price	\$112,500.00
Low Price	\$10.39
High/Low Price Variance in \$	\$112,489.61
High/Low Price Variance in %	1,082,771.90%
Average Price	\$9,015.12
Median Price	\$44.25
Retail Price [if available]	None available
Potential Cost Avoidance if All Purchased at Low Price	\$178,844.57

**Figure 10:** BUYSMARTER UI Example 5

American Council for Technology-Industry Advisory Council (ACT-IAC)  
3040 Williams Drive, Suite 500, Fairfax, VA 22031  
[www.actiac.org](http://www.actiac.org) • (p) (703) 208.4800 • (f) (703) 208.4805

## Constraints

This AI based solution was designed and built from the ground-up by piecing together multiple AI tools that all had to work together to produce the desired results. HHS leaders had to be willing to the ‘develop fast, fail fast, and learn’ model with daily interactions with the AI vendors to mold the tool into what HHS needed. By being the ‘first in kind’ AI tool, HHS had no rulebook or recipe to follow, so the development of the tool took longer than desired. Also, developing a simplified and automated process for ingesting the data from the five contract writing systems would greatly improve the challenges of data loading and structuring.

## Transformation Enablers

The most key enabler has been unwavering leadership support, including the visionary leadership of the Office of Acquisition, the Heads of Contracting Activity from the OpDivs/StaffDivs, and the Re-Imagine HHS Transformation Management Office. Going forward, BUYSMARTER must really focus on the value delivered, in terms of “cost savings” in addition to clearly articulating the “why” behind continued investment will prove to be critical.

## Stakeholders

Stakeholders vary significantly across the HHS Operating Divisions and Agencies. However, primary stakeholders include the Heads of Contracting Activity, acquisition workforce, and acquisition systems developers and maintainers (to include the maintainers of legacy systems). Also, the cost avoidances that these large enterprise acquisitions will realize will help the OpDivs and StaffDivs better deliver their mission by keeping more of their funds focused on mission delivery.

## Business Process Model

The entire enterprise acquisition process follows the HHS *BUYSMARTER* Operating Model, which mirrors the Healthcare Industry’s Group Purchasing Organization model for leveraging the buying power of member groups of hospitals and non-acute care sites.

## Organizational Readiness

HHS Deputy Secretary Eric Hargan signed the *BUYSMARTER* Program Proclamation on January 14, 2020, formally establishing *BUYSMARTER* as an official program with the Office of Acquisitions.

## Technology Selection

Full suite of 40+ AI Open Source Code Tools brought together to achieve the mission.

## Lessons Learned

AI, by its very nature, is a new and customized capability when applied to any new use case. The best lesson was constant developer/customer interactions. Developing in a model that is ‘more agile than Agile’, in that, the full team meets every day to review progress on development, data, and testing. It has become affectionately known as ‘Full Contact Development’ and is the key to our success.

## Conclusion

### ***People produce permanent processes***

No matter what modernization effort is being undertaken, the adoption of emerging technologies is often a difficult task beyond the technical aspects of the initiative. Technical issues aside, buy-in is critical for emerging technologies to be successful as the workforce needs to be able to lead this change and leverage the capabilities of these technologies to transform government.

For example, the Defense Logistics Agency (DLA) was able to successfully leverage the use of RPA by overcoming workforce reluctance about how these technologies affected their positions. The DLA RPA team undertook a comprehensive strategy to communicate with workers that not only were there not going to be reductions in jobs, but they were also going to be able to do more interesting work — and that the bots would be able to be a force multiplier against empty positions that were never going to be filled due to budgetary reasons. These efforts would enable the department to get more strategic work done, the epitome of the “Shifting from Low-Value to High-Value Work” CAP goal.<sup>22</sup>

However, with the possibilities that these emerging technologies bring especially the promise of AI, the federal workforce is not ready to execute these solutions. According to a 2019 Accenture survey:<sup>23</sup>

61 percent of federal workers said they haven't been adequately trained to work alongside AI technologies. According to the new research, while 85% of federal executives believe that 'collaboration between humans and machines will be critical to innovation in the future,' only 18% said they were currently preparing their workforce to interact with collaborative, interactive and explainable AI-based systems.

Reskilling the workforce to implement emerging technology solutions is required. Achieving this training goal begins with federal leaders engaging their workforces early to provide a vision on the outcomes to be achieved by the technology and the requirements of the workforce to successfully achieve these objectives. Ultimately, federal leaders need to remove barriers to make it easier for the workforce to obtain the skills they need to be successful. Several areas include:

- Designing a more user-friendly learning environment to increase convenience and opportunities for the federal workforce that accelerates training. This area focuses on time and easing the overburdened workforce to prioritize training and incentives life-learning beyond the heavy compliance of daily operations.
- Creating organizational cultures that make employee engagement and knowledge transfer easier, empower communications, and improve collaboration to create positive changes that people both feel and grow.

- Reevaluating the way training programs are developed and delivered, as well as understanding how agencies need to train adult learners to grasp new topics and skills. For example, Beth Killoran, GSA's Deputy Chief Information Officer, is considering how short, YouTube-style videos and podcasts could be used as training tools for employees to learn at their own pace in digestible chunks.<sup>24</sup> In addition, agencies can leverage online educational platforms like LinkedIn Learning.<sup>25</sup>

Through a concerted effort by federal leaders to focus on the workforce versus the technology, the implementation of these capabilities will be successful across the federal government.

### **A starting point for today's practitioners**

One of the advantages of these technologies is the relative ease of getting started with their use. Today's practitioners have numerous opportunities to begin understanding these tools through product demonstrations, either through verifying general capabilities, or better yet, pre-coordinated demonstration that allows a specific issue to be addressed through a more targeted demonstration. Nonetheless, the goal of using these technologies should be to integrate emerging technologies into a larger digital transformation and modernization strategy rather than approaching these projects piecemeal. While using them this way may be an effective strategy in the short term, these technologies are not suitable in all situations.

One of the most important planning steps needed before undertaking any emerging technology initiative is to understand the outcomes of using the technology and the problem needing to be solved. The technology is not a be-all-end-all and this planning and understanding helps organizations ensure they are using these tools effectively and not misapplying it to processes when another solution would be a better fit. Considerations for practitioners to get started with emerging technologies:<sup>26</sup>

- One of the most important starting points for emerging technology initiatives is to avoid focusing on large scale projects. By looking for "quick wins" through smaller projects and lower risk opportunities, measurable return on investment is possible to allow for momentum to grow towards larger implementations.
- Fit the problem to the solution to increase opportunities for successful projects. For example, processes that would be suitable for RPA implementations are rules-based and repetitious. Work with stakeholders on process automation workshops to assess processes and choose your starting point.
- Develop business cases to prioritize which processes to automate and analyze the best opportunities for value and return on investment. Focus on the value of increasing efficiency and accuracy of the tasks looking to be automated, and cost savings as a result
- Determine the right operating model for hosting the technology, since the technologies can be executed in the same ways as any other software. This analysis would be look at models such as internally hosted data centers, or via cloud, for example.

- Having the right partners, or mentors, can be the difference between success and failure.  
As stated by Prem Jadhwan, Chief Technology Officer, of Government Acquisitions, Inc.:  
<sup>27</sup>

"Mentors should be more than a sounding board or advisor – they can also be valuable partners, guiding and helping you throughout the AI solution lifecycle – from inception to mission outcome. Think of that when choosing an integrator or solutions provider; getting into AI is not like buying a computer, but more like adopting a new way of doing business. You'll get further faster if your partner has done that before."

### ***Next steps for a team new to these technologies***

An important factor for federal agencies is to focus on understanding how these technology solutions impact the priorities of the agency. Agencies should research how others in the public sector have implemented these solutions under consideration, since these initiatives have broad exposure across the federal government. Research should also be conducted with vendors and other industry experts to see what capabilities exist that can be implemented. It can be tempting to jump in headfirst when a solution promises to push the agency toward meeting a certain goal. Adoption of new technology should be well-conceived and methodical for it to be successful.

In addition to a well-thought-out implementation process, make sure the team is supported and setup for success. This includes training so that the workforce can get the most value out of new technologies and can best leverage the capability. Training should be both ongoing and integrated into operations.

This training focus should be two-phased. Phase one would constitute more traditional platforms and mediums such as expert consultants, the technology vendor, etc. Phase two then pivots to focus on training-the-trainer and identifying employees who are comfortable using the new solution and encourage them to both train other employees and be coaches to others when questions arise.<sup>28</sup> Ultimately, practitioners will just have to dig in and try new things. Agencies can and should experiment with single-point solutions and not get locked into a technology that cannot be easily discontinued if it is not working or solving a problem. Getting real people using it as soon as possible is the best way to realize the potential of an emerging technology.

## Authors

This paper was written by a consortium of government and industry representatives. The organizational affiliations of these contributors are included for information purposes only. The views expressed in this document do not necessarily represent the official views of the individuals and organizations that participated in its development.

Jaime Gracia	United States Government (Project Lead)
Mike Rice	CornerStone IT (Project Lead)
Polly Hall	United States Government
Janelle Billingslea	United States Government
Marisa Roinestad	United States Government
David Dastvar	Eagle Tech, Inc.
David Hernandez	Riva Solutions, Inc.
Ken Thomson	Unissant, Inc.
Ali Loveys	BT Block Health Group

## References

---

- <sup>1</sup> Federal Acquisition Regulation, 1.102(a), [https://www.acquisition.gov/far/part-1#FAR\\_1\\_102](https://www.acquisition.gov/far/part-1#FAR_1_102)
- <sup>2</sup> Federal Acquisition Regulation, 1.102-4(e), <https://www.acquisition.gov/far/1.102-4>
- <sup>3</sup> Office of Federal Procurement Policy website: <https://www.whitehouse.gov/omb/management/office-federal-procurement-policy/>
- <sup>4</sup> Mitchell, Billy. (MAR 29, 2019). Acquisition is ‘ripe for emerging technology,’ says OFPP’s Newhart’. fedscoop. <https://www.fedscoop.com/emerging-technology-ofpp-joanie-newhart/>
- <sup>5</sup> President’s Management Agenda, <https://www.whitehouse.gov/omb/management/pma/>
- <sup>6</sup> Cross Agency Priorities, <https://www.performance.gov/CAP/overview/>
- <sup>7</sup> Ibid
- <sup>8</sup> CAP Goal Shifting From Low-Value to High-Value Work , <https://www.performance.gov/CAP/low-value-to-high-value-work/>
- <sup>9</sup> CAP Goal Frictionless Acquisition, <https://www.performance.gov/CAP/frictionless-acquisition/>
- <sup>10</sup> Ibid
- <sup>11</sup> Gansler, Jacques S., Greenwalt, William C. and Lucyshyn, William. (November, 2013) Non-Traditional Commercial Defense Contractors. University of Maryland, School of Public Policy, Center for Public Policy and Private Enterprise. <https://apps.dtic.mil/dtic/tr/fulltext/u2/a613239.pdf>
- <sup>12</sup> Kumar, Vivek (Jun 25, 2018). The Need for Data Infrastructure for Best Utilization of Artificial Intelligence. towards data science. <https://towardsdatascience.com/the-need-for-data-infrastructure-for-best-utilization-of-artificial-intelligence-72612c1026e0>
- <sup>13</sup> Ibid
- <sup>14</sup> DATA Act Information Model Schema (DAIMS) <https://fiscal.treasury.gov/data-transparency/DAIMS-current.html>
- <sup>15</sup> Fiscal Year 2019 Digital Accountability and Transparency Act Reporting Compliance, November 7, 2019, Reference Number: 2020-10-003. <https://www.treasury.gov/tigta/auditreports/2020reports/202010003fr.pdf>
- <sup>16</sup> About Pilot IRS: <https://www.irs.gov/about-irs/procurement/about-pilot-irs>
- <sup>17</sup> Boyd, Aaron (August 5, 2019). IRS Doesn’t Have An R&D Shop, So It Built A \$7M Procurement Vehicle Instead. Nextgov. <https://www.nextgov.com/emerging-tech/2019/08/irs-doesnt-have-rd-shop-so-it-built-7m-procurement-vehicle-instead/158952/>
- <sup>18</sup> Office of Federal Procurement Policy (JUN 14, 2012). Contracting Guidance to Support Modular Development. <https://obamawhitehouse.archives.gov/sites/default/files/omb/procurement/guidance/modular-approaches-for-information-technology.pdf>

<sup>19</sup> Periodic Table of Acquisition Innovations, Federal Acquisition Institute. [https://www.fai.gov/periodic-table/pdfs/Pilot\\_IRS.pdf](https://www.fai.gov/periodic-table/pdfs/Pilot_IRS.pdf)

<sup>20</sup> Department of Homeland Security, Commercial Solutions Opening Pilot Program Guide, Office of The Chief Procurement Officer (JUL 17, 2019).

[https://www.dhs.gov/sites/default/files/publications/commercial\\_solutions\\_opening\\_pilot\\_program\\_guide.pdf](https://www.dhs.gov/sites/default/files/publications/commercial_solutions_opening_pilot_program_guide.pdf)

<sup>21</sup> HHS BUYSMARTER website: <https://www.hhs.gov/grants/contracts/buysmarter/journey-to-program/index.html>

<sup>22</sup> Ibid

<sup>23</sup> Accenture Federal Services (February 15, 2019). Federal workers ready to thrive in the age of AI. <https://www.accenture.com/us-en/insights/us-federal-government/ready-thrive-ai?src=SOMS>

<sup>24</sup> Blake, Johnson Nicole (January 30, 2020). Are Government Employees Ready for Ai? Govloop. <https://www.govloop.com/are-government-employees-ready-for-ai/>

<sup>25</sup> LinkedIn Learning - Overview. <https://www.linkedin.com/help/learning/answer/71918/linkedin-learning-overview?lang=en>

<sup>26</sup> Greer, Kevin. (August 8, 2018). Getting started with robotic process automation in government is easier than you think. CGI Blog. <https://www.cgi.com/us/en-us/blog/getting-started-with-robotic-process-automation-in-government-is-easier-than-you-think>

<sup>27</sup> Jadhwanı, Prem. 7 Steps for Getting Started in AI. Government Executive Leadership Voices. <https://www.govexec.com/media/sponsored/leadershipvoices/seven-steps-for-getting-started-in-ai/index.html#article>

<sup>28</sup> Power, Rhett. (Jul 14, 2019). 3 Best Practices for Incorporating Emerging Tech. Forbes. <https://www.forbes.com/sites/rhettpower/2019/07/14/3-best-practices-for-incorporating-emerging-tech/?sh=1d8f83287256>

## Confronting Bias: BSA's Framework to Build Trust in AI

## CONTENTS

<b>Introduction</b> . . . . .	1
<b>What Is AI Bias?</b> . . . . .	3
Sources and Types of AI Bias . . . . .	4
<b>The Need for AI Risk Management</b> . . . . .	8
What Is Risk Management? . . . . .	8
Managing the Risk of Bias . . . . .	9
<b>Foundations for Effective Risk Management</b> . . . . .	10
Governance Framework . . . . .	11
Impact Assessment . . . . .	13
<b>AI Bias Risk Management Framework</b> . . . . .	14
AI Lifecycle Phases . . . . .	15
Framework Structure . . . . .	17
Stakeholder Roles and Responsibilities . . . . .	18
Spectrum of AI Development and Deployment Models . . . . .	18
<b>BSA AI Bias Risk Management Framework</b> . . . . .	19
<b>Foundational Resources</b> . . . . .	28
<b>Endnotes</b> . . . . .	29

# Introduction

Tremendous advances in artificial intelligence (AI) research and development are quickly transforming expectations about how the technology may shape the world. The promise that AI may one day impact every industry is quickly turning into a commercial reality. From financial services to healthcare, AI is increasingly leveraged to improve customer experiences, enhance competitiveness, and solve previously intractable problems. For instance, AI is enabling medical researchers to diagnose early-stage Alzheimer's Disease years before debilitating symptoms arise,<sup>1</sup> and it is helping ecologists analyze impossibly large datasets to better track the impact of their efforts to preserve critical habitat and prevent illegal elephant poaching in Malawi.<sup>2</sup>

As used in this report, the term "artificial intelligence" refers to systems that use machine learning algorithms that can analyze large volumes of training data to identify correlations, patterns, and other metadata that can be used to develop a model that can make predictions or recommendations based on future data inputs. For example, developers used machine learning to create "Seeing AI," an app that helps people who are blind or visually impaired navigate the world by providing auditory descriptions of objects in photographs.<sup>3</sup> Users of the app can use their smartphone to take pictures, and Seeing AI describes what appears in the photograph. To develop the computer vision model capable of identifying the objects in a picture, the system was trained using data from millions of publicly available images depicting common objects, such as trees, street signs, landscapes, and animals. When a user inputs a new image, Seeing AI in effect predicts what objects are in the photo by comparing it to the patterns and correlations that it derived from the training data.

**The proliferation of AI across industries is also prompting questions about the design and use of the technology and what steps can be taken to ensure it is operating in a manner that accounts for any potential risks it may pose to the public.**

The use of advanced technologies in connection with high-stakes decisions presents both opportunities and risks. On the one hand, the adoption of AI by financial institutions has the potential to reduce discrimination and promote fairness by facilitating a data-driven approach to decision-making that is less vulnerable to human biases.<sup>4</sup> For instance, the use of AI can improve access to credit and housing to historically marginalized communities by enabling lenders to evaluate a greater array of data than is ordinarily accounted for in traditional credit reports. At the same time, researchers caution that flaws in the design, development, and/or deployment of AI systems have the potential to perpetuate (or even exacerbate) existing societal biases.<sup>5</sup>

Developing mechanisms for identifying and mitigating the risks of AI bias has therefore emerged as an area of intense focus for experts in industry, academia, and government. In just the past few years, a vast body of research has identified a range of organizational best practices, governance safeguards, and technical tools that can help manage the risks of bias throughout the AI lifecycle. Static evaluations of AI models cannot account for all potential issues that may arise when AI systems are deployed in the field, so experts agree that mitigating risks of AI bias requires a lifecycle approach that includes ongoing monitoring by end-users to ensure that the system is operating as intended.

**This document sets forth an AI Bias Risk Management Framework that organizations can use to perform impact assessments to identify and mitigate potential risks of bias that may emerge throughout an AI system's lifecycle.** Similar to impact assessments for data privacy, AI impact assessments can serve as an important assurance mechanism that promotes

accountability and enhances trust that high-risk AI systems have been designed, developed, tested, and deployed with sufficient protections in place to mitigate the risk of harm. AI impact assessments are also an important transparency mechanism that enables the many potential stakeholders involved in the design, development, and deployment of an AI system to communicate about its risks and ensure that responsibilities for mitigating those risks are clearly understood.

**In addition to setting forth a process for performing an AI impact assessment, the Bias Risk Management Framework:**

- Sets out the key corporate governance structures, processes, and safeguards that are needed to implement and support an effective AI risk management program; and
- Identifies existing best practices, technical tools, and resources that stakeholders can use to mitigate specific AI bias risks that can emerge throughout an AI system's lifecycle.

⋮

**This Framework is intended to be a flexible tool that organizations can use to enhance trust in their AI systems through risk management processes that promote fairness, transparency, and accountability.**

# What Is AI Bias?

References to “AI bias” in this document refer to AI systems that systematically and unjustifiably yield less favorable, unfair, or harmful outcomes to members of specific demographic groups.

At its core, the goal of machine learning is to create a model that derives generalized rules from historical examples in order to make predictions about future data inputs. For instance, an image recognition system designed to identify plants would likely be trained on large volumes of photographs depicting each of the many species of vegetation. The system would look for general rules, like leaf patterns, that are common across the photographs of each species, thereby creating a model that can evaluate whether new data inputs (i.e., user-submitted photos) include any of the species it has been trained to identify. In other words, machine learning

works by drawing generalizations from past data to make predictions about future data inputs. However, when AI is used to model human behavior, concerns about unintended bias take on an entirely different dimension. As AI is integrated into business processes that can have consequential impacts on people’s lives, there is a risk that “biased” systems will systematically disadvantage members of historically marginalized communities. AI bias can manifest in systems that perform less accurately or treat people less favorably based on a sensitive characteristic, including but not limited to race, gender identity, sexual orientation, age, religion, or disability.

## Sources and Types of AI Bias



### DESIGN

AI bias can be introduced at multiple stages in the AI lifecycle.<sup>6</sup> Decisions made at the earliest stages of the conception and design of an AI system can introduce bias:

- **Problem Formulation Bias.** In some instances, the basic assumptions underlying a proposed AI system may be so inherently biased that they render it inappropriate for any form of public deployment.

#### EXAMPLES

In 2016, researchers at Shanghai Jiao Tong University published a highly controversial paper<sup>7</sup> detailing their effort to train an AI system to predict “criminality” through a facial imaging system. By training the system on a large volume of police mugshots, the researchers alleged that their system could predict “criminality” with close to 90 percent accuracy merely by analyzing a person’s facial structure. Unsurprisingly, the paper quickly became the subject of scathing criticism, and commentators rightfully noted that the model relied on the profoundly disturbing (and causally unsupportable) assumption that criminality can be inferred from a person’s appearance.<sup>8</sup>

.....

Problem formulation bias can also arise when an AI system’s target variable is an imprecise or overly simplistic proxy for what the system is actually trying to predict. For example, in 2019 researchers discovered that an AI system widely used by hospitals to triage patients<sup>9</sup> by predicting the likelihood that they required urgent care systematically prioritized the needs of healthier white patients to the detriment of less-healthy minority patients. In this instance, bias arose because the system sought to predict “healthcare needs” using historical data about “healthcare costs” as an easy-to-obtain stand-in for the actual data about the healthcare needs of patients. Unfortunately, because minority patients have historically had less access to healthcare, using “healthcare costs” as a proxy for the current needs of those patients paints an inaccurate picture that can result in dangerously biased outcomes.

- **Historical Bias.** There is a risk of perpetuating historical biases reflected in data used to train an AI system.

#### EXAMPLE

A medical school in the United Kingdom set out to create a system that would help identify good candidates for admission. The system was trained using data about previously admitted students. It was discovered, however, that the school's historical admissions decisions had systematically disfavored racial minorities and females whose credentials were otherwise equal to other applicants. By training the model using data reflecting historical biases, the medical school inadvertently created a system that replicated those same biased admission patterns.<sup>10</sup>

- **Sampling Bias.** If the data used to train a system is misrepresentative of the population in which it will be used, there is a risk that the system will perform less effectively on communities that may have been underrepresented in the training data. This commonly occurs when sufficient quantities of representative data are not readily available, or when data is selected or collected in ways that systematically over- or under-represent certain populations.

#### EXAMPLES

As the pathbreaking research by Joy Buolamwini and Timnit Gebru demonstrated, facial recognition systems trained on datasets composed disproportionately of white and male faces perform substantially less accurately when evaluating the faces of women with darker complexions.<sup>11</sup>

.....

Sampling bias can also arise as a result of data collection practices. The City of Boston's attempt to create a system capable of automatically detecting and reporting potholes in need of repair is an illustrative case in point. Because early versions of the program relied heavily on data supplied by users of a smartphone app called "StreetBump," it received a disproportionate number of reports from affluent neighborhoods with residents who could afford smartphones and data plans. As a result of the sampling bias, potholes in poorer neighborhoods were underrepresented in the dataset, creating a risk that the system would allocate repair resources in a manner that would treat members of those communities unfairly.<sup>12</sup>

- **Labeling Bias.** Many AI systems require training data to be “labeled” so that the learning algorithm can identify patterns and correlations that can be used to classify future data inputs. The process of labeling the training dataset can involve subjective decisions that can be a vector for introducing human biases into the AI system.

**EXAMPLE**

ImageNet is a database of more than 14 million images that have been categorized and labeled to enable AI researchers to train vision recognition systems. Although ImageNet has been a critical tool for advancing the state of the art in AI object recognition, recent scholarship has shone a light on how the database’s categorization and labeling system can create significant risks of bias when it is used to train systems involving images of people. In *Excavating AI*,<sup>13</sup> Kate Crawford and Trevor Paglen demonstrated that the categories and data labels associated with the images of people in ImageNet reflect a range of “gendered, racialized, ableist, and ageist” biases that could be propagated in any AI system that uses them as training data. For instance, an AI system trained on ImageNet data was more likely to classify images of Black subjects as “wrongdoers” or “offenders.”<sup>14</sup>



## DEVELOPMENT

Once the necessary data has been collected, the development team must clean, process, and normalize the data so that it can be used to train and validate a model. Developers must also select a machine learning approach, or adapt an off-the-shelf model, that is appropriate for the nature of the data they are using and the problem they are trying to solve. This may involve building many different models using different approaches and then choosing the most successful among them.<sup>15</sup> Usually, the development team must also make choices about data parameters to make the model functional. For instance, data reflecting a numerical score may be converted to a “yes” or “no” answer by assigning a threshold—for example, scores equal or greater to X may be re-designated as a “yes,” and scores below that threshold designated “no.” Biases that can emerge during the development stage include the following:

- **Proxy Bias.** The process of selecting the input variables (i.e., “features”) that the model will weigh as it is being trained is another critical decision point that can introduce bias. Even when sensitive demographic data is excluded, bias may be introduced if the system relies on features that are closely correlated to those traits, called proxies.

**EXAMPLE**

Even the use of seemingly benign features can introduce proxy bias due to their correlation with sensitive attributes. Researchers have shown, for instance, that information about whether a person owns a Mac or PC laptop may be predictive of their likelihood to pay back a loan.<sup>16</sup> A financial institution might therefore seek to include such a variable when building an AI system to screen potential loan applicants. However, the inclusion of that feature also introduces a significant risk of proxy bias because Mac ownership correlates closely to race. As a result, its inclusion could result in a system that systematically disfavors applicants based on a feature that is closely correlated to race but that is unrelated to actual credit risk.

- **Aggregation Bias.** Using a “one-size-fits-all” model that overlooks key variables can result in system performance that is optimized only for the dominant sub-group. Aggregation bias can arise if the model fails to account for underlying differences between sub-groups that materially impact a system’s accuracy rates. Rare phenomena may be lost in averages and aggregates. Worse, models of aggregated populations may correctly predict different or even opposite behavior to modes of sub-groups of the same population, a phenomenon known as Simpson’s Paradox.

#### EXAMPLE

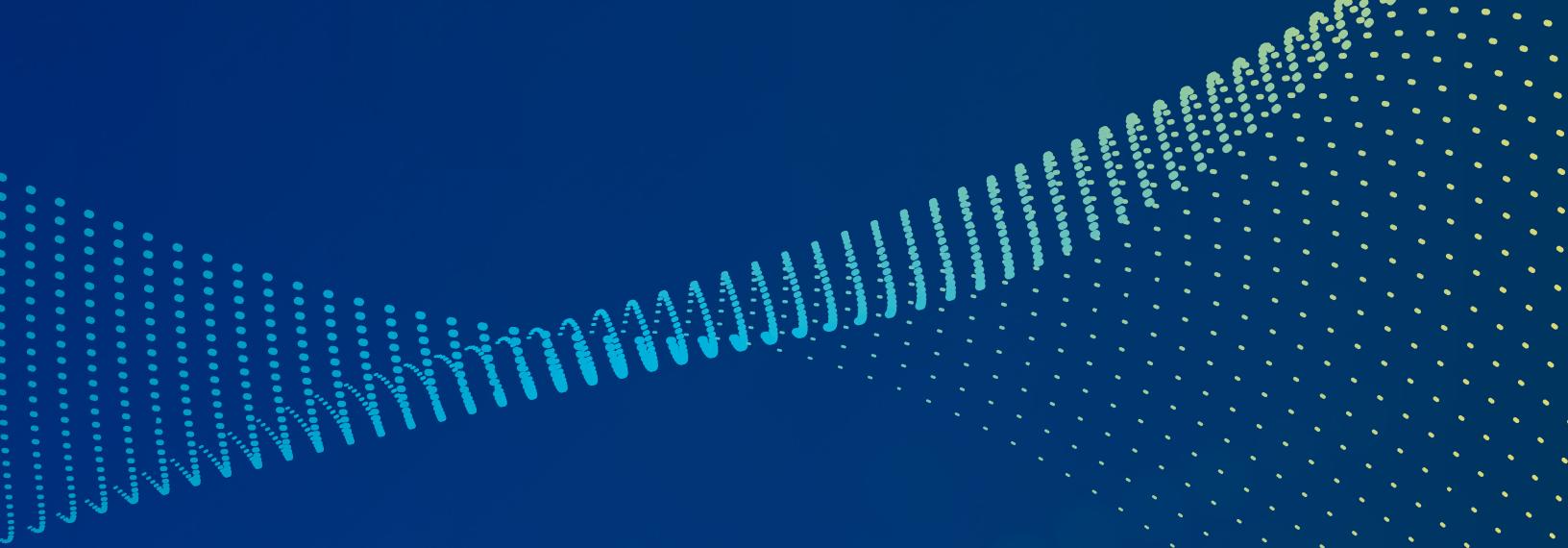
The risk of aggregation bias is particularly acute in healthcare settings where diagnosis and treatment must often account for the unique manner in which medical conditions may impact people across racial and ethnic lines. For instance, because the risk of complications posed by diabetes varies wildly across ethnicities, an AI system used to predict the risks associated with diabetes may underperform for certain patients unless it accounts for these differences.<sup>17</sup>



## DEPLOYMENT, MONITORING, AND ITERATION

AI systems inevitably encounter real world scenarios that differ from the data used to train the model. As a result, even a system that has been thoroughly validated and tested prior to deployment may suffer performance degradation when it is put into production. Therefore, it is important that AI systems undergo ongoing evaluation and assessment throughout their lifecycles.

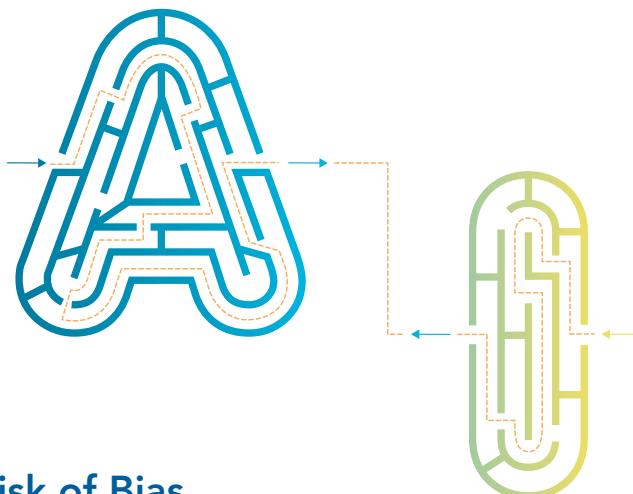
- **Deployment Bias.** Bias can arise in various ways after a system has been deployed, including when the data used to train or evaluate an AI system differs markedly from the population the system encounters when it is deployed, rendering the model unable to perform as intended. Deployment bias can emerge when a model is unable to reliably generalize beyond the data on which it was trained, either because the model was overfitted at the time of training (i.e., the prediction model learned so much detail about the training data that it is unable to make accurate generalizations about other data inputs) or because of concept drift (i.e., performance degradation was brought on by a shift in the relationship between the target variable and the training data).
- **Misuse Bias.** Deployment bias can also arise when an AI system or feature built for one purpose is used in an unexpected or unintended manner.



# The Need for AI Risk Management

## What Is Risk Management?

Risk management is a process for ensuring systems are trustworthy by design by establishing a methodology for identifying risks and mitigating their potential impact. Risk management processes are particularly important in contexts, such as cybersecurity and privacy, where the combination of quickly evolving technologies and highly dynamic threat landscapes render traditional “compliance” based approaches ineffective. Rather than evaluating a product or service against a static set of prescriptive requirements that quickly become outdated, risk management seeks to integrate compliance responsibilities into the development pipeline to help mitigate risks throughout a product or service’s lifecycle. Effective risk management is anchored around a governance framework that promotes collaboration between an organization’s development team and its compliance personnel at key points during the design, development, and deployment of a product.



## Managing the Risk of Bias

Organizations that develop and use AI systems must take steps to prevent bias from manifesting in a manner that unjustifiably yields less favorable or harmful outcomes based on someone's demographic characteristics. Effectively guarding against the harms that might arise from such bias requires a risk management approach because:

### **"BIAS" AND "FAIRNESS" ARE CONTEXTUAL**

It is impossible to eliminate bias from AI systems because there is no universally agreed upon method for evaluating whether a system is operating in a manner that is "fair." In fact, as Professor Arvind Narayanan has famously explained, there are at least 21 different definitions<sup>18</sup> (i.e., mathematical criteria) that can be used to evaluate whether a system is operating fairly, and it is *impossible* for an AI system to simultaneously satisfy all of them. Because no universal definition of fairness exists, developers must instead evaluate the nature of the system they are creating to determine which metric for evaluating bias is most appropriate for mitigating the risks that it might pose.

### **EFFORTS TO MITIGATE BIAS MAY INVOLVE TRADE-OFFS**

Interventions to mitigate bias for one group can increase it for other groups and/or reduce a system's overall accuracy.<sup>19</sup> Risk management provides a mechanism for navigating such trade-offs in a context-appropriate manner.

### **BIAS CAN ARISE POST-DEPLOYMENT**

Even if a system has been thoroughly evaluated prior to deployment, it may produce biased results if it is misused or deployed in a setting in which the demographic distribution differs from the composition of its training and testing data.

# Foundations for Effective Risk Management

The aim of risk management is to establish repeatable processes for identifying and mitigating potential risks that can arise throughout an AI system's lifecycle. A comprehensive risk management program has two key elements:

⋮

**1**  
A **governance framework** to support the organization's risk management functions.

⋮

**2**  
A scalable process for performing an **impact assessment** to identify and mitigate risks.

## Governance Framework

Effective AI risk management should be underpinned by a governance framework that establishes the policies, processes, and personnel that will be used to identify, mitigate, and document risks throughout the system's lifecycle. The purpose of such a governance framework is to promote understanding across organizational units—including product development, compliance, marketing, sales, and senior management—about each entity's role and responsibilities for promoting effective risk management during the design, development, and deployment of AI systems. Key features of a risk management governance framework include:

### Policies and Processes

At the core of the governance framework is a set of formal policies setting forth the organization's approach to risk management. These policies should define the organization's risk management objectives, the procedures that it will use to meet those objectives, and the benchmarks it will rely on for evaluating compliance.

- **Objectives.** AI risk management should be contextualized within an organization's broader risk management functions with the goal of ensuring that the organization is developing and using AI in a manner that aligns with its core values. To that end, the governance framework should identify how the organization will manage risks that could undermine those values.
- **Processes.** The governance framework should establish processes and procedures for identifying risks, assessing the materiality of those risks, and mitigating risks at each stage of the AI lifecycle.
- **Evaluation Mechanisms.** The governance framework should establish mechanisms, such as metrics and benchmarks, that the organization will use to evaluate whether policies and procedures are being carried out as specified.
- **Periodic Review.** As AI capabilities continue to mature and the technology is put to new uses, it is important that organizations periodically review and update their AI governance framework so that it remains fit-for-purpose and capable of addressing the evolving landscape of risk.



**Executive Oversight.** AI Developers and AI Deployers should maintain a governance framework that is backed by sufficient executive oversight. In addition to developing and approving the substance of the governance framework's policies, senior management should play an active role in overseeing the company's AI product development lifecycle. For high-risk systems that may negatively impact people in consequential ways, company leadership should be accountable for making "go/no-go" decisions.

## Personnel, Roles, and Responsibilities

The effectiveness of risk management depends on establishing a cross-functional group of experts that can guide decisions throughout the AI lifecycle. Depending on the size of an organization and the nature of the systems it is developing or deploying, the responsibilities for risk management may involve staff from multiple business units. The governance framework should therefore identify the personnel within the organization who have roles and responsibilities related to AI risk management and clearly map reporting lines, authorities, and necessary expertise. In assigning roles and responsibilities, organizations should prioritize independence, competence, influence, and diversity.

- **Independence.** Risk management is most effective when personnel are structured in a manner that facilitates separate layers of independent review. For instance, risk management responsibilities may be split between multiple teams, including:
  - **Product Development Team.** Engineers, data scientists, and domain experts involved in designing and developing AI products and services.
  - **Compliance Team.** A diverse team of legal, compliance, domain experts, and data professionals who are responsible for overseeing compliance with the company's AI development policies and practices, such as the development of impact assessments for high-risk AI systems.
  - **Governance Team.** Ideally a senior management-led team with responsibility for developing, maintaining, and ensuring effective oversight of the organization's AI Governance Framework and risk management processes.
- **Competence, Resourcing, and Influence.** Personnel with risk management responsibilities must be provided with adequate training and resources to fulfill their governance functions. It is equally important to ensure that personnel are empowered and have the right incentives to make decisions to address and/or escalate risks. For instance, the organization should establish a clear escalation path that enables risk management personnel to engage with executive decision-makers so that there is executive-level visibility into key risk areas and decisions.



**Diversity.** The sociotechnical nature of AI systems makes it vitally important to prioritize diversity within the teams involved in a system's development and oversight. Development and oversight processes are most effective when team members bring diverse perspectives and backgrounds that can help anticipate the needs and concerns of users who may be impacted by or interact with an AI system. Because "algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values," it is vital that organizations establish teams that reflect a diversity of lived experiences and that traditionally underrepresented perspectives are included throughout the lifecycle of the AI design and development process.<sup>20</sup> To the extent an organization is lacking in diversity, it should consult with outside stakeholders to solicit feedback, particularly from underrepresented groups that may be impacted by the system.

## Impact Assessment

To effectively manage AI risks, organizations should implement a robust process for performing impact assessments on any system that may materially impact members of the public. Impact assessments are widely used in a range of other fields—from environmental protection to data protection—as an accountability mechanism that promotes trust by demonstrating that a system has been designed in a manner that accounts for the potential risks it may pose to the public. In short, the purpose of an impact assessment is to identify the risks that a system may pose, quantify the degree of harm the system could generate, and document any steps that have been taken to mitigate those risks to an acceptable level.

Impact assessment processes should be tailored to address the nature of the system that is being evaluated and the type of harms it may pose. For truly low-risk systems—for example, a system used to predict the type of fonts being used on a document—a full impact assessment may not be necessary. But for systems that pose an inherent risk of material harm to the public, a full impact assessment should be performed. Given the incredible range of applications to which AI can be applied, there is no “one-size-fits-all” approach for identifying and mitigating risks. Instead, impact assessment processes should be tailored to address the nature of an AI system and the type of inherent risks and potential harms it may pose. To determine whether a system poses an inherent risk of material harm, stakeholders should consider:

- **Potential Impact on People.** Impact assessments are likewise important in circumstances where an AI system will be used in decision-making processes that may result in consequential impacts on people, such as their ability to obtain access to credit or housing.
- **Context and Purpose of the System.** Evaluating the nature of the AI system and the setting in which it will be used is a good starting point for determining both the necessity and appropriate scope of an impact assessment. Impact assessments are particularly critical for high-risk AI systems that will be used in domains (e.g., healthcare, transportation, finance) where the severity and/or likelihood of potential harms is high.
- **Degree of Human Oversight.** The degree to which an AI system is fully automated may also impact the inherent risks that it poses. A system designed to provide recommendations to a highly skilled professional is likely to pose fewer inherent risks than a similarly situated fully automated system. Of course, the mere existence of a human-in-the-loop certainly does not mean that an AI system is free from risk. It is necessary instead to examine the nature of the human-computer interaction holistically to determine the extent to which human oversight may mitigate an AI system’s inherent risks.
- **Type of Data.** The nature of the data used to train a system can also shed light on a system’s inherent risks. For instance, using training data relating to human characteristics or behaviors is a signal that a system may require closer scrutiny for bias.



# AI Bias Risk Management Framework

We outline below an AI Bias Risk Management Framework that is intended to aid organizations in performing impact assessments on systems with potential risks of AI bias. In addition to setting forth processes for identifying the sources of bias that can arise throughout an AI system's lifecycle, the Framework identifies best practices that can be used to mitigate those risks.

**The Framework is an assurance-based accountability mechanism that can be used by AI Developer and AI Deployer organizations for purposes of:**

- **Internal Process Guidance.** AI Developers and AI Deployers can use the Framework as a tool for organizing and establishing roles, responsibilities, and expectations for internal processes.
- **Training, Awareness, and Education.** AI Developers and AI Deployers can use the Framework to build internal training and education programs for employees involved in developing and using AI systems. In addition, the Framework may provide a useful tool for educating executives about the organization's approach to managing AI bias risks.
- **Assurance and Accountability.** AI Developers and AI Deployers can use the Framework as a basis for communicating and coordinating about their respective roles and responsibilities for managing AI risks throughout a system's lifecycle.
- **Vendor Relations.** AI Deployers may choose to use the Framework to guide purchasing decisions and/or developing vendor contracts that ensure AI risks have been adequately accounted for.
- **Trust and Confidence.** AI Developers may wish to communicate information about a product's features and its approach to mitigating AI bias risks to a public audience. In that sense, the Framework can help organizations communicate to the public about their commitment to building ethical AI systems.
- **Incident Response.** Following an unexpected incident, the processes and documentation set forth in the Framework provide an audit trail that can help AI Developers and AI Deployers identify the potential source of system underperformance or failure.

## AI Lifecycle Phases

The Framework is organized around the phases of the AI lifecycle, which represent the key iterative steps involved in the creation and use of an AI system.



### DESIGN PHASE

---

- **Project Conception.** The initial stage of AI design involves identifying and formulating the “problem” that the system is intended to address and initially mapping how the model will achieve that objective. During this phase, the design team will define the purpose and structure of the system. Depending on the nature of the system, the design team will identify a target variable that the system is intended to predict. For instance, a fitness app that analyzes a consumer’s heart rate to monitor for irregularities that might predict whether that person is at risk of a stroke or heart disease (i.e., the target variable). At this early stage of the system design process, the goal of the Bias Risk Management Framework is to identify whether using AI is appropriate for the project at hand. Potential risks include:
  - **Problem Formulation Bias.** Target variables may reflect inherent prejudices or faulty assumptions that can perpetuate harmful biases. In some instances, the basic assumptions underlying a proposed AI system may be so inherently biased as to render it inappropriate for any form of public deployment.
- **Data Acquisition.** Once the system objectives have been defined, developers must assemble a corpus of data that will be used to train the model to identify patterns that will enable it to make predictions about future data inputs. This training data can inadvertently introduce biases into an AI system in many ways. Potential risks include:
  - **Historical Bias.** Training an AI system using data that itself may reflect historical biases creates a risk of further entrenching those inequities.
  - **Sampling Bias.** The risk of bias also arises when the data used to train an AI system is not representative of the population in which it will be deployed. An AI system trained on unrepresentative data may not operate as effectively when making predictions about a member of a class that is either over- or under-represented.
  - **Labeling Bias.** Many AI systems require training data to be labeled so that it can identify what patterns it should be looking for. The process of labeling the training dataset can be a vector for introducing bias into the AI system.



## DEVELOPMENT PHASE

---

- **Data Preparation and Model Definition.** The next step of the AI lifecycle involves preparing the data so that it is ready to train the model. During this process, the development team will clean, normalize, and identify the variables (i.e., “features”) in the training data that the algorithm will evaluate as it looks for patterns and relationships as the basis of a rule for making future predictions. The team must also establish the system’s underlying architecture, including selecting the type of algorithmic model that will power the system (e.g., linear regression, logistic regression, deep neural network).<sup>21</sup> Once the data is ready and the algorithm is selected, the team will train the system to produce a functional model that can make predictions about future data inputs. Potential risks include the following:
  - **Proxy Bias.** The process of selecting features in the training data and choosing a modeling approach involves human decisions about what variables should be considered as relevant for making predictions about the model’s target variable. These interventions can inadvertently introduce bias to the system, including by relying on variables that act as proxies for protected classes.
  - **Aggregation Bias.** Aggregation bias can arise if the model fails to account for underlying differences between sub-groups that materially impact a system’s accuracy rates. Using a “one-size-fits-all” model that overlooks key variables can result in system performance that is optimized only for the dominant sub-group.
- **Model Validation, Testing, and Revision.** After the model has been trained, it must be validated to determine if it is operating as intended and tested to demonstrate that the system’s outputs do not reflect unintended bias. Based on outcome of validation and testing, the model may need to be revised to mitigate risks of bias that are deemed unacceptable.



## DEPLOYMENT PHASE

---

- **Deployment and Use.** Prior to deployment, the AI Developer should evaluate the system to determine whether risks identified in earlier stages of design and development have been sufficiently mitigated in a manner that corresponds to the company’s governance policies. To the extent identified risks may arise through misuse of the system, the AI Developer should seek to control for them by integrating product features (e.g., user interfaces that reduce risk of misuse) to mitigate those risks, prohibiting uses that could exacerbate risks (e.g., end-user license agreements), and providing AI Deployers with sufficient documentation to perform their own impact assessments.

Prior to using an AI system, an AI Deployer should review documentation provided by the AI Developer to assess whether the system corresponds with its own AI governance policies and to determine whether deployment-related risk management responsibilities are clearly assigned.

Although some post-deployment risk management responsibilities may be addressed by the AI Developer, the AI Deployer will often bear responsibility for monitoring system performance and evaluating whether it is operating in a manner that is consistent with its risk profile. Potential risks include:

- **Deployment Bias.** AI systems are trained on data that represents a static moment in time and that filters out “noise” that could undermine the model’s ability to make consistent and accurate predictions. Upon deployment in the real world, AI systems will necessarily encounter conditions that differ from those in the development and testing environment. Further, because the real-world changes over time, the snapshot in time that a model represents may naturally become less accurate as the relationship between data variables evolves. If the input data for a deployed AI system differs materially from its training data, there is a risk that the system could “drift” and that the performance of the model could be undermined in ways that will exacerbate the risks of bias. For instance, if an AI system is designed (and tested) for use in a specific country, the system may not perform well if it is deployed in a country with radically different demographics.
- **Misuse Bias.** Deploying an AI system into an environment that differs significantly from the conditions for which it was designed or for purposes that are inconsistent with its intended use cases can exacerbate risks of bias.

## Framework Structure

The Framework identifies best practices for identifying and mitigating risks of AI bias across the entire system lifecycle. It is organized into:

- **Functions**, which denote fundamental AI risk management activities at their highest level, dividing them between Impact Assessment and Risk Mitigation Best Practices.
- **Categories**, which set out the activities and processes that are needed to execute upon the Functions at each phase of the AI Lifecycle. In other words, the Categories set forth the steps for performing an Impact Assessment and identify the corresponding Risk Mitigation Best Practices that can be used to manage associated risks.
- **Diagnostic Statements**, which set forth the discrete actions that should be taken to execute upon the Categories. They provide a set of results that help support achievement of the outcomes in each Category.
- **Comments on Implementation**, which provide additional information for achieving the outcomes described in the Diagnostic Statements.
- **Tools and Resources**, which identify a range of external guidance and toolkits that stakeholders can use to mitigate the bias risks associated with each phase of the AI lifecycle. The specific tools and resources identified in the framework are non-exhaustive and are highlighted for informational purposes only.

## Stakeholder Roles and Responsibilities

Reflecting the inherently dynamic nature of AI systems, the Framework is intended to account for the array of stakeholders that may play a role in various aspects of a system’s design, development, and deployment. Because there is no single model of AI development or deployment, it is impossible in the abstract to assign roles or delegate specific responsibilities for many of the Framework’s risk management functions. However, in general, there are three sets of stakeholders that may bear varying degrees of responsibility for certain aspects of AI risk management throughout a system’s lifecycle:

- **AI Developers.** AI Developers are organizations responsible for the design and development of AI systems.
- **AI Deployers.** AI Deployers are the organizations that adopt and use AI systems. (If an entity develops its own system, it is both the AI Developer and the AI Deployer.)
- **AI End-Users.** AI End-Users are the individuals—oftentimes an employee of an AI Deployer—who are responsible for overseeing the use of an AI system.

The allocation of risk management responsibilities between these stakeholders will in many cases depend on an AI system’s development and deployment model.

## Spectrum of AI Development and Deployment Models

The appropriate allocation of risk management responsibilities between stakeholders will vary depending on the nature of the AI system being developed and which party determines the purposes and means by which the underlying model is trained. For instance:

- **Universal, Static Model.** The AI Developer provides all its customers (i.e., AI Deployers) with a static, pre-trained model.
  - The AI Developer will bear responsibility for most aspects of model risk management.
- **Customizable Model.** The AI Developer provides a pre-trained model to AI Deployers who can customize and/or retrain the model using their own data.
  - Risk management will be a shared responsibility between the AI Developer and the AI Deployer.
- **Bespoke Model.** The AI Developer trains a bespoke AI model on behalf of an AI Deployer using the AI Deployer’s data.
  - Risk management will be a shared responsibility between the AI Developer and the AI Deployer, with the bulk of obligations falling on the AI Deployer.

# BSA AI Bias Risk Management Framework

 <b>DESIGN</b>			
Function	Category	Diagnostic Statement	Comments on Implementation
<b>PROJECT CONCEPTION</b>			
<b>Impact Assessment</b>	Identify and Document Objectives and Assumptions	<p>Document the intent and purpose of the system.</p>	<ul style="list-style-type: none"> <li>• What is the purpose of the system—i.e., what “problem” will it solve?</li> <li>• Who is the intended user of the system?</li> <li>• Where and how will the system be used?</li> <li>• What are the potential misuses?</li> </ul>
		<p>Clearly define the model’s intended effects.</p>	What is the model intended to predict, classify, recommend, rank, or discover?
		<p>Clearly define intended use cases and context in which the system will be deployed.</p>	
	Select and Document Metrics for Evaluating Fairness	<p>Identify “fairness” metrics that will be used as a baseline for assessing bias in the AI system.</p>	The concept of “fairness” is highly subjective and there are dozens of metrics by which it can be evaluated. Because it is impossible to simultaneously satisfy all fairness metrics, it is necessary to select metrics that are most appropriate for the nature of the AI system that is being developed and consistent with any applicable legal requirements. It is important to document the rationale by which fairness metrics were selected and/or excluded to inform latter stages of the AI lifecycle.
	Document Stakeholder Impacts	<p>Identify stakeholder groups that may be impacted by the system.</p>	Stakeholder groups include AI Deployers, AI End-Users, Affected Individuals (i.e., members of the public who may interact with or be impacted by an AI system).
		<p>For each stakeholder group, document the potential benefits and potential adverse impacts, considering both the intended uses and reasonably foreseeable misuses of the system.</p>	
		<p>Assess whether the nature of the system makes it prone to potential bias-related harms based on user demographics.</p>	User demographics may include, but are not limited to race, gender, age, disability status, and their intersections.
	Document Risk Mitigations	<p>If risk of bias is present, document efforts to mitigate risks.</p>	

DESIGN			
Function	Category	Diagnostic Statement	Comments on Implementation
<b>PROJECT CONCEPTION</b>			
<b>Impact Assessment</b> (continued)	Document Risk Mitigations	Document how identified risks and potential harms of each risk will be measured and how the effectiveness of mitigation strategies will be evaluated.	
		If risk of bias is present, document efforts to mitigate risks.	
		If risks are unmitigated, document why the risk was deemed acceptable.	
<b>Risk Mitigation Best Practices</b>	Independence and Diversity	Seek feedback from a diverse set of stakeholders to inform the impact assessment.	Because risks identified during this initial phase will inform later aspects of the development and impact assessment processes, it is vital to develop a holistic understanding of potential harms that may arise by soliciting diverse perspectives from people with a range of lived experiences, cultural backgrounds, and subject matter expertise. To the extent in-house personnel lack subject matter or cultural diversity, it may be necessary to consult with third-party experts or to solicit feedback from members of communities that may be adversely impacted by the system.
	Transparent Documentation	Share impact assessment documentation with personnel working on later stages of the AI pipeline so that risks and potential unintended impacts can be monitored throughout the development process.	
	Accountability and Governance	Ensure that senior leadership has been adequately briefed on potential high risk AI systems.	Impact assessment documentation for systems deemed "high risk" should be shared with senior leadership to facilitate a "go/no-go" decision.
<b>DATA ACQUISITION</b>			
<b>Impact Assessment</b>	Maintain Records of Data Provenance	Maintain sufficient records to enable "recreation" of the data used to train the AI model, verify that its results are reproducible, and monitor for material updates to data sources.	<p>Records should include:</p> <ul style="list-style-type: none"> <li>• Source of data</li> <li>• Origin of data (e.g., Who created it? When? For what purpose? How was it created?)</li> <li>• Intended uses and/or restrictions of the data and data governance rules (e.g., What entity owns the data? How long can it be retained (or must it be destroyed)? Are there restrictions on its use?)</li> <li>• Known limitations of data (e.g., missing elements?)</li> <li>• If data is sampled, what was the sampling strategy?</li> <li>• Will the data be updated? If so, will any versions be tracked?</li> </ul>

DESIGN			
Function	Category	Diagnostic Statement	Comments on Implementation
<b>DATA ACQUISITION</b>			
<b>Impact Assessment</b> <i>(continued)</i>	Examine Data for Potential Biases	<p>Scrutinize data for historical biases.</p> <p>Evaluate “representativeness” of the data.</p>	<p>Examine sources of data and assess potential that they may reflect historical biases.</p> <ul style="list-style-type: none"> <li>• Compare demographic distribution of training data to the population where the system will be deployed.</li> <li>• Assess whether there is sufficient representation of subpopulations that are likely to interact with the system.</li> </ul>
	Document Risk Mitigations	Scrutinize data labeling methodology.	<ul style="list-style-type: none"> <li>• Document personnel and processes used to label data.</li> <li>• For third-party data, scrutinize labeling (and associated methodologies) for potential sources of bias.</li> </ul>
<b>Risk Mitigation Best Practices</b>	Independence and Diversity	To facilitate robust interrogation of the datasets, data review teams should include personnel that are diverse in terms of their subject matter expertise and lived experiences.	Effectively identifying potential sources of bias in data requires a diverse set of expertise and experiences, including familiarity with the domain from which data is drawn and a deep understanding of the historical context and institutions that produced it. To the extent in-house personnel lack diversity, consultation with third-party experts or potentially affected stakeholder groups may be necessary.
	Re-Balancing Unrepresentative Data	<p>Consider re-balancing with additional data.</p> <p>Consider re-balancing with synthetic data.</p>	<p>Improving representativeness can be achieved in some circumstances by collecting additional data that improves the balance of the overall training dataset.</p> <p>Imbalanced datasets can potentially be rebalanced by “oversampling” data from the underrepresented groups. A common oversampling method is the Synthetic Minority Oversampling Technique, which generates new “synthesized” data from the underrepresented group.</p>

DESIGN			
Function	Category	Diagnostic Statement	Comments on Implementation
<b>DATA ACQUISITION</b>			
<b>Risk Mitigation Best Practices</b> <i>(continued)</i>	Data Labeling	Establish objective and scalable labeling guidelines.	<ul style="list-style-type: none"> <li>To mitigate the potential of labeling bias, the personnel responsible for labeling the data should be provided with clear guidelines establishing an objective and repeatable process for individual labeling decisions.</li> <li>In domains where the risk of bias is high, labelers should have adequate subject matter expertise and be provided training to recognize potential unconscious biases.</li> <li>For high-risk systems, it may be necessary to set up a quality assurance mechanism to monitor label quality.</li> </ul>
	Accountability and Governance	Integrate data labeling processes into a comprehensive data strategy.	Establishing an organizational data strategy can help ensure that data evaluation is performed consistently and prevent duplication of effort by ensuring that company efforts to scrutinize data are documented for future reference.

## DESIGN: RISK MITIGATION TOOLS AND RESOURCES

### Project Conception

- Aequitas Bias and Fairness Audit Toolkit**  
Pedro Saleiro, Abby Stevens, Ari Anisfeld, and Rayid Ghani, University of Chicago Center for Data Science and Public Policy (2018), <http://www.datasciencepublicpolicy.org/projects/aequitas/>.
- Diverse Voices Project | A How-To Guide for Facilitating Inclusiveness in Tech Policy**  
Lassana Magassa, Meg Young, and Batya Friedman, University of Washington Tech Policy Lab, <https://techpolicylab.uw.edu/project/diverse-voices/>.

### Data Compilation

- Datasheets for Datasets**  
Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, arXiv:1803.09010v7, (March 19, 2020), <https://arxiv.org/abs/1803.09010>.
- AI FactSheets 360**  
IBM Research, <https://aif360.mybluemix.net/>.

DEVELOPMENT			
Function	Category	Diagnostic Statement	Comments on Implementation
<b>DATA PREPARATION AND MODEL DEFINITION</b>			
<b>Impact Assessment</b>	Document Feature Selection and Engineering Processes	Document rationale for choices made during the feature selection and engineering processes and evaluate their impact on model performance.	Examine whether feature selection or engineering choices may rely on implicitly biased assumptions.
		Document potential correlation between selected features and sensitive demographic attributes.	For features that closely correlate to a sensitive class, document the relevance to the target variable and the rationale for its inclusion in the model.
	Document Model Selection Process	Document rationale for the selected modeling approach.	
		Identify, document, and justify assumptions in the selected approach and potential resulting limitations.	
<b>Risk Mitigation Best Practices</b>	Feature Selection	Examine for biased proxy features.	<ul style="list-style-type: none"> <li>Simply avoiding the use of sensitive attributes as inputs to the system—an approach known as “fairness through unawareness”—is not an effective approach to mitigating the risk of bias. Even when sensitive characteristics are explicitly excluded from a model, other variables can act as proxies for those characteristics and introduce bias into the system. To avoid the risk of proxy bias, the AI Developer should examine the potential correlation between a model’s features and protected traits and examine what role these proxy variables may be playing in the model’s output.</li> <li>The ability to examine statistical correlation between features and sensitive attributes may be constrained in circumstances where an AI Developer lacks access to sensitive attribute data and/or is prohibited from making inferences about such data.<sup>22</sup> In such circumstances, a more holistic analysis informed by domain experts may be necessary.</li> </ul>

DEVELOPMENT			
Function	Category	Diagnostic Statement	Comments on Implementation
<b>DATA PREPARATION AND MODEL DEFINITION</b>			
<b>Risk Mitigation Best Practices</b> <i>(continued)</i>	Feature Selection	Scrutinize features that correlate to sensitive attributes.	<ul style="list-style-type: none"> <li>Features that are known to correlate to a sensitive attribute should only be used if there is a strong logical relationship to the system's target variable.</li> <li>For example, income—although correlated to gender—is reasonably related to a person's ability to pay back a loan. The use of income in an AI system designed to evaluate creditworthiness would therefore be justified. In contrast, the use of "shoe size"—which also correlates to gender—in a model for predicting creditworthiness would be an inappropriate use of a variable that closely correlates to a sensitive characteristic.</li> </ul>
	Independence and Diversity	Seek feedback from diverse stakeholders with domain-specific expertise.	The feature engineering process should be informed by personnel with diverse lived experiences and expertise about the historical, legal, and social dimensions of the data being used to train the system.
	Model Selection	Avoid inscrutable models in circumstances where both the risk and potential impact of bias are high.	Using more interpretable models can mitigate the risks of unintended bias by making it easier to identify and mitigate problems.
<b>VALIDATING, TESTING, AND REVISING THE MODEL</b>			
<b>Impact Assessment</b>	Document Validation Processes	Document how the system (and individual components) will be validated to evaluate whether it is performing consistent with the design objectives and intended deployment scenarios.	<ul style="list-style-type: none"> <li>Establish cadence at which model will be regularly re-validated.</li> <li>Establish performance benchmarks that will trigger out-of-cycle re-validation.</li> </ul>
	Document Testing Processes	<p>Document re-validation processes.</p> <p>Test the system for bias by evaluating and documenting model performance.</p>	<p>Testing should incorporate fairness metrics identified during Design phase and examine the model's accuracy and error rates across demographic groups.</p>
		<p>Document how testing was performed, which fairness metrics were evaluated, and why those measures were selected.</p> <p>Document model interventions.</p>	If testing reveals unacceptable levels of bias, document efforts to refine the model.

DEVELOPMENT			
Function	Category	Diagnostic Statement	Comments on Implementation
<b>VALIDATING, TESTING, AND REVISING THE MODEL</b>			
<b>Risk Mitigation Best Practices</b>	Model Interventions	Evaluate potential model refinements to address bias surfaced during testing.	<p>In circumstances where testing reveals that the system is exhibiting unacceptable levels of bias based on the selected fairness metric, it will be necessary to refine the model. Potential model refinements include:</p> <ul style="list-style-type: none"> <li><b>Pre-Processing Interventions.</b> Such refinements can involve revisiting earlier stages of the Design and Development lifecycle (e.g., seeking out additional training data).</li> <li><b>In-Processing Interventions.</b> Bias can also be mitigated by imposing an additional fairness constraint directly on the model. Traditional machine learning models are designed to maximize for predictive accuracy. Emerging techniques enable developers to build constraints into the model to reduce the potential for bias across groups. The addition of a fairness constraint, in effect, instructs the model to optimize both for accuracy and a specific fairness metric.</li> <li><b>Post-Processing Interventions.</b> In some cases, bias can be addressed through the use of post-processing algorithms that manipulate the model's output predictions to ensure that it adheres to a desired distribution.</li> </ul>
Independence and Diversity		Validation and testing documentation should be reviewed by personnel who were not involved in the system's development.	The independent team should compare the validation and testing results to the system specifications developed during earlier phases of the design and development process.

## DEVELOPMENT: RISK MITIGATION TOOLS AND RESOURCES

- **Model Cards for Model Reporting**

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, (January 2019): 220–229, <https://arxiv.org/abs/1810.03993>.

- **AI Factsheets 360**

Aleksandra Mojsilovic, IBM Research (August 22, 2018), <https://www.ibm.com/blogs/research/2018/08/factsheets-ai/>.

- **AI Explainability 360**

IBM Research, <https://aix360.mybluemix.net/>.

- **AI Fairness 360**

IBM Research, <https://aif360.mybluemix.net/>.

- **Responsible Machine Learning with Error Analysis**

Besmira Nushi, Microsoft Research (February 18, 2021), <https://techcommunity.microsoft.com/t5/azure-ai/responsible-machine-learning-with-error-analysis/ba-p/2141774>.

- **Aequitas Open Source Bias Audit Toolkit**

Pedro Saleiro, Abby Stevens, Ari Anisfeld, and Rayid Ghani, University of Chicago Center for Data Science and Public Policy, <http://www.datasciencelppublicpolicy.org/projects/aequitas/>.

- **FairTest: Discovering Unwarranted Associations in Data-Driven Applications**

Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels and Huang Lin, ArXiv, (2015), <https://github.com/columbia/fairtest>.

- **Bayesian Improved Surname Geocoding**

Consumer Finance Protection Bureau (2014), [https://files.consumerfinance.gov/f/201409\\_cfpb\\_report\\_proxy-methodology.pdf](https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf).

Deployment and Use			
Function	Category	Diagnostic Statement	Comments on Implementation
<strong>PREPARING FOR DEPLOYMENT AND USE</strong>			
<strong>Impact Assessment</strong>	Document Lines of Responsibility	Define and document who is responsible for the system's outputs and the outcomes they may lead to, including details about how a system's decisions can be reviewed if necessary.	
		Establish management plans for responding to potential incidents or reports of system errors.	<ul style="list-style-type: none"> <li>• What does it mean for the system to fail and who might be harmed by a failure?</li> <li>• How will failures be detected?</li> <li>• Who will respond to failures when they are detected?</li> <li>• Can the system be safely disabled?</li> <li>• Are there appropriate plans for continuity of critical functions?</li> </ul>
	Document Processes for Monitoring Data	Document what processes and metrics will be used to evaluate whether production data (i.e., input data the system encounters during deployment) differs materially from training data.	
	Document Processes for Monitoring Model Performance	For static models, document how performance levels and classes of error will be monitored over time and benchmarks that will trigger review.	
		For models that are intended to evolve over time, document how changes will be inventoried; if, when, and how versions will be captured and managed; and how performance levels will be monitored (e.g., cadence of scheduled reviews, performance indicators that may trigger out-of-cycle review).	
	Document Audit and End-of-Life Processes	Document the cadence at which impact assessment evaluations will be audited to evaluate whether risk mitigation controls remain fit for purpose.	
		Document expected timeline that system support will be provided and processes for decommissioning system in event that it falls below reasonable performance thresholds.	
	Monitoring for Drift and Model Degradation	Input data encountered during deployment can be evaluated against a statistical representation of the system's training data to evaluate the potential for data drift (i.e., material differences between the training data and deployment data that can degrade model performance).	
<strong>Risk Mitigation Best Practices</strong>			

DEPLOYMENT AND USE			
Function	Category	Diagnostic Statement	Comments on Implementation
<b>PREPARING FOR DEPLOYMENT AND USE</b>			
<b>Risk Mitigation Best Practices</b> <i>(continued)</i>	Product Features and User Interface	Integrate product and user interface features to mitigate risk of foreseeable unintended uses—e.g., interface that enforces human-in-the-loop requirements, alerts to notify when a system is being misused.	
	System Documentation	AI Developers should provide sufficient documentation regarding system capabilities, specifications, limitations, and intended uses to enable AI Deployers to perform independent impact assessment concerning deployment risks.	If necessary, AI Developers can also provide AI Deployers with a technical environment to perform an independent impact assessment.
		Consider incorporating terms into the End-User License Agreement that set forth limitations designed to prevent foreseeable misuses (e.g., contractual obligations to ensure end-user will comply with acceptable use policy).	
		Sales and marketing materials should be closely reviewed to ensure that they are consistent with the system's actual capabilities.	
	AI User Training	AI Deployers should provide training for AI Users regarding a system's capabilities and limitations, and how outputs should be evaluated and integrated into a workflow.	For human-in-the-loop oversight of AI system to be an effective risk mitigation measure, AI Users should be provided adequate information and training so they can understand how the system is operating and make sense of the model's outputs.
	Incident Response and Feedback Mechanisms	AI Deployers should maintain a feedback mechanism to enable AI Users and Affected Individuals (i.e., members of the public that may interact with the system) to report concerns about the operation of a system.	For consequential decisions, Affected Individuals should be provided with an appeal mechanism.

#### DEPLOYMENT AND USE: RISK MITIGATION TOOLS AND RESOURCES

- **AI Incident Response Checklist**  
BNH.AI, <https://www.bnhi.ai/public-resources>.
- **Watson OpenScale**  
IBM, <https://www.ibm.com/cloud/watson-openscale>.
- **Detect Data Drift on Datasets**  
Microsoft Azure Machine Learning (June 25, 2020), <https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?tabs=python#create-dataset-monitors>.

# Foundational Resources

## ***A Framework for Understanding Unintended Consequences of Machine Learning***

Harini Suresh and John V. Guttag, arXiv (February 2020), <https://arxiv.org/abs/1901.10002>.

## ***AI Fairness***

Trisha Mahoney, Kush R. Varshney, and Michael Hind, O'Reilly (April 2020), <https://www.oreilly.com/library/view/ai-fairness/9781492077664/>.

## ***Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models***

Andrew Burt, Brenda Leong, Stuart Shirrell, and Xiangnong (George) Wang, Future of Privacy Forum (June 2018), <https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf>.

## ***Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI***

Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach, CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (April 2020): 1–14, <https://doi.org/10.1145/3313831.3376445>.

## ***Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing***

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P., FAT\* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, (January 2020): 33–44, <https://doi.org/10.1145/3351095.3372873>.

## ***Supervisory Guidance on Model Risk Management***

US Federal Reserve Board (April 2011), <https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf>.

## ***Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector***

David Leslie, The Alan Turing Institute (2019), <https://doi.org/10.5281/zenodo.3240529>.

## ENDNOTES

- <sup>1</sup> Gina Kolata, "Alzheimer's Prediction May Be Found in Writing Tests," *New York Times* (February 1, 2021), <https://www.nytimes.com/2021/02/01/health/alzheimers-prediction-speech.html>.
- <sup>2</sup> Dina Temple-Raston, *Elephants under Attack Have an Unlikely Ally: Artificial Intelligence*, NPR (October 25, 2019), <https://www.npr.org/2019/10/25/760487476/elephants-under-attack-have-an-unlikely-ally-artificial-intelligence>.
- <sup>3</sup> *Seeing AI: An App for Visually Impaired People That Narrates the World Around You*, Microsoft, <https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/>.
- <sup>4</sup> See e.g., Jennifer Sukis, *The Origins of Bias and How AI May Be the Answer to Ending Its Reign*, Medium (January 13, 2019), <https://medium.com/design-ibm/the-origins-of-bias-and-how-ai-might-be-our-answer-to-ending-it-acc3610d6354>.
- <sup>5</sup> See e.g., Nicol Turner Lee, Paul Resnick, and Genie Barton, *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*, Brookings (May 22, 2019), <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>.
- <sup>6</sup> Harini Suresh and John V. Guttag, *A Framework for Understanding Unintended Consequences of Machine Learning* (February 17, 2020), <https://arxiv.org/pdf/1901.10002.pdf>.
- <sup>7</sup> See Xiaolin Wu and Xi Zhang, *Automated Inference on Criminality Using Face Images*, Shanghai Jiao Tong University (November 13, 2016), <https://arxiv.org/pdf/1611.04135v1.pdf>.
- <sup>8</sup> Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov, *Physiognomy's New Clothes*, Medium (May 6, 2017), <https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a>.
- <sup>9</sup> Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* (October 25, 2019), <https://science.sciencemag.org/content/366/6464/447>.
- <sup>10</sup> Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California University Law Review* 104, no. 3 (September 30, 2016): 671, <http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf>.
- <sup>11</sup> Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77–91, <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>.
- <sup>12</sup> Kate Crawford, *The Hidden Biases in Big Data*, Harvard Business Review (April 1, 2013), <https://hbr.org/2013/04/the-hidden-biases-in-big-data>.
- <sup>13</sup> Kate Crawford and Trevor Paglen, *Excavating AI: The Politics of Images in Machine Learning Training Sets* (September 19, 2019), <https://excavating.ai/>.
- <sup>14</sup> Cade Metz, "'Nerd,' 'Non-smoker,' 'Wrongdoer': How Might A.I. Label You?" *New York Times* (September 20, 2019), <https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html>.
- <sup>15</sup> Jessica Zosa Forde, A. Feder Cooper, Kweku Kwagyir-Aggrey, Chris De Sa, and Michael Littman, *Model Selection's Disparate Impact in Real-World Deep Learning Applications*, arXiv:2104.00606 (April 1, 2021), <https://arxiv.org/abs/2104.00606>.
- <sup>16</sup> Aaron Klein, *Credit Denial in the Age of AI*, Brookings Institution (April 11, 2019), <https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/>.
- <sup>17</sup> J. Vaughn, A. Baral, M. Vadari "Analyzing the Dangers of Dataset Bias in Diagnostic AI systems: Setting Guidelines for Dataset Collection and Usage," ACM Conference on Health, Inference and Learning, 2020 Workshop, [http://juliev42.github.io/files/CHIL\\_paper\\_bias.pdf](http://juliev42.github.io/files/CHIL_paper_bias.pdf).
- <sup>18</sup> Arvind Narayanan, *21 Fairness Definitions and Their Politics*, ACM Conference on Fairness, Accountability and Transparency (March 1, 2018), <https://www.youtube.com/watch?v=jIXluYdnyyk>.
- <sup>19</sup> Reuben Binns and Valeria Gallo, *AI Blog: Trade-Offs*, UK Information Commission's Office (July 25, 2019), <https://ico.org.uk/about-the-ico/news-and-events/ai-blog-trade-offs/>.
- <sup>20</sup> Inioliwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, FAT\* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (January 2020): 33–44, <https://doi.org/10.1145/3351095.3372873>.
- <sup>21</sup> Sara Hooker, Moving Beyond "Algorithmic Bias Is a Data Problem," *Patterns* (April 9, 2021), <https://www.sciencedirect.com/science/article/pii/S2666389921000611>.
- <sup>22</sup> McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang, "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness, arXiv:2011.02282 (January 23, 2021), <https://arxiv.org/abs/2011.02282>.



[www.bsa.org](http://www.bsa.org)

BSA Worldwide Headquarters  
20 F Street, NW  
Suite 800  
Washington, DC 20001

 +1.202.872.5500  
 @BSAnews  
 @BSATheSoftwareAlliance

BSA Asia-Pacific  
300 Beach Road  
#30-06 The Concourse  
Singapore 199555

 +65.6292.2072

BSA Europe, Middle East & Africa  
44 Avenue des Arts  
Brussels 1040  
Belgium

 +32.2.274.13.10

**DISCUSSION PAPER SERIES**

IZA DP No. 14021

**Public Procurement and Innovation for  
Human-Centered Artificial Intelligence**

Wim Naudé  
Nicola Dimitri

JANUARY 2021

## DISCUSSION PAPER SERIES

IZA DP No. 14021

# Public Procurement and Innovation for Human-Centered Artificial Intelligence

**Wim Naudé**

*University College Cork, RWTH Aachen University and IZA*

**Nicola Dimitri**

*University of Siena*

JANUARY 2021

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9  
53113 Bonn, Germany

Phone: +49-228-3894-0  
Email: publications@iza.org

[www.iza.org](http://www.iza.org)

## ABSTRACT

# Public Procurement and Innovation for Human-Centered Artificial Intelligence

The possible negative consequences of Artificial Intelligence (AI) have given rise to calls for public policy to ensure that it is safe, and to prevent improper use and misuse. Human-centered AI (HCAI) draws on ethical principles and puts forth actionable guidelines in this regard. So far however, these have lacked strong incentives for adherence. In this paper we contribute to the debate on HCAI by arguing that public procurement and innovation (PPal) can be used to incentivize HCAI. We dissect the literature on PPal and HCAI and provide a simple theoretical model to show that procurement of innovative AI solutions underpinned by ethical considerations can provide the incentives that scholars have called for. Our argument in favor of PPal for HCAI is also an argument for the more innovative use of public procurement, and is consistent with calls for mission-oriented and challenge-led innovation policies. Our paper also contributes to the emerging literature on public entrepreneurship, given that PPal for HCAI can advance the transformation of society, but only under uncertain.

**JEL Classification:** H57, D02, O38, O32

**Keywords:** artificial intelligence, data, innovation, public procurement, ethics

**Corresponding author:**

Wim Naudé  
Technology and Innovation Management (TIM)  
RWTH Aachen University  
Kackertstraße 7  
52072 Aachen  
Germany  
E-mail: naude@time.rwth-aachen.de

# 1 Introduction

Artificial intelligence (AI) is<sup>1</sup> the “simulation of human intelligence processes by machines, especially computer systems.” These processes include learning, reasoning and self-correction. Although the term Artificial Intelligence (AI) was first used in 1956, when attempts to build AI was based on computational logic, modern AI is based on a statistical-probabilistic approach and is specialized in narrow domains (Moor, 2006; Naudé, 2021). It is also known as narrow AI, to distinguish it from the not-yet-existing Artificial General Intelligence (AGI) which would be indistinguishable from human intelligence (Stanford University, 2016). The use of Machine Learning (ML) and in particular Deep Learning (DL) based on big data has been particularly successful in generating technologies that have become as good and, in many cases, better than humans at pattern recognition, prediction and natural language processing (NLP). As a result, AI models are now routinely and widely used to provide services and products such as online search engines, chatbots and virtual assistants, recommender systems, reputation systems, news curation and aggregation, hyper-personalization of marketing, translation, credit scoring, predictive policing and spam filters, amongst others. Progress has also been made in developing autonomous vehicles and medical diagnostic tools.

Given the ubiquity and growth in data for ML, allowing for AI models to improve their performance over time, and the wide range of terrains where pattern recognition prediction and NLP are required, it is clear why there has been high expectations put on AI, and this is why it has even been described as a new general-purpose technology (Trajtenberg, 2018). The high expectations of AI are reflected not only in the declarations of scientists and government officials, but also in the growing investments and research in the field of AI specifically, and data science more generally. With these expectations have however also come warnings and concerns about the possible long-term existential threats of AI (Bostrom, 2014) as well as the shorter-term negative consequences. The latter include intrusive surveillance<sup>2</sup> and erosion of privacy, AI weapons, job losses due to automation, higher inequality, discrimination and biased policy making (Frey and Osborne, 2017; Korinek and Stiglitz, 2017; Feldstein, 2019; Russel et al., 2015). The GitHub site “Awful AI” contains a repository of some of the negative consequences of AI.<sup>3</sup>

---

<sup>1</sup>See [www.WhatIs.com](http://www.WhatIs.com); Note however that there is no single universal definition of AI (Van de Gevel and Noussair, 2013).

<sup>2</sup>As Smith and Neupane (2018, p.25) point out, “AI algorithms supercharge surveillance by processing data faster than previously possible and detecting patterns too subtle for human analysts to uncover.”

<sup>3</sup>For Awful AI, see <https://github.com/daviddao/awful-ai>.

These negative consequences of AI have given rise to calls for proper AI governance<sup>4</sup> (Dafoe, 2018) to ensure that AI-based applications are safe, and that the development and diffusion of AI do not suffer from improper use and misuse. What would count as improper use and misuse, and safety-risks such as that could be caused by accidents or unintended consequences of AI systems, would be determined by the extent to which AI systems are human-centered. Human-centered AI (HCAI) draws on ethical principles and puts forth actionable guidelines for reducing the risks mentioned (Shneiderman, 2020). As such, HCAI is concerned with Ethical AI<sup>5</sup> and Responsible AI.<sup>6</sup>

In recent years there have been a growing number of initiatives to elaborate on Ethical and Responsible AI, see e.g. Boddington (2017), Etzioni and Etzioni (2017), Floridi (2018) and Yu et al. (2018). These include the Asilomar AI Principles of the Institute of the Future of Life<sup>7</sup> (2017), the European Union's April 2019 Ethics Guidelines for Trustworthy AI (EC, 2019), the OECD's May 2019 Principles on Artificial Intelligence (OECD, 2019), the G-20's June 2019 Human-Centered AI Principles (G-20, 2019) and the Institute of Electrical and Electronics Engineers' Ethically Aligned Design Principles (IEEE, 2019). In addition to these cross-country and multinational initiatives, the European Union Agency for Fundamental Rights (FRA) has documented more than 290 AI policy initiatives in individual EU Member States between 2016 and 2020.<sup>8</sup>

A shortcoming of these is that they tend to lack strong incentives for developers and users of AI to adhere to them (Askell et al., 2019; Calo, 2017; Hagendorff, 2020). To incentivize the ethics for HCAI, Eitel-Porter (2020, p.1) argues for “strong, mandated governance controls including tools for managing processes and creating associated audit trails.” Floridi et al. (2018, p.704) argues that governments should “incentivise financially the inclusion of ethical, legal and social considerations in AI research projects.”

In this paper we contribute to the debate on HCAI by arguing that public procurement and innovation is a potentially relevant tool with which to incentivize HCAI. We provide a simple

<sup>4</sup>“The field of AI governance studies how humanity can best navigate the transition to advanced AI systems, focusing on the political, economic, military, governance, and ethical dimensions” (Dafoe, 2018, p.5). In 2016, an open letter by an eminent group of scientists ignited the field of AI governance by calling for “expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do” - see <https://futureoflife.org/ai-open-letter/>.

<sup>5</sup>It is also sometimes still referred to as “Machine Ethics”, which has been defined as being “concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable” (Anderson and Anderson, 2007, p.15)

<sup>6</sup>Responsible AI has been described as AI systems that “have an acceptably low risk of harming their users or society and, ideally, to increase their likelihood of being socially beneficial” (Askell et al., 2019, p.2).

<sup>7</sup>See <https://futureoflife.org/ai-principles/>

<sup>8</sup>See <https://tinyurl.com/y8juagop>.

theoretical model to show that such procurement of innovative AI solutions underpinned by ethical considerations, can provide both the tools and audit trails as well as the financial incentives that scholars such as Eitel-Porter (2020) and Floridi et al. (2018) have called for.

With the concept *public procurement and innovation* (PPAI) we encompass three dimensions of the relationship between public procurement and innovation (see Obweger and Müller (2018, p.5)). The first is *public procurement for innovation* (PPfI), which deals with the question, how can public procurement drive innovation? The second is *public procurement of innovation* (PPoI), which deals with the question, how can public services be innovated? And the third is *innovative public procurement* (IPP) which deals with the question, how can public institutions procure innovatively? Our paper is relevant for all three of these dimensions of innovation and public procurement although we will lay more stress on the first, namely PPfI, and in particular one of its relatively new tools in the EU, the pre-commercial procurement of innovation (PCP).

Our argument in favor of public procurement and innovation (PPAI) to advance HCAI is also an argument for the more innovative use of public procurement. Most often in the past, public procurement has been used ex post to support new innovations by creating a demand for the product (being a customer for new products). This is also the way in which Lin (2020, p.26) conceives of the relationship between public procurement for innovation support. However, this is a rather restricted view: we argue in this paper that through PPfI of innovation, governments can also ex ante support research and development of goods and services that do not yet exist. This potential instrument for steering HCAI has been neglected in both the AI as well as innovation literature.<sup>9</sup> Miller and Lehoux (2020, p.2) quoting from Uyarra et al. (2017, p.828) confirms that the underlying mechanisms of how public procurement of innovation can shape the incentives of private agents in innovation is still “under-theorized.” Our paper contributes by addressing this neglect.

Furthermore, our argument for PPAI as a policy towards HCAI is consistent with the calls for mission-oriented and challenge-led innovation policies, as made for instance by Mazzucato (2013, 2018). Mazzucato (2020, p.101) describe challenge-led policies as “policies that use investment and innovation to solve difficult problems.” Certainly, the issue of HCAI is a challenge and a difficult problem. It is pre-eminently a challenge facing humanity and requires challenge-led innovation policies. PPfI and PPoI can fulfil this purpose, as it “uses public

---

<sup>9</sup>For instance, Bloom et al. (2019) provides an innovation policy toolkit, discussing eight policy tools: direct R&D grants, R&D tax credits, patent boxes, skilled immigration, supporting universities’ research, competition policy, intellectual property rights, and mission-oriented policies. They do not discuss the role of public procurement of innovation.

procurement strategically to address a need which cannot be met by conventional solutions” (Lenderink et al., 2019, p.7).

Finally, our paper also contributes to the emerging literature of public entrepreneurship (see e.g., Hayter et al. (2018)), in that the use of public procurement of innovation for HCAI is an example of public entrepreneurship as it entails innovation, it aims to contribute towards transformation of society, and it is subject to uncertainty. These three elements – innovation, transformation and uncertainty, are according to Hayter et al. (2018, p.676) what characterises public entrepreneurship.

The rest of the paper will proceed as follows. In sections 2 and 3 we review the relevant literature dealing respectively with human-centered AI and innovation policy. In section 4 we provide a simple theoretical model wherein we show how public procurement of innovation can incentivize the development of HCAI. Section 5 concludes.

## 2 Human-Centered Artificial Intelligence

We start by reviewing the state of the literature on HCAI, in particular the role of ethical principles and actionable guidelines. This literature is part of the broader emerging literature on AI governance.

The downsides of AI, some of which was mentioned in the introduction, include intrusive surveillance and erosion of privacy, AI weapons, cybercrime, fake news and misinformation, job losses and tax losses due to automation, higher inequality, discrimination and biased policy making (Smith and Neupane, 2018). These are due to the misuse, accidents as well as systemic risks attached to the use of AI. It is clear that these downsides or risks posed by AI are of both an intentional nature (as in cybercrime, fake news and AI weapons) or unintentional nature (as in discrimination, inequality and accidents). In this paper we are largely concerned with the unintentional harm that AI can cause, and which can be more broadly analyzed as being a “systemic” or “accident” AI risks (Dafoe, 2018).

Systemic AI risk refers to “the risks of undesired outcomes - some of which may be very traditional - that can emerge from a system of competing and cooperating agents and can be amplified by novel forms of AI. For example, AI could increase the risk of inadvertent nuclear war, not because of an accident or misuse, but because of how AI could rapidly shift crucial strategic parameters, before we are able to build up compensating understandings,

norms, and institutions” (Dafoe, 2018, p.28).

One aspect of systemic risk is that AI models can unintentionally result in biased decisions and recommendations due to being based on biased data (or data with gaps and absences) as well as design bias, where the values of the designer influences the model design and functioning. AI models also suffer, due to the nature of ML from lack of transparency and accountability resulting in the problem that decisions and outcomes from AI models very often cannot be easily explained. This is also known as the “black-box problem” (Castelvecchi, 2016; Nguyen et al., 2014). With bias and lack of explainability, “inequality in application” is a consequence: AI does not benefit everyone equality or fairly (Calo, 2017). Moreover, these shortcomings “disproportionately affect groups that are already disadvantaged by factors such as race, gender and socio-economic back-ground” (Crawford and Calo, 2016, p.312).

As far as accident risks are concerned, an AI accident occurs if “a human designer had in mind a certain (perhaps informally specified) objective or task, but the system that was designed and deployed for that task produced harmful and unexpected results” (Amodei et al., 2016, p.2). “Normal” accidents from AI should be expected. According to Maas (2018, p.4) it even “appears plausible that many AI applications may be even more susceptible to normal accidents than past ‘textbook’ case technologies such as nuclear power or aviation.” There are at least three remedies for systemic and accident risks due to AI that have been given attention in the literature, and that forms part of the challenge of establishing a HCAI.

A first remedy is regulations and laws - to outlaw and police misuse – such as for instance through data privacy laws, combating cybercrime and other clearly malicious uses of AI – but also to lay down regulations for improving the safety of AI and limiting accidents. This remedy may be necessary but not sufficient. As Askell et al. (2019, p.7) remarks, the difficulties are that “AI regulation seems particularly tricky to get right, as it would require a detailed understanding of the technology on the part of regulators” and that “regulation that is reactive and slow may also be insufficient to deal with the challenges raised by AI systems.”

A second remedy is to improve the safety of AI, and hence limit accidents and unintentional consequences, through for instance addressing technical issues in design, monitoring and operation (Maas, 2018; Leike et al., 2017) and by subjecting AI systems to regular monitoring and redesign (van de Poel, 2020). The problem is that issues of AI technical safety are

complex,<sup>10</sup> and moreover as Leike et al. (2017, p.1) pointed out, “This nascent field of AI safety still lacks a general consensus on its research problems.”

Thirdly, AI risks – both system and accident risks - can be addressed through the adoption and adherence to ethical principles and guidelines for the development and use of AI – for Ethical and Responsible AI. Such principles are needed for combatting the misuse and under-use of AI, for instance by making clear when the use of AI would constitute a fair or unfair usage, as well as for informing the technical safety of AI, for instance in identifying questions or problems that the design of an AI should incorporate.<sup>11</sup> The requirement of fairness has led to the concept of “Fair AI,” a subset of Ethical AI, which refers to refers to “systems that both quantify bias and mitigate discrimination against subgroups” (Feuerriegel et al., 2020, p.379). HCAI extends beyond merely reducing misuse and negative side-effects and ensuring fairness. HCAI need to be able to contribute to “a more diverse, fair, and equitable society” (AI and Inclusion Symposium, 2017, p.2) and encourage and foster the diffusion and actual use of AI for promotion of societal development (Vinuesa et al., 2020). Thus, ethical AI will increase trust, and increased use of AI. Safe and ethical AI will also help firms<sup>12</sup> to avoid costly mistakes and so facilitate the uptake and diffusion of AI (Baker-Brunnbauer, 2020). Thus, HCAI approaches aim to establish the societal and commercial requisites for the greater use of AI (Floridi et al., 2018).

Eitel-Porter (2020, p.3) argue that most proposals for HCAI have similar pillars, namely fairness,<sup>13</sup> accountability, transparency, explainability and privacy. And according to Floridi et al. (2018, p.689) the ethical principles<sup>14</sup> proposed by the *AI4People* initiative are of beneficence, non-maleficence, autonomy, justice and explicability. In June 2019 the *G-20* proposed five *Human-Centered AI Principles* (G-20, 2019). These principles require AI to be consistent with (i) inclusive growth, sustainable development and well-being, (ii) human- centered values and fairness, (iii) transparency and explainability, (iv) robustness, security and safety, and (v) accountability (G-20, 2019, p.4).

---

<sup>10</sup>Everitt et al. (2018) considers the safety challenges involved over the long-term in the case of an AGI. In this paper we are only concerned with the shorter and medium risks arising from narrow AI.

<sup>11</sup>Leike et al. (2017) poses eight questions or problems that need to be answered during the design of AI to ensure that it will be safe for humans, for example “How do we ensure that an agent behaves robustly when its test environment differs from the training environment?”

<sup>12</sup>It is no surprise that a growing number of multinational firms have been coming up with their own ethical AI principles. For example, Baker-Brunnbauer (2020) relates the case of German automotive company *Continental* who in June 2020 “announced its intention to develop a code of ethics for its internal development and usage of AI that is based on the EC Trustworthy AI guideline.”

<sup>13</sup>See Feuerriegel et al. (2020, pp.381-382) for a discussion of mathematical notions of fairness in AI.

<sup>14</sup>These are similar, with the exception of explicability, to principles of bioethics (Floridi et al., 2018).

Although the establishment of widely accepted and agreed on pillars and principles for ethical and responsible AI is a necessary and important milestone for the development and use of HCAI, it has a problem. The problem with this third remedy is, as we already stated in the introduction, that these principles and ethical guidelines lack strong incentives for developers and users of AI to adhere to them (Askell et al., 2019; Calo, 2017; Hagendorff, 2020). According to Hagendorff (2020, p.99) “Do those ethical guidelines have an actual impact on human decision-making in the field of AI and machine learning? The short answer is: No, most often not.” He concludes, from a review of the field of AI ethics that AI ethics “lacks mechanisms to reinforce its own normative claims.” Calo (2017, p.6-7) concurred with this assessment, noting that “... even assuming moral consensus, ethics lacks a hard enforcement mechanism. A handful of companies dominate the emerging AI industry. They are going to prefer ethical standards over binding rules for the obvious reason that no tangible penalties attach to changing or disregarding ethics should the necessity arise.” In fact, AI developers and users may under highly competitive conditions face the incentive to “underinvest in ensuring their systems are safe” (Askell et al., 2019, p.1). They may “skimp” on AI ethics also when lured by the potential winner-takes-all effects of eventually inventing an AGI, as Armstrong et al. (2016) emphasized.

To incentivize the ethics for HCAI, Eitel-Porter (2020, p.1) argues for a strong, mandated governance controls including tools for managing processes and creating associated audit trails.” Floridi et al. (2018, p.704) argues that governments should “incentivise financially the inclusion of ethical, legal and social considerations in AI research projects.” This then, is a key challenge facing HCAI: how to incentivize adherence to ethical principles? In the rest of the paper, we argue that the nature of this challenge is such that public procurement and innovation is very relevant.

In the next section, we argue that the literature on innovation policy suggests that public procurement can make an important contribution towards steering society towards its goals, in this case, the goal of a HCAI.

### **3 Innovation Policy and Artificial Intelligence**

In the previous section we came to the conclusion that it is unlikely that HCAI will be forthcoming automatically from the market. This requires government intervention, and in particular, through steering the incentives for the private sector to develop and disseminate

AI models that conform to the ethical and responsibility requirement of HCAI. For this, innovation policy will be required. In this section we discuss how directional innovation policies may be appropriate and effective, what the contribution of one specific innovation policy – innovation procurement – may be, and finally how public procurement and innovation, and specifically public procurement for innovation may be utilised.

### 3.1 Directional innovation policies and AI

According to Edler and Fagerberg (2017, p.4) “Innovation is understood as the introduction of new solutions in response to problems, challenges, or opportunities that arise in the social and/or economic environment.” According to Bloom et al. (2019, pp.167-168) the empirical evidence from the literature suggests that social returns to innovation “are much higher than private returns, which provides a justification for government-supported innovation policy.” More generally, an innovation such as AI generates both positive and negative externalities and markets do not adequately capture these in market prices. As such there is, as Korinek (2019, p.4) points out “no theoretical reason to believe that the free market will direct innovative efforts to the most socially desirable innovations.... the market may thus guide innovation in the wrong direction.” AI innovation will thus not automatically result in HCAI.

The implication is that HCAI is a challenge for innovation policies. Innovation policies are the “public interventions to support the generation and diffusion of innovation, whereby an innovation is understood as the transformation of an invention into marketable products and services, the development of new business processes and methods of organization, and the absorption, adaptation and dissemination of novel technologies and know-how” (WTO, 2020, p.24).

There are many innovation policy tools aimed at the “generation and diffusion of innovation,” from both the supply and demand side.<sup>15</sup> Tools such as R&D subsidies and tax credits have typically been described as supply-side tools (Edler and Fagerberg, 2017), whilst regulations and public procurement have been described as demand-side tools (Aschhoff and Sofka, 2009). In recent years there has been subtle shift towards the increasing use of demand-side tools, in particular as these are seen as being potentially useful for industrial policy purposes and to facilitate economic restructuring (Crespi and Guarascio, 2019; Edler and Fagerberg, 2017). This growing use is part of what is termed the “directional turn” in innovation policy – not to encourage innovation for the sake of innovation, but for addressing a pressing societal

---

<sup>15</sup>See e.g. Edler and Fagerberg (2017) who presents and discuss 15 major innovation policy instruments.

need or challenge (Miller and Lehoux, 2020).

Present concerns with directional innovation and industrial policies are based on recognition that government intervention was in the past important for the development of important technologies, such as “jet engines, radar, nuclear power, the Global Positioning System (GPS), and the internet” (Bloom et al., 2019, p.166). Clemens and Rogers (2020) studied how government procurement influenced innovation in prosthetics during the U.S. Civil War and World War I, conclude that the nature of government procurement can influence significantly whether innovation will be on costs or quality – which is a finding with much relevance for the challenge of HCAI which is primarily a challenge relating to the quality of the technology. Another oft-quoted example of government steering the development of new technologies is of the USA’s efforts during the Second World War, which started with the creation of the National Defense Research Committee (NDRC) in 1940. These efforts resulted in many impactful new technologies and their diffusion during and after the war, including radar, mass-produced penicillin, radio communications, and pesticides such as DDT (Gross and Sampat, 2020). After the Second World War, three notable innovations where government incentives were important, included the creation of the internet through the USA’s DARPA-program,<sup>16</sup> the creation of modern biomedical research through funding grants in the UK, and the creation of Google’s search engine algorithm by a government grant (Lenderink et al., 2019). Interesting discussions of the specific role of government procurement in the establishment of the computer industry, the semi-conductor industry, and the US commercial aircraft industry is contained in (Geroski, 1990).

Azoulay et al. (2018, pp.69-70) discusses the USA’s DARPA-program as a catalyst for many modern-day technologies, not only the internet but also such as the personal computer, lasers and Microsoft Windows. DARPA’s approach may be argued to be eminently suitable to the challenge facing the development and diffusion of HCAI. Fouse et al. (2020) discuss how DARPA, through its various initiatives, had an “outsized” influence on the development of AI, and continues to do so. For instance, DARPA is at the time of writing reported to be “investing more than US \$2 billion in AI through its initiative, AI Next” (Fouse et al., 2020, p.4).

As discussed by Azoulay et al. (2018) DARPA’s projects typically are focused on generating new technological solutions where three characteristics of the technology and markets overlap, namely 1) there has to be a societal challenge; 2) the technology must be new and on the

---

<sup>16</sup>DARPA is the acronym for the USA’s Department of Defense’s “Defense Advanced Research Projects Agency” which was launched in 1958 and first called the Advanced Research Projects Agency (ARPA) (Azoulay et al., 2018).

beginning of the technology *S-curve*, and 3) there are significant frictions in the markets for ideas and technology that will hinder a spontaneous market solution. In the case of HCAI, all three these conditions are present, as it presents a significant societal challenge, the technology is new (in particular the alignment with human values) and the high degree to which the technology, being data-based, are subject to knowledge spillover and data network effects. These three conditions are also amenable to be influenced through public procurement for and of innovation as we will discuss in greater detail in the next sub-section. For instance, public procurement can effectively stimulate from the demand-side the production of socially desirable technologies (Miller and Lehoux, 2020); it is better suited to technologies that are still early in the product-life cycle (such as AI) (Geroski, 1990); and it overcomes knowledge spillover externalities by creating new markets and networks.

In the next sub-section, we discuss the emerging literature on public procurement and innovation.

### 3.2 Public procurement and innovation

Public procurement refers to “the direct purchase of goods and services by the public sector” (Crespi and Guarascio, 2019, p.783). The potential impact of public procurement is enormous. Gerdon and Molinari (2020) document for instance that there are more than 250,000 public authorities in the EU alone, who spent more than EUR 2 trillion annually on procurement.

Public procurement in general, has long since been a tool to promote innovation (and industrialization). Most often in the past, it has been used *ex post* to support new innovations by creating a demand for the product (being a customer for new products). This is also the way in which Lin (2020, p.26) conceives of the relationship between public procurement for innovation support. However, this is a rather restricted view: through the more targeted approaches of public procurement for and of innovation, governments can also *ex ante* support research and development of goods and services that do not yet exist – for present purposes, HCAI solutions. In this government can fulfil the function of a lead user role in innovation, as for instance argued by von Hippel (1976), and it can also create new networks and assist in the diffusion of innovations for instance by helping SMEs to adopt and use AI<sup>17</sup>

---

<sup>17</sup>In the case where the public sector procures a new innovation for its own use, it is termed *intrinsic procurement* and when it is procured for use outside of the government, it is termed *extrinsic procurement* (Czarnitzki et al., 2020).

Czarnitzki et al. (2020). There has been a growing interest in recent years by governments to support innovation through public procurement – the WTO (2020, p.67) documents that 81 percent of OECD countries have adopted initiatives to stimulate innovation through public procurement.

Public procurement and innovation (PPaI) are, as we explained in the introduction, a term encompassing three modes of innovation from the perspective of public procurement. As discussed by Obwegeneser and Müller (2018, p.5), the first is public procurement for innovation (PPfI), which deals with the question, how can public procurement drive innovation? This refers broadly to all public procurement initiatives aimed at innovation. There are many types and modalities of public procurement for innovation depending on focus, use, output, and interaction with suppliers (see Lenderink et al. (2019)). A full discussion of the perturbations and typologies of procurement of and for innovation falls outside the scope of this paper – the reader is referred to Lenderink et al. (2019) for an extensive overview.

For the present purposes though, public procurement for innovation includes procurement of products and services that do not yet exist. In the European Union the explicit use of public procurement in order to procure such products or services is relatively recent (Czarnitzki et al., 2020). Edler and Georghiou (2007) relate how the public procurement for innovation, after having been neglected in the EU, got onto the EU's agenda around 2003. This eventually resulted in the European Commission's *Handbook on Public Procurement for Innovation* in 2007. One of the innovations in public procurement that the EU introduced herein was the legal instrument of *Pre-Commercial Procurement of innovation* (PCP). PCP “concerns the procurement of R&D services prior to commercialisation, where new solutions for a specific social need or challenge are developed in competition with risk-benefit sharing between the public organisation and potential suppliers” (Lenderink et al., 2019, p.8). PCP is “a competitive and selective” instrument, as firms have to compete and come up with the best solution on their own (Aschhoff and Sofka, 2009, p.1236), and it “emphasizes the need for new development of a solution to solve a specific requirement” (Obwegeneser and Müller, 2018, p.11).

The second mode of innovation from the perspective of public procurement is public procurement of innovation (PPoI). This deals with the question, how can public services be innovated? For instance, government departments such as health, education and energy, to name but a few, could increase efficiency through adopting AI solutions, for instance AI diagnostic tools for hospitals and energy monitoring systems for reducing emissions from government buildings. A third mode of innovation from the perspective of public procure-

ment is innovative public procurement (IPP). It deals with the question how can public institutions procure innovatively? The demands of steering societal outcomes through public procurement, including to steer advanced technologies such as AI, will require that public procurement itself be conducted in a more advanced manner – that the public sector innovate itself in how it procures innovation. This is not least due to the complexity of AI and the need for government agencies to understand the technology, but also due to the need to continually monitor and upgrade AI technologies, which are never static, a result of AI models constantly learning and changing behavior as they get exposed to more data.

PPAI have been shown to be an effective tool to steer innovation.<sup>18</sup> Czarnitzki et al. (2020) studies the case of Germany, where the government already since 2009 allowed for public procurement of innovation. They find from an analysis of 3410 cases, between 2010 to 2012, that the use of this instrument by the German government “increased turnover with new products and services in the German business sector by EUR 13 billion in 2012, which represents 0.37% of GDP. Standard procurement tenders without innovation-related components, by contrast, show no detectable relationship” (Czarnitzki et al., 2020, p.2). Moreover, based on calculations from German public procurement and R&D spending, they conclude that “The quantitative potential of public procurement of innovation is about ten times larger than the amount of public R&D subsidies distributed to the business sector” (Czarnitzki et al., 2020, p.3). Other interesting studies of public procurement and innovation include Miller and Lehoux (2020) who study the use of public procurement of innovation in the Canadian healthcare sector.

We can also report examples of how government procurement is already having an impact on the development and diffusion of AI. Simonite (2020) reports that in the USA, the Centers for Medicare and Medicaid Services (CMS) will facilitate the dissemination of particular AI models for use in medical diagnostics in US hospitals. Through this, the government is paying the health service to use particular AI algorithms. Beraja et al. (2020) studies the case of facial recognition AI in China and notes the significant role that the government’s steering of the technology played. They argue from this that because data is central to AI, and furthermore that “the state is a key collector of data” as well as that there are economies of scope from sharing data across firms, that the government is well positioned to stimulate AI innovation. Specifically, they find that “following the receipt of a government contract to supply AI software, firms produce more software both for government and commercial

---

<sup>18</sup>Geroski (1990, p.189) argues that PPI can be even better than R&D subsidies and tax credits to steer innovation, because “procurement programmes link the production of innovations to their use” while “subsidies focus only on the production of innovations.”

purposes when the contract provides access to more government data” (Beraja et al., 2020, p.1).

Finally, Stojčić et al. (2020) argue that public procurement of and for innovation can help build innovation capability, unlike R&D grants and tax credits which depends on existing innovation capability. In the novel case of AI ethical embeddedness, the innovation capability may not be broadly present amongst firms. Therefore, the suitability for public procurement of and for innovation in the context of steering human-centered AI is particularly appropriate.

### **3.3 How can public procurement and innovation steer AI?**

We have noted on a few occasions in the preceding sub-sections why public procurement of and for innovation (PPoI, PPfI) can be relevant and appropriate for the steering of AI towards HCAI. Having established the why of public procurement for and of innovation, we will deal in the rest of this section with the how. We will in particular focus on the use of PCP.

For ordering the discussing it is useful to start with the salient features of PCP, and then indicate what elements the contracting authority should emphasize in the case of steering towards HCAI, where after we will note some additional requirements – innovations in procurement – that may be particularly useful for the success of PCP with respect to human-centered AI.

#### **3.3.1 Pre-Commercial Procurement of Innovation**

First, we discuss the salient features of PCP as a tool to steer HCAI. Pre-Commercial Procurement (PCP) for innovation is the very first European Union (EU) legal provision, available to the public sector, to steer and procure innovative solutions, which are not available in the market, and that require R&D activity. As such, it is very relevant to steer innovation towards human-centered AI. The 2007 PCP Communication was followed by additional, clarifying, documents released to specify its correct interpretation and for its implementation. Yet the main message has been clear from the very beginning, namely that PCP can be used to procure R&D services only, needed to develop an innovative solution which is unavailable in the market. More specifically, PCP can be used to develop prototypes, up to few units of the final product, to make sure the solution is as desired by the procurer. However, PCP

cannot be used by a contracting authority (CA) to procure the needed quantity of the final product. Indeed, to do so, at the end of PCP the CA should follow-up with a commercial procurement, opening a competitive tendering, or some type of negotiated, procedure within the legal framework defined by the 2014 EU Public Procurement Directive (EC, 2014). In such procurement any eligible firm could compete for the commercial contract, and not just those which participated to the PCP.

The underlying idea, behind the separation of the R&D phase from the commercial procurement phase, is that the European Commission (EC) did not want to restrict the commercial procurement of potential innovative solutions only to PCP participants. As a matter of fact if, while a PCP procedure is taking place, other firms in the market developed interesting new solutions they should be allowed to participate in the commercial procurement of the final product. This would be compliant with the principle of fair treatment with respect to potential participants, with the principle of open competition, in the best interest of CA and of the entire society.

After more than a dozen years from its introduction PCP is gradually diffusing across EU countries, and AI based solutions are naturally being proposed within such purchasing procedures. A distinguishing feature of PCP is that the initial need is formulated directly by the public authorities, as much as the US Department of Defence (DoD) does it with DARPA, which implies that the solutions must satisfy technical but also ethical requirements, in compliance with the mission of the State. For this reason, AI based solutions originated within the PCP would avoid innovations that are potentially harmful for the society. The 2007 Communication elaborates on the fact that PCP should be founded on a risk sharing principle, between companies and CA, concerning the possibility of project failure. Furthermore, EC suggests that as an incentive to participate in the PCP, the Intellectual Property Rights (IPR) behind the new solutions should be left with the companies invited by CA, rather than being appropriated by CA. Though the reason behind the suggestion is clear, with some AI based innovative solutions contracting authorities may consider keeping the IPR, in exchange of a compensation to the companies. This could take place when the contracting authority wants to have under its own control the diffusion and continuous updating of the new product, which in the case of ethical and responsible AI solutions that it wants to disseminate to SMEs and update over time (as was shown to be necessary in section 2.1) would be necessary.

Article 31 of the 2014 Public Procurement Directive the EC introduced an additional legal provision for procuring innovative solutions, the so called *Innovation Partnership* (IP). Unlike

PCP in the IP the contracting authority may not unbundle the procurement of the R&D phase from the commercial phase, that is the procurement of the desired quantity of the final product. As a result, currently public officers in EU have two alternative legal instruments to procure innovative solutions, which have not been developed yet and require R&D activity: PCP and IP.

### **3.3.2 Using PCP to steer innovation towards HCAI**

In what follows we discuss some of the main elements of PCP and case studies to articulate how public procurement can steer AI solutions in EU public sector. In the 2007 Communication the EC suggests a model procedure for PCP composed of three phases to implement the PCP. Contracting Authorities (CA) are by no means mandated to follow such model, but it can certainly be a very helpful benchmark for them. A PCP is typically preceded by one, or more, sessions of market consultation. These sessions are publicly announced and organized by CA to inform all the potentially interested subjects of its needs and of the problem – e.g., the need for an AI solution that is consistent with certain ethical principles - that will be the object of the PCP.

Then a call for manifestation of interest to participate in a PCP for an ethical AI solution is openly announced. After having received the initial manifestation of interest CA can opt to invite a restricted number of companies for the first phase. In such phase CA asks the invited companies to propose an initial draft of the would-be final solution. Once received the proposals, a subset of them will be selected for the second phase where a prototype of the solution will be presented. Then, a subset of firms which submitted a prototype will be invited to the third phase, where the selected companies are asked by CA to produce a few units of the final solution, to check if it works as desired. To prevent the emergence of dominant/unique solutions, the 2007 Communication suggests that at least two companies should be invited to the third phase.

As said, if the solution is satisfactory, and the CA interested in procuring a solution such as the one emerged from the PCP, then after the conclusion of the PCP the contracting authority will have to open a follow-up procurement procedure, for products which no longer requires R&D. For this reason, such procedure would fall under the 2014 Public Procurement Directive, and the CA will now be able to describe precisely in the tender documentation what it wishes to purchase. Of course, in this case, those companies which were able to reach the third phase of the PCP will have some advantage, but no certainty to obtain the

contract. Still in this case, since the procured solution would be innovative, according to the EC terminology it will be called PPI, that is public procurement of an innovative solution, which however does not require R&D activity. It is still considered innovation procurement because it purchases an object which was not in the market yet.

The above short description should clarify how AI based innovative solutions could be driven in the desired direction by the public sector, in all the PCP phases. From the original formulation of the problem, to the gradual screening and selection of proposals taking place across the three phases, until PCP ends, CA can navigate the whole process towards solutions which are beneficial for the society. In this navigation process, a number of elements are particularly relevant for human-centered AI solutions – and some of them may require innovation in public procurement.

The first is that the CA should from the outset approach the use of PCP for an ethical, human-centered AI solution with the intention to foster cooperation between firms and research organizations, as well as and foster multi-disciplinary cooperation. According to Askell et al. (2019, p.1) “competition between AI companies could decrease the incentives of each company to develop responsibly by increasing their incentives to develop faster. As a result, if AI companies would prefer to develop AI systems with risk levels that are closer to what is socially optimal- as we believe many do- responsible AI development can be seen as a collective action problem.” As said, the PCP benchmark is based on a procedure with sequential selection of companies, hence in principle competitive. Yet, despite this, in using PCP, the CA can design calls for HCAI solutions that requires and facilitates such cooperation. Askell et al. (2019) discuss various ways in which this can be incentivized, for instance by creating high shared gains from cooperation and penalizing non-cooperation. Finally, if the PCP model turns out to be unsuitable for convincingly establish cooperative behaviour among invited firms, then the Innovation Partnership could be used instead as legal instrument.

The second is that the CA should insist on ethical design approaches be followed through all the stages of the procurement process. In this respect Crawford and Calo (2016) proposed two broad approaches to the ethical design of AI, which may also be applied by a CA using PCP. The first is to require developers of AI solutions to use frameworks such as value sensitive design and responsible innovation, and the second is to require a social-systems approach to building AI, which would entail consideration of “the social and political history of the data” (Crawford and Calo, 2016, p.313). The latter approach would be in line with the need for data justice as discussed below.

The third is that the CA should promote greater openness in AI development – for instance by requiring open scrutiny of solutions that it purchases and requiring certain degrees of openness and outside participation in and evaluation of the solutions developed as part of the procurement action. Bostrom (2017) concludes that over the medium-term, greater openness in AI development would be unambiguously positive by speeding up the development and diffusion of AI.<sup>19</sup> He is more concerned that over the long run, the openness would discourage innovation to the extent that the effort may be motivated by having a monopoly in the market from a proprietary innovation. However, through PPI this long-run disincentive effect of openness may be compensated for by government not only be reducing the costs and risks of developing a new AI solution, but also offering a firm the chance to build its capacity for developing and using AI, and even for a company to benefit in complementary areas to the AI being procured (Bostrom, 2017).

Fourth, and related to the need for openness, is the requirement that the PCP call specifically take care of the reliance of AI models on large datasets. Beraja et al. (2020) for instance distinguish between “data-rich versus data-scarce government contracts.” We want to argue that PCP contracts for HCAI should be of the data-rich variety. A data-rich PCP contract should in particular address the *data parity problem* (Calo, 2017) which refers to the fact that “a few large firms have disproportionate access to big data necessary to train AI models. Mostly, small businesses do not have the scale and scope and resources to gather data and benefit from data network economies. This could mean that machine learning applications will bend systematically toward the goals of profit-driven companies and not society at large” (Calo, 2017, p.19).

The data parity problem is mirrored in society as a digital divide, with data gaps, and data absences creating and perpetuity inequality in access to the use and knowledge of AI, and in benefiting from AI. In recent times the imperative of addressing these as a requirement for HCAI has resulted in the notion of data justice, which goes beyond data gaps and data absences to recognition of structural inequalities. According to Dencik et al. (2019, p.874) “The framework of data justice broadens the terms of the debate in a way that accounts for a host of issues that are compounded in the datafied society, as evidenced in recent scholarship relating to democratic procedures, the entrenchment and introduction of inequalities, discrimination and exclusion of certain groups, deteriorating working conditions, or the dehumanisation of decision-making and interaction around sensitive issues. These

---

<sup>19</sup> According to Bostrom (2017, p.137) “openness would improve static efficiency, by making products available at marginal cost (e.g., in the form of open-source software) and allowing a given level of state-of-the-art technical capability to diffuse more quickly through the economy.”

discussions suggest a need to position data in a way that engages more explicitly with questions of power, politics, inclusion and interests, as well as established notions of ethics, autonomy, trust, accountability, governance and citizenship.” The discussion of data justice and the relevance of societal concerns about fairness, diversity and equity in the application of potential general-purpose technologies such as AI reflects the fact that ultimately, AI is sociotechnical in that the value that it offers “depend on not only technical hardware but also human behavior and social institutions for their proper functioning” (van de Poel, 2020, p.7).

Fifth, the contract authority (CA) should make sure that all documentation<sup>20</sup> and audits are captured, as this “is important to provide an audit trail in case of subsequent issues with the model” (Eitel-Porter, 2020, p.4). This is relevant not only for adjudicating the quality of the solution offered, but to support efforts to continually review the solution so as to ensure that “ethical parameters are not breached over time” (Eitel-Porter, 2020, p.5). This is not so much due to the ethical norms changing, but because of the nature of machine learning (ML), which adapts and changes as it learns more (gathers more data). The important point is that ethical AI solutions may not automatically remain ethical given the possibility of learning, due to AI essentially being an intelligent agent (Riedl, 2019). Anderson and Anderson (2007) refer to this feature of AI to argue that the ultimate challenge is to create explicit ethical machines, which are able to learn and formulate “its own ethical choices based on its own set of principles.”

Having ensured that the above five elements are followed, the CA purchasing an AI solution would need finally to be able to verify that the prototype solution offered during stage 3 of the PCP process conforms to the specifications set out in the call – in other words, the technical AI solution needed as well as the requirement that ethical principles are observed. In this regard, it is important to set up adequate measures for verification. As Brundage et al. (2020, p.1) stresses “The capacity to verify claims made by developers, on its own, would be insufficient to ensure responsible AI development. Not all important claims admit verification, and there is also a need for oversight agencies such as governments and standards organizations to align developers’ incentives with the public interest.” Values<sup>21</sup> in AI models need to be “intended, embodied, and realized values” (van de Poel, 2020, p.5). For present purposes, drawing on van de Poel (2020, p.9) we can state that for a procured AI solution to embody a particular value, denoted by  $V$ , it should be the case that at least two conditions

---

<sup>20</sup>This should include all documentation including AI impact assessments (Gerdon and Molinari, 2020).

<sup>21</sup>See the discussion in van de Poel (2020) on the meaning of values and how to assess whether the desired values are indeed intended, embodied and realized.

be met: first, the AI must be designed for  $V$ , and second that in using the AI solution, the values  $V$  would be promoted. Thus, AI models need to be designed in order to meet a particular value, for instance human health, and second, utilising the AI model should then indeed promote that value, e.g. human health.

What measures can be used for verification and assessing whether desired values are intended, embodied and realized? Brundage et al. (2020) discuss: third party auditing, red teaming exercises, the piloting of bias and safety bounties, and the sharing of information about AI accidents. In all of these actions, governments can benefit from having independent oversight bodies tracking the work of contractors, and subject these to risk-analyses and public scrutiny. Tzachor et al. (2020) advocates the use of “red teaming” by oversight bodies to stress test any AI solutions that contractors may come up with, where “red teaming is a way of challenging the blind spots of a team by explicitly looking for flaws from an outsider or adversarial perspective” (Tzachor et al., 2020, p.366). According to Tzachor et al. (2020) one legacy of the COVID-19 pandemic is the realisation that circumstances can require the application of new AI solutions under extremely short time horizons and with urgency, and that “doing ethics with urgency” becomes important. Such doing ethics with urgency will require according to the authors that government should be able to “rapidly conduct robust testing and verification of systems.” (Ibid, p.366)

Even if these various measures and methods for verification and assessment is successful in finding that the conditions for the AI solution to embody human-centered values are met, a further problem is that as was mentioned, AI models may have unintended results. Moreover, value may change over time. To deal with such unintended results and changes in values, require continuous involvement and monitoring by the contracting authority. According to van de Poel (2020) the design of technologies can never remain static, and they may need to be redesigned at some future stage in light if possible unintended effects as value as possible shifts in values.

At this stage, once the CA is satisfied with the final prototype of the HCAI solution from the PCP process, it will have to ensure that the public sector more broadly has in place complementary mechanisms for dealing with remaining safety risks posed by AI. As we pointed out in section 2.2.1, normal accidents are one of the salient risks posed by AI. It is unfortunately the case that the risks of accidents “cannot simply be ‘designed out’ of the technology -at least not without giving up on many of their benefits” (Maas, 2018, p.5). Maas (2018) concludes that this requires two complementary approaches: one, that governments ensure that disaster and accident insurance schemes are appropriate to cover AI risks, and

second that more research be encouraged into AI safety and AI ethics - see also Anderson and Anderson (2007).

Finally, for the PCP process to be ultimately successful in contributing to innovations in the field of AI, the government and its contracting authorities will need to upgrade and continually invest in its own AI expertise (Calo, 2017). With this they could even use data science and AI techniques within the public sector to explore and test the ethical consequences of AI. For example, Yu et al. (2018) survey various technical solutions to support evaluation of ethical concerns in AI. These include useful resources and initiatives on which the public sector's scientists and advisors could draw, such as the *GenEth* ethical dilemma analyser that can be used to explore ethical dilemmas; the *MoralDM* tool helping to resolve ethical dilemmas through first-principles reasoning and analogical reasoning; and crowdsourcing of self-reported preferences such as through the *Moral Machine* project.<sup>22</sup>

## 4 A Model of Public Procurement and Responsible AI Innovation

As we outlined in section 2.3, a contracting authority (CA) in the EU could use the provisions of the EU's PCP instrument to procure an AI-solution that conforms to high standards of ethical and responsible AI, as was identified in section 2.1. By issuing an explicit call for such a solution and offering to fund the R&D costs of the successful company, the public sector could incentivize the development and diffusion of human-centered AI (HCAI). This incentive is important, in the absence of which companies may skimp on adequate ethical and safety standards, as we argued in the introduction and section 2.1.

We can now formalize these ideas. First, we can note that in the absence of the public sector's financial incentives, companies may not take the risks to provide the socially optimal AI solution. Let us suppose that a company  $X$  is willing, and has the competence, to develop an AI solution which could result in a socially desirable, human-centered AI system. The project requires resources to be invested and success in its development process is intrinsically uncertain. More specifically, if  $C > 0$  is the amount of money invested in the project by company  $X$  and  $\pi(C)$  the success probability, with  $\pi'(C) \geq 0$  and  $\pi''(C) \leq 0$ , then typically  $0 \leq \pi(C) < 1$  for any level of  $C$ . That is, no matter how large the company's effort, success will remain uncertain.

---

<sup>22</sup>See <https://www.moralmachine.net>.

Here  $\pi(C)$  formalizes the technology available to develop the project and exhibit non-increasing returns of scale.<sup>23</sup> Suppose now that, if development is successful and the project is commercialized that the company's revenue would be  $R$ . Therefore, its profit would be a random variable defined as (see Dimitri (2012)):

$$\Pi(C, R) = \begin{cases} R - C & \text{with probability } \pi(C) \\ -C & \text{with probability } 1 - \pi(C) \end{cases} \quad (1)$$

and the related expected profit is

$$E\Pi(C, R) = R\pi(C) - C \quad (2)$$

If the company maximizes (2) then the first order condition is

$$R\pi'(C) - 1 = 0 \quad (3)$$

and the optimal investment level  $C^*$  given by

$$\pi'(C^*) = \frac{1}{R} \quad (4)$$

as long as  $E\Pi(C^*, R)$  is non-negative. Expression (4) immediately suggests that, due to the concavity of the success probability, the larger the revenue the higher is  $C^*$ . However, there is no guarantee that, regardless of the value of  $R$ , that the company would invest at all. Indeed, a non-negative expected profit requires that

$$\pi'(0) > \frac{1}{R} \quad (5)$$

which if not satisfied would discourage the company to undertake any investment.

For example, suppose  $\pi(C) = a(1 - e^{-bC})$  with  $0 < a < 1$  and  $b > 0$ . It is immediate

---

<sup>23</sup>In reality, several technologies exhibit increasing returns of scale, at least for some investment levels. However, to keep the exposition simple, with no major loss of generality, in what follows we do not consider this possibility.

to see that the larger are both,  $a$  and  $b$ , the higher is  $\pi(C)$  and so the less challenging is the development process of the AI project. Interestingly the parameters  $a$  and  $b$  play two different roles. Indeed while  $a = \lim_{C \rightarrow \infty} \pi(C)$ , that is it defines the maximum level that success probability can take,  $b$  only defines how  $\pi(C)$  changes as  $C$  changes. This can be seen by considering for example the elasticity of the function  $\pi(C)$ , namely the percentage change in  $\pi(C)$  due to 1% change in  $C$ , which is given by

$$\frac{\pi'(C)C}{\pi(C)} = \frac{be^{-bC}C}{(1 - e^{-bC})} \quad (6)$$

As this is independent of  $a$ , (5) would become

$$ab > \frac{1}{R} \quad (7)$$

which suggests that the company decides to invest if the project is sufficiently rewarding, or technically not too difficult, or both. In this case the optimal level of investment  $C^*$  would be

$$C^* > \frac{\log(abR)}{b} \quad (8)$$

which, consistently with (7), is positive for  $abR > 1$ .

However, suppose now that (7) is not satisfied and the company decides not to invest. Failure to develop the AI project may damage the society and so  $A$  could consider supporting the company, however only if the ethical level of the solution is sufficiently high. Broadly speaking, suppose that company  $X$  has to compete for funding with other companies as per the EU's PCP, and that the ethical level of the solution under development, summarized by the indicator  $e \geq 0$ , is positively related to the probability  $\theta(e)$  of being funded by  $A$ . We assume  $\theta(e = 0) = 0$  if differentiable and  $\theta'(e) > 0$  and that  $\lim_{e \rightarrow \infty} \theta(e) = 1$ . Finally, we now assume that the R&D total costs for the company are given by  $C + ce$ , where  $ce$  is total cost due to reaching the ethical level  $e$  and  $c > 0$  is the marginal cost for doing so. Under these assumptions, in the analysis conducted so far  $X$  undertook no, or very low, investment to take care of the ethical features of the solution under development. Therefore, company  $X$  needs to decide the level of both  $C$  and  $e$ , and meticulously document  $e$  as per the CA's requirements.

For this, suppose  $A$  and  $X$  agree to write the following contract: “if the AI project is funded by  $A$ , with probability  $\theta(e)$ , and successfully developed then  $A$  pays to  $X$  the sum  $p$  while if the project fails to be developed then  $A$  pays to  $X$  the sum of  $q$ , with  $p > q$ . Moreover, the intellectual property rights will remain with  $X$ , which will still enjoy reward  $R$ ”. As we shall see below, such admittedly simple contract will induce the company to try developing a solution that contains consideration of the ethical requirements.

Before proceeding, it is worth noticing that conditional on  $A$  funding  $X$  the above contract operates as a mixture of pull ( $p - q$ ) and push ( $q$ ) incentives. That is, although we assume  $A$  will only pay at the end of the project, as it may be with the EU’s PCP,  $q$  are guaranteed to  $X$  and it is as if they would be paid upfront, hence as a push incentive. For this reason we could also imagine that should  $X$  lack resources for engaging in the process then  $A$  may indeed pay upfront the sum  $q$ . Nonetheless, the additional sum  $p - q$  will be paid only conditionally upon success of the project, hence as a pull incentive.

Then the company’s profit  $\Pi(C, R, p, q, c, e)$  is defined as the following random variable:

$$\Pi(C, R, p, q, c, e) = \begin{cases} (p + R) - C - ce & \text{with probability } \theta(e)\pi(C + ce) \\ q - C - ce & \text{with probability } \theta(e)(1 - \pi(C + ce)) \\ 0 & \text{with probability } 1 - \theta(e) \end{cases} \quad (9)$$

Therefore, its expected profit will be

$$E\Pi(C, R, p, q, c, e) = \begin{cases} \theta(e)\pi(C + ce)(p + R - q) + \theta(e)(q - C - ce) & \text{if } e > 0 \\ 0 & \text{if } e = 0 \end{cases} \quad (10)$$

and the optimal level of investment  $C^{**}$  chosen by the firm would now be such that

$$\pi'(C^{**} + ce) = \frac{1}{(R + p - q)} \quad (11)$$

Condition (11) looks like (4) except that the right-hand side now has the additional term  $p - q > 0$  at the denominator. This basically increases  $X$ ’s expected revenues providing a stronger incentive for the company to invest resources and engage in developing a solution, provided it is ethically acceptable to  $A$ . This will now take place if

$$\lim_{(C+ce) \rightarrow 0} \pi'(C + ce) = ab > \frac{1}{R + p - q} \quad (12)$$

Condition (12) indicates that  $A$  could always find a large enough amount  $p - q$  such that (5) is satisfied. Moreover, notice that on the left hand side of (12) we are taking  $\lim_{(C+ce) \rightarrow 0} \pi'(C + ce)$  because expression (10) is equal to 0 for  $e = 0$ , and so  $\pi'(0)$  would not even be defined. Moreover, with an acceptable estimation of  $\pi(C)$ , authority  $A$ , from (11), could also determine the effort  $(C^{**} + ce)$  of  $X$ .

Considering again the previous example  $\pi(C) = a(1 - e^{-bC})$ , the optimal investment level for  $X$  would now be

$$C^{**} + ce = \frac{\log(ab(R + p - q))}{b} \quad (13)$$

If  $C^{**} + ce > 0$  it is interesting to notice that  $\frac{\log(ab(R + p - q))}{b} < p - q$ , namely the company will invest less than the additional incentive term  $p - q$ , for all levels of  $p - q$  if  $abR < 1$ , that is if the company was not investing at all before  $A$ 's intervention. However, if  $abR > 1$  then the inequality would be true for  $p - q > z$ , with  $z$  solving  $\frac{\log(ab(R + z))}{b} = z$ .

So, the element  $p - q$  will work as a co-funding term for  $X$  to engage in the AI project. To summarize, what is important for  $A$  to motivate  $X$  to invest more in R&D activities are not the values of  $p$  and  $q$  separately, but rather only their difference. Indeed, this is true even if  $q = 0$ , as long as  $p$  is large enough.

It is worth stressing that the optimal level of R&D investment obtained from (13) must include the component  $ce > 0$ , needed by  $X$  to receive funding from  $A$ , which represents the expenditure undertaken by  $X$  to introduce ethical features into the solution. The desirable level of the indicator  $e$  resolves a trade-off between increasing the probability of being funded by  $A$ , as well as of finding the innovative solution, and of increasing the costs. It would be found by maximising (10) also with respect to  $e$ .

To summarise, the above equations describe a simple framework where public procurement can induce HCAI solutions, by conditioning the funding of a company to the presence of ethical features, for example in the outcome of a PCP as well as of Innovation Partnership procedures. This use of public procurement to promote HCAI as we modelled it here, is complementary to the approach in Naudé and Dimitri (2020), who model the use of public

procurement to reduce the likelihood of arms races for AI.

The mechanism that we outlined above for incentivizing R&D for HCAI is relevant not only for addressing the general reluctance of companies to invest in HCAI, but more generally the fact that the concerns about technologies such as AI's have led to increasing calls in recent years to make "Technology Impact Assessments" (TIAs) obligatory for entrepreneurs before implementing and investing in new technologies (Korinek, 2019). This will have particularly negative effects however, on innovation, over and above the impact of the risks already illustrated here. It may be worth pointing out that one reason why innovation in the digital economy have not been declining as innovation elsewhere, is due to the fact that until now, such innovation has been relatively unencumbered by bureaucratic influence, i.e., it has been rather more "permissionless." Rather than risk dis-incentivizing innovation, we want to argue that PCP, organised for example within a framework such as the one expressed by equations (9) and (10), can obtain the same results, but through rewarding entrepreneurs who do perform technology impact assessments. In other words, PCP may reduce the fixed costs which a requirement such as TIA can cause, by requesting the solution to incorporate the requested features, possibly co-funding the engaged companies for this.

Finally, given the potential of PCP to steer HCAI as this simple model show, we can conclude this section by pointing out that although public procurement for innovation as instrument to steer AI is still relatively neglected, there are already a small sample of PCP AI based undertaken in EU since 2007 when PCP was launched. It is useful to briefly mention these projects, and to recommend further research to document the lessons learned from them.

A first example is that during the years 2015-16 a consortium of five hospitals, in different European countries implemented an EU funded PCP project to develop a highly interoperable, AI based, telemedicine system for the cockpit of intensive care units in hospitals. The project, called *THALEA*, was successfully completed and as a result the Dutch Company called *NewCompliance*, which developed its solution for the operating room cockpit, was able to attract venture capital and sell the solutions to several hospitals in EU and in the US. Moreover, it established partnerships with some major companies in the field to improve the solution, scaling up its business volume and size.

Among the PCP AI projects which have not been funded by the EU, a very timely and successful project is the PCP sponsored by the Danish Market Development Fund for self-driving-robots, based on ultraviolet light, to disinfect hospitals. The project was initially promoted in 2014, and completed in 2017, by a group of Danish hospitals, hence much earlier

than 2020 when the COVID-19 pandemic hit the world. As a result of the PCP project, the Danish company *Blue Ocean Robotics* attracted millions of dollars in venture capital, and sold its robots across several countries, to be used in hospitals across the world to help disinfecting hospitals during the COVID-19 pandemic.

Another interesting example, of a non-EU funded PCP, is an AI-based smart mobility PCP project developed in the period 2014-2016 and promoted by the Dutch Province of North-Brabant, to try solving shock-wave traffic jams problems in the area. Traffic jams often occur when a vehicle suddenly stops, inducing also the following vehicles to suddenly use their breaks even more strongly, originating traffic jams. As the outcome of PCP, the company *Be-Mobile* produced an application merging and analysing real time data on traffic, reported by various sources, data on vehicles speed, and to advise drivers on which roads/lanes to take. This innovative solution has been so successful that *Be-Mobile* was able to meaningfully scale-up in terms of employers and raised funds.

## 5 Concluding Remarks

There are high expectations of AI as a general-purpose technology that can contribute towards economic growth, innovation and sustainable development. There are, however, also concerns about the negative consequences and uses of AI. These negative consequences of AI have given rise to calls for public policy to ensure that AI-based applications are safe, and that the development and diffusion of AI does not suffer from improper use and misuse. What would count as improper use and misuse, and safety-risks such as that could be caused by accidents or unintended consequences of AI systems, would be determined by the extent to which AI systems are human-centered. Human-centered AI (HCAI) draws on ethical principles and puts forth actionable guidelines for reducing the risks mentioned. As such, HCAI is concerned with Ethical AI and Responsible AI.

In recent years there have been a growing number of proposals for Ethical and Responsible AI. A shortcoming of these is that they lack strong incentives for developers and users of AI to adhere to them. In this paper, we argued that public procurement and innovation (*PPaI*) is a potentially relevant tool with which to incentivize HCAI. Indeed, if the governance of AI deals with “how decisions are made about AI,” and what “institutions and arrangements could help in this decision-making (Dafoe, 2018, 2020), then public procurement and innovation as presented in this paper is a key element of AI governance – however an element that we

argued is still neglected.

We provided a simple theoretical model to show that the procurement for an innovative AI solution, underpinned by ethical considerations, can provide both the tools and audit trails as well as the financial incentives necessary to steer AI towards HCAI. Additionally, we highlighted five elements that the EU's PCP instrument should specifically take into account when aiming to steer HCAI, as well as complementary initiatives, such as supporting more multi-disciplinary research into ethics and into making ethics computable. This could inform further innovations in public procurement itself.

In this, we believe our paper made a number of contributions to the literature. First, our argument and model illustrating that public procurement for innovation can promote HCAI adds to the literature on innovation policy. This potential instrument for steering HCAI has been neglected in both the AI as well as innovation literature. Our paper contributed to addressing this neglect by providing arguments based on an analysis of the literature as well as providing a theoretical model. Secondly, our paper contributed to the literature on innovation policy by locating the role of public procurement for innovation of ethical, HCAI within the recent literature on mission-oriented and challenge-led innovation policies. Finally, we think our paper also contributed to the emerging literature of public entrepreneurship, in that the use of public procurement of innovation is an example of public entrepreneurship as it consists of innovation, it aims to contribute towards transformation of society, and it is subject to uncertainty.

## Acknowledgements

We are grateful to the participants of various workshops of the Research Unit for the Diffusion of Quality AI at Paderborn University for their comments on earlier versions of this paper, and in particular to Thomas Gries and Margarete Redlin. The financial assistance of the *Volkswagen Stiftung*, through planning grant *AZ 97042* from their project on *Artificial Intelligence and the Society of the Future* is gratefully acknowledged. The usual disclaimer applies.

## References

- AI and Inclusion Symposium (2017). An Evolving Reading List. Available at <https://drive.google.com/file/d/0ByG0FdgytPz7YkxRMHF1d19ibE0/view>.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete Problems in AI Safety. *ArXiv:1606.06565 [Cs]*, June:<http://arxiv.org/abs/1606.06565>.
- Anderson, M. and Anderson, S. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 28(4):15.
- Armstrong, S., Bostrom, N., and Schulman, C. (2016). Racing to the Precipice: A Model of Artificial Intelligence Development. *AI & Society*, 31:201–206.
- Aschhoff, B. and Sofka, W. (2009). Innovation on Demand: Can Public Procurement Drive Market Success of Innovations? *Research Policy*, 38(8):1235–1247.
- Askell, A., Brundage, M., and Hadfield, G. (2019). The Role of Cooperation in Responsible AI Development. *ArXiv*, July.
- Azoulay, P., Fuchs, E., Goldstein, A., and Kearney, M. (2018). Funding Breakthrough Research: Promises and Challenges of the ARPA Model. *Innovation Policy and the Economy*, 19:69–96.
- Baker-Brunnbauer, J. (2020). Management Perspective of Ethics in Artificial Intelligence. *AI and Ethics*, 11 September:<https://doi.org/10.1007/s43681-020-00022-3>.
- Beraja, M., Yang, D., and Yuchtman, N. (2020). Data-Intensive Innovation and the State: Evidence from AI Firms in China. *NBER Working Paper No. 27723, National Bureau of Economic Research*.
- Bloom, N., van Reenen, J., and Williams, H. (2019). A Toolkit of Policies to Promote Innovation. *Journal of Economic Perspectives*, 33(3):163–184.
- Boddington, P. (2017). Towards a Code of Ethics for Artificial Intelligence. *Heidelberg: Springer*.
- Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. *Oxford: Oxford University Press*.

- Bostrom, N. (2017). Strategic Implications of Openness in AI Development. *Global Policy*, pages 1–14.
- Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., and et al (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *ArXiv [cs.CY]*, April:https://arxiv.org/abs/2004.07213.
- Calo, R. (2017). Artificial Intelligence Policy: A Primer and Roadmap. *SSRN*, At https://ssrn.com/abstract=3015350.
- Castelvecchi, D. (2016). The Black Box of AI. *Nature*, 538:21–23.
- Clemens, J. and Rogers, P. (2020). Demand Shocks, Procurement Policies, and the Nature of Medical Innovation: Evidence from Wartime Prosthetic Device Patents. *NBER Working Paper No. 2667, National Bureau of Economic Research*.
- Crawford, K. and Calo, R. (2016). There is a Blind Spot in AI Research. *Nature*, 538:311–313.
- Crespi, F. and Guarascio, D. (2019). The Demand-Pull Effect of Public Procurement on Innovation and Industrial Renewal. *Industrial and Corporate Change*, 28(4):793–815.
- Czarnitzki, D., Hünermund, P., and Moshgbar, N. (2020). Public Procurement of Innovation: Evidence from a German Legislative Reform. *International Journal of Industrial Organization*, 71:1–16.
- Dafoe, A. (2018). AI Governance: A Research Agenda. *Centre for the Governance of AI, University of Oxford*.
- Dafoe, A. (2020). AI Governance: Opportunity and Theory of Impact. *Effective Altruism Forum*, 17th September.
- Dencik, L., Hintz, A., Redden, J., and Treré, E. (2019). Exploring Data Justice: Conceptions, Applications and Directions. *Information, Communication & Society*, 22(7):873–881.
- Dimitri, N. (2012). R&D Incentives for Neglected Diseases. *Plos One*, 7(12):e50835.
- EC (2014). On the Coordination of Procedures for the Award of Public Works Contracts, Public Supply Contracts and Public Service Contracts. *European Commission Directive 24. Brussels*.
- EC (2019). Ethics Guidelines for Trustworthy AI: High-Level Expert Group on Artificial Intelligence. *European Commission, Brussels*.

- Edler, J. and Fagerberg, J. (2017). Innovation Policy: What, Why, and How. *Oxford Review of Economic Policy*, 33(1):2–23.
- Edler, J. and Georghiou, L. (2007). Public Procurement and Innovation - Resurrecting the Demand Side. *Research Policy*, 36:949–963.
- Eitel-Porter, R. (2020). Beyond the Promise: Implementing Ethical AI. *AI and Ethics*, 6 October(<https://doi.org/10.1007/s43681-020-00011-6>).
- Etzioni, A. and Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *Journal of Ethics*, 21(4):403–418.
- Everitt, T., Lea, G., and Hutter, M. (2018). AGI Safety Literature Review. *International Joint Conference on Artificial Intelligence (IJCAI)*, arXiv: 1805.01109.
- Feldstein, S. (2019). The Global Expansion of AI Surveillance. *Working Paper, Carnegie Endowment for International Peace*, Sept.
- Feuerriegel, S., Dolata, M., and Schwabe, G. (2020). Fair AI: Challenges and Opportunities. *Business and Information Systems Engineering*, 62(4):379–384.
- Floridi, L. (2018). The Ethics of Artificial Intelligence. In Franklin, D. ed. *Megatech: Technology in 2050*. London: Profile Books. Chapter 13, pp. 155-163.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. (2018). AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28:689–707.
- Fouse, S., Cross, S., and Lapin, Z. (2020). DARPA’s Impact on Artificial Intelligence. *AI Magazine*, Summer:3–9.
- Frey, C. and Osborne, M. (2017). The Future of Employment: How Susceptible are Jobs to Computerization? *Technological Forecasting and Social Change*, 114:254–280.
- G-20 (2019). G20 Ministerial Statement on Trade and Digital Economy. 9 May.
- Gerdon, S. and Molinari, V. (2020). How Governments can use Public Procurement to Shape the Future of AI Regulation - and Boost Innovation and Growth. *World Economic Forum Blog*, 8th June.
- Geroski, P. (1990). Innovation, Technological Opportunity, and Market Structure. *Oxford Economic Papers*, (42):586–602.

- Gross, D. and Sampat, B. (2020). Inventing the Endless Frontier: The Effects of the World War II Research Effort on Post-War Innovation. *Harvard Business School Strategy Unit Working Paper No. 20-126*.
- Hagendorff (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines*, 30:99–120.
- Hayter, C. S., Link, A., and Scott, J. (2018). Public-Sector Entrepreneurship. *Oxford Review of Economic Policy*, 34(4):676–694.
- IEEE (2019). Ethically Aligned Design. 1st Edition at <https://ethicsinaction.ieee.org>.
- Korinek, A. (2019). Integrating Ethical Values and Economic Value to Steer Progress in Artificial Intelligence. *NBER Working Paper no. 26130, National Bureau of Economic Research*.
- Korinek, A. and Stiglitz, J. (2017). Artificial Intelligence and its Implications for Income Distribution and Unemployment. *NBER Working Paper no. 24174. National Bureau for Economic Research*.
- Leike, J., Martic, M., Krakovna, V., Ortega, P., Everitt, T., Lefrancq, A., Orseau, L., and Legg, S. (2017). AI Safety Gridworlds. *ArXiv:1711.09883 [Cs]*, November 27:<http://arxiv.org/abs/1711.09883>.
- Lenderink, B., Johannes, I., and Voordijk, H. (2019). Innovation and Public Procurement: From Fragmentation to Synthesis on Concepts, Rationales and Approaches. *Innovation: The European Journal of Social Science Research*, (<https://doi.org/10.1080/13511610.2019.1700101>).
- Lin, J. (2020). Industrial Policy Revisited. In *WTO, World Trade Report 2020: Government Policies to Promote Innovation in the Digital Age*. Geneva. World Trade Organization, page 26.
- Maas, M. (2018). Regulating for Normal AI Accidents: Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 223–228.
- Mazzucato, M. (2013). The Entrepreneurial State: Debunking the Public vs. Private Myth in Risk and Innovation. *Anthem Press: London, UK*.

- Mazzucato, M. (2018). Mission-Oriented Innovation Policies: Challenges and Opportunities. *Industrial and Corporate Change*, 27(5):803–815.
- Mazzucato, M. (2020). Mission-Oriented Innovation and Industrial Policies. In *WTO, World Trade Report 2020: Government Policies to Promote Innovation in the Digital Age*. Geneva. World Trade Organization, page 101.
- Miller, F. and Lehoux, P. (2020). The Innovation Impacts of Public Procurement Offices: The Case of Healthcare Procurement. *Research Policy*, 49:1–13.
- Moor, J. (2006). The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine*, 27(4):87–91.
- Naudé, W. (2021). Artificial Intelligence: Neither Utopian nor Apocalyptic Impacts Soon. *Economics of Innovation and New Technology*, 30(1):1–23.
- Naudé, W. and Dimitri, N. (2020). The Race for an Artificial General Intelligence: Implications for Public Policy. *AI & Society*, 35(2):367–379.
- Nguyen, A., Yosinski, J., and Clune, J. (2014). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *ArXiv*, <https://arxiv.org/abs/1412.1897>.
- Obwegeneser, N. and Müller, S. (2018). Innovation and Public Procurement: Terminology, Concepts, and Applications. *Technovation*, 74-75:1–17.
- OECD (2019). Recommendation of the Council on Artificial Intelligence. *OECD/LEGAL/0449, Adopted 22 May 2019. Paris: Organization for Economic Cooperation and Development*.
- Riedl, M. (2019). Human-Centered Artificial Intelligence and Machine Learning. *Human Behavior and Emerging Technologies*, 1(1):33–36.
- Russel, S., Dewey, D., and Tegmark, M. (2015). Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine, Association for the Advancement of Artificial Intelligence*, Winter:105–114.
- Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction*, 36(6):495–504.
- Simonite, T. (2020). The US Government Will Pay Doctors to Use These AI Algorithms. *Wired Magazine*, 11 October.

- Smith, M. L. and Neupane, S. (2018). Artificial Intelligence and Human Development: Towards a Research Agenda. *White Paper : International Development Research Centre*.
- Stanford University (2016). Artificial Intelligence and Life in 2030. *Stanford University at <https://ai100.stanford.edu>*, September.
- Stojčić, N., Srhoj, S., and Coad, A. (2020). Innovation Procurement as Capability-Building: Evaluating Innovation Policies in Eight Central and Eastern European Countries. *European Economic Review*, 121:1–18.
- Trajtenberg, M. (2018). AI as the Next GPT: A Political-Economy Perspective. *NBER Working Paper no. 24245. National Bureau for Economic Research*.
- Tzachor, A., Whittlestone, J., Sundaram, L., and ÓhÉigeartaigh, S. (2020). Artificial Intelligence in a Crisis Needs Ethics with Urgency. *Nature Machine Intelligence*, 2:365–366.
- Uyarra, E., Flanagan, K., Magro, E., and Zabala-Iturriagagoitia, J. (2017). Anchoring the Innovation Impacts of Public Procurement to Place: The Role of Conversations. *Environmental Planning*, C35:828–848.
- Van de Gevel, A. and Noussair, C. (2013). The Nexus between Artificial Intelligence and Economics. *Springer Briefs in Economics*.
- van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds and Machines*, September:<https://doi.org/10.1007/s11023-020-09537-4>.
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S., Tegmark, M., and Nerini, F. F. (2020). The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature Communications*, 11(233).
- von Hippel, E. (1976). The Dominant Role of Users in the Scientific Instrument Innovation Process. *Research Policy*, 5(3):212–239.
- WTO (2020). E-commerce, Trade and the COVID-19 Pandemic. *Information Note . Geneva: World Trade Organization*, 4 May.
- Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V., and Yang, Q. (2018). Building Ethics into Artificial Intelligence. *ArXiv*, 7 Dec.

## Data ethics requirements for RM6200 Artificial Intelligence Suppliers

It is important that Suppliers follow the data ethics framework, mitigate bias and ensure diversity in the team that developed/ is developing a solution; as well as transparency/ interpretability and explainability of the results, including audits.

The supplier will need to be open around how the AI service was built.

Below is a list of screening questions that could be added to an invitation to tender and asked of a Supplier where there is an ethical dimension to the tender.

### **Purpose:**

Describe the area of the problem space that is addressed by your solution

Is your approach built on an existing AI system (COTS) or will it be custom made?

Describe what algorithms or techniques you anticipate this service to implement?

Describe the approach to ensuring that use of AI/ML is both necessary and proportionate in the solution

Describe how you have ensured that the data used to power the AI solution is sufficient in quantity, accuracy and relevance to the data available, and what measures have been taken to mitigate bias in the model

Explain how end users have been considered throughout the design and implementation process

Explain how you will demonstrate accountability for the goals and outcomes of the solution

### **Consent and control:**

Provide evidence that you have adopted legally sound and ethical consent for processing and capturing the data throughout the full lifecycle of the solution

Describe the level of human decision-making at critical control points

### **Privacy and cybersecurity**

Describe your privacy and cybersecurity approach for the proposed solution, in particular how the data will be protected

Describe the potential threats and vulnerabilities to the system or solution from external or internal adversaries

Explain your test processes, including the specialist expertise used to assess the solution

Provide, if applicable, evidence of previous case studies of where the solution has been implemented and how the output has been interpreted, highlighting best practice

Describe how your organisation draws on specialist knowledge and expertise to develop and maintain the solution

### **Explainability**

Describe the capabilities in the business to ensure that the outputs of AI technology are explainable and that this explanation is widely available and understandable to a non-expert audience

Would you allow independent, third party audit(s) of the AI solution? If your answer is no, please explain

### **Ethical considerations relation to data limitations, fairness and bias**

What data limitations have you identified and what strategies will you implement to address these data limitations?

(Applicable where solution is COTS and/or government has shared data as part of the procurement process)

How will you ensure that the AI system fits the requirements of the Data Ethics Framework (or can you ensure that you meet the requirements of the Data Ethics Framework during the tool development?)

Describe the approach to eliminate (or minimise) bias, ethical issues, or other safety risks as a result of using the service

Describe the process for ensuring that the development team adopts an ethical mindset

Explain how the solution will be checked to detect bias and the remediation steps if it is introduced

What training data was used, which variables have contributed most to a result, and the types of audit and assurance the model went through, with respect to intrinsic attributes such as considerations of fairness and mitigation of bias. This should be included in documentation supplied by your provider.

### **Concept drift**

Explain how you will ensure the solution or service does not drift from its intended purpose or outcome

### **Interoperability**

Explain how your system or service conforms to specific international or local open interoperability standards or other relevant standards relating to cyber security, coding quality, safety or testing for example

### **Due diligence on existing algorithms or COTS solutions:**

Describe the architecture of the solution, including use of external COTS or open source components and the function they provide in the solution. This should consider the data used by each component of the solution and how the output of that component was validated

### **Documentation on toolkit and auditability**

Please provide information about:

1. the available toolkit, including the list of software tools the provider proposes to use
2. the origin and nature of any data the provider plans on bringing to the project
3. data used to train algorithms the provider will bring to the project
4. and algorithms used
5. documentation that provides information about the algorithms used (e.g. data used for training algorithms, whether the model is based on supervised, unsupervised, or reinforcement learning, and any limitations).

provide information on their model building methodology, including how they select variables, build samples (where applicable), and validate the model.

Documentation that provides information about the algorithms used (e.g. data used for training algorithms, whether the model is based on supervised, unsupervised, or reinforcement learning, and any limitations).

provide information on their model building methodology, including how they select variables, build samples (where applicable), and validate the model.

What training data was used, which variables have contributed most to a result, and the types of audit and assurance the model went through, with respect to intrinsic attributes such as considerations of fairness and mitigation of bias. This should be included in documentation supplied by your provider.

Enable end-to-end auditability by implementing process logs that gather the data across the modelling, training, testing, verifying, and implementation phases of the project lifecycle. Such a log should allow for the variable accessibility and presentation of information with different users in mind to achieve interpretable and justifiable AI.

### **Lifecycle management**

How do you envisage training and knowledge transfer of the AI system development and deployment to the public sector delivery team?

Explain how you will ensure usability for non-trained staff

Explain how the AI system will be maintained, how its accuracy and integrity will be sustained over time, and whether third party providers could be engaged for these activities

Monitoring and feedback loops: For AI-powered solutions in the public sector, implementation plans, sustainable and ongoing evaluation methods, and mechanisms to feed back into the data model are crucial to ensure ethical use.

Make clear in your invitation to tender that such considerations by the provider count and will be discussed during procurement

Testing: establish with the provider how the efficacy of the model will be monitored once deployed

### **Skills**

Can you demonstrate how you will assess the skills, qualifications and diversity of the team that will develop and deploy the AI system?



# U.S. DEPARTMENT OF HOMELAND SECURITY

## ARTIFICIAL INTELLIGENCE STRATEGY

December 3, 2020

*Vision: The Department of Homeland Security will enhance its capability to safeguard the American people, our homeland, and our values through the responsible integration of artificial intelligence (AI) into the Department's activities and by mitigating new risks posed by AI.*

## TABLE OF CONTENTS



<b>INTRODUCTION.....</b>	<b>1</b>
STRATEGIC VISION .....	2
DHS APPROACH.....	3
DEVELOPMENT AND IMPLEMENTATION.....	4
<b>GOAL ONE - ASSESS POTENTIAL IMPACTS OF AI.....</b>	<b>6</b>
OBJECTIVE 1.1: DEVELOP KNOWLEDGE OF TECHNICAL APPLICATIONS OF AI.....	6
OBJECTIVE 1.2: IDENTIFY OPPORTUNITIES FOR AI USE .....	6
OBJECTIVE 1.3: IDENTIFY CRITICAL IMPACT ON CRITICAL INFRASTRUCTURE .....	6
<b>GOAL TWO - INVEST IN AI CAPABILITIES .....</b>	<b>7</b>
OBJECTIVE 2.1: DEVELOP PHASED PLANS TO UPGRADE DEPARTMENT INFRASTRUCTURE .....	7
OBJECTIVE 2.2: DEVELOP BUDGET REQUIREMENTS FOR AI .....	7
OBJECTIVE 2.3: SURVEY EXISTING CAPABILITIES FOR SECURITY AND STORAGE .....	7
<b>GOAL THREE – MITIGATE AI RISKS .....</b>	<b>9</b>
OBJECTIVE 3.1: PRODUCE AND RELEASE PUBLIC AI DATA USE GUIDANCE .....	9
OBJECTIVE 3.2: DESIGN A PROGRAM TO CURATE NATIVE DATA SETS.....	9
OBJECTIVE 3.3: DOCUMENT SMART ENGINEERING AND OPERATING PRACTICES.....	10
OBJECTIVE 3.4: DESIGN A PROGRAM TO CURATE NATIVE DATA SETS.....	10
OBJECTIVE 3.5: FORMALIZE AI GOVERNANCE PROCESS.....	10
OBJECTIVE 3.6: TARGETED ENGAGEMENT AND EDUCATION.....	10
OBJECTIVE 3.7: COMMUNITY STANDARDS.....	11
OBJECTIVE 3.8: DEVELOPMENT OF INTERNATIONAL STANDARDS.....	11
OBJECTIVE 3.9: COUNTER ADVERSARIAL USE OF AI .....	11
<b>GOAL FOUR – WORKFORCE DEVELOPMENT .....</b>	<b>12</b>
OBJECTIVE 4.1: IDENTIFY CURRENT AI EXPERTISE AND GAPS.....	12
OBJECTIVE 4.2: IDENTIFY EXTERNAL AI TRAINING COURSES .....	12
OBJECTIVE 4.3: DEVELOP PUBLIC/PRIVATE SECTOR FELLOWSHIP PROGRAM .....	12
<b>GOAL FIVE – IMPROVE PUBLIC TRUST AND ENGAGEMENT.....</b>	<b>14</b>
OBJECTIVE 5.1: DEVELOP STRATEGIC COMMUNICATIONS PLAN.....	14
OBJECTIVE 5.2: ESTABLISH A FRAMEWORK FOR RELEASING AI INFORMATION.....	14
OBJECTIVE 5.3: COMMUNICATE IDENTIFIED AI-ENABLED THREATS.....	15
<b>CONCLUSION.....</b>	<b>16</b>



## INTRODUCTION

Artificial Intelligence (AI) refers to automated, machine-based technologies with at least some capacity for self-governance that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. The increased use of AI is inevitable as part of the ongoing global race to leverage new technologies for competitive advantages by nations and to increase economic prosperity by private sector entities. Use of AI can transform global economies, effect U.S. national security, and impact American citizens in their daily lives. The potential impact of AI also extends to critical infrastructure sectors like manufacturing, financial services, transportation, healthcare, energy, and food and agriculture.

AI presents opportunities for the Department of Homeland Security (DHS) to more effectively or efficiently accomplish our mission to secure the homeland. Yet with increased use of AI systems across the homeland security enterprise, comes increased risk. These risks include compromised or poorly designed AI systems, as well as adversarial use of AI technologies by unfriendly nations or criminals to increase their malicious capabilities. The potential impacts from AI on the security of the homeland and upon our Department's operational activities—both positive and negative—make it imperative for DHS to take a proactive role in the use of AI systems and to contribute to the national conversation on the secure use of this transformative technology. Therefore, DHS must act to ensure it is positioned to capitalize on the opportunities and benefits of AI, while constantly evaluating risks associated with the use of AI across the homeland security enterprise and the adversarial use of AI to cause us harm.

AI offers rich opportunities to improve the way we accomplish our mission across DHS Components. Efforts to secure the border, identify and interdict criminal actors, and secure cyberspace will be aided by use of AI systems. This strategy therefore seeks to prioritize the responsible use of AI by DHS while also mitigating AI-related risks to our homeland, citizens, and values.

While we work to reap the potential benefits of AI as a Department, we must also ensure that our use of AI comports with best practices and promotes trust and confidence of the public and of our domestic and international partners. DHS will be guided by the principles set forth in Executive Order 13690, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government* (December 3, 2020). Trust in the Department's expertise to identify and mitigate security risks and in its responsible use of its own AI systems is at the core of DHS's future success and leadership in AI. Furthermore, public input, especially in those instances where AI uses sensitive personal information, will improve the Department's accountability and increase the trust and confidence of the American people.

However, we must also be aware of other risks associated with the use of AI by partners and stakeholders across the homeland security enterprise, and the risks associated with the malicious use of AI to threaten the homeland. Potential adversarial use of AI will continue to evolve at pace with the development of the technology. Adversaries can increasingly use AI-enabled systems to exploit or overcome security measures currently in place at our physical borders including at ports-of-entry, in cyberspace, in election systems, and beyond. DHS will work to

make our nation more secure and resilient against the malicious use of AI and other emerging technologies by other nations and by criminals.

DHS will take a proactive role in the ongoing national conversation on AI by issuing this strategy and through the development of a subsequent implementation plan. The strategy sets out five goals to govern the Department's approach to successfully integrating AI into our mission in a responsible and trustworthy manner and successfully mitigating risks associated with AI across the homeland security enterprise.

### **DHS AI Goals**

- 1. Assess Potential Impact of AI on the Homeland Security Enterprise*
- 2. Invest in DHS AI Capabilities*
- 3. Mitigate AI Risks to the Department and to the Homeland*
- 4. Develop a DHS AI Workforce*
- 5. Improve Public Trust and Engagement*

*Figure 1. DHS AI Goals*



These concrete and achievable goals for DHS are consistent with the Department's strategic vision which is set out below and is grounded in a risk-informed model that fosters accountability, transparency, trust, and security.

### **Strategic Vision**

Our strategic vision is that DHS will become a global leader in policy development, governance, and the use of AI systems as we lead national efforts to mitigate against institutional and enterprise risks posed by AI. DHS Components will take actions to identify opportunities for purpose-driven and tailored AI solutions for mission enhancement and continuously identify risks when considering the use case for a new AI system or when evaluating an existing system. DHS will simultaneously engage with external partners, including the public, to effectively communicate its AI policies and to grow and maintain an expert workforce to lead the responsible and accountable use of AI throughout the system's life cycle.

In furtherance of its mission, DHS will strive for the highest standards of transparency, accountability, and public outreach. Public trust in the Department's expertise to operate its own AI systems and to identify and mitigate security risks is at the core of DHS's future success and

leadership in AI. Innovation and growth in AI should therefore be prioritized, in all instances, in accordance with fundamental rules and best practices to address the creation, acquisition, privacy, integrity, security, quality, and use of AI systems and any associated data sources.

We will remain always vigilant to emerging AI risks in carrying out the vital missions of protecting the American people as the Department becomes a leading practitioner of AI. DHS will develop comprehensive risk assessments and mitigation measures to ensure the continued success of the homeland security mission and protection against adversarial use of AI. DHS will leverage existing and new partnerships across the AI and homeland security communities to inform and manage AI risks to the homeland.

## DHS Approach

DHS has a responsibility to the American people to innovate in support of its mission and to do so responsibly and deliberately. AI has already emerged as a technology and is increasingly ubiquitous in applications used every day across private sector, academia, and the U.S. Government. As is the case across the Federal Government, DHS is currently deploying and operating various AI systems. AI brings tremendous potential to improve processes that yield increased efficiency and effectiveness across the public and private sectors alike. As with any technology, however, opportunities for innovation come with risks.

For DHS to accomplish our mission to safeguard the American people, our homeland, and our values, we must use AI consistent with law and respectful of those values we seek to safeguard. We must ensure, for example, that DHS Components have measures in place to increase transparency, accountability, and to regularly monitor AI systems for potential bias and error.

The integration of AI into the current technology landscape combined with the rapid pace of development led the White House to develop EO 13960. That executive order sets forth the principles that will guide the use of AI across the Federal Government and by the Department. The principles were designed to provide a framework for DHS and other federal agencies to consider when weighing the adoption of AI solutions. This DHS AI Strategy is guided by and builds upon the principles in the EO. The DHS AI Strategy seeks to position the Department to take advantage of creative and innovative AI solutions consistent with the principles in EO 13960.

While the principles of EO 13960 will serve as our foundation, this strategy sets forth broader strategic goals that will prepare the Department to be a responsible and trustworthy user of AI and to secure the homeland against risks presented by AI technology and by adversarial use of AI. AI has shown great promise at DHS to enhance a variety of missions such as cybersecurity protections, law enforcement investigations, and a range of other operational efficiencies. Currently, however, these DHS AI systems operate in the absence of a unified, enterprise-level strategic approach to AI usage and investment.

The first goal of this strategy calls for the Department to take a comprehensive, wholistic, and strategic look at AI in terms of the potential impacts—positive and negative—that it can have across our mission set. Such an assessment, leveraging expertise inside and outside the

Department, will better position DHS to effectively make informed decisions about its AI-related efforts. Goal 2 then focuses on the need for strategic investment by DHS in AI, including an assessment of current capabilities, a Department-wide effort to invest in core infrastructure needed to run advanced AI algorithms securely and efficiently, consideration of realistic research and development opportunities, and associated budget priorities. The objectives under Goal 2 will drive a concentrated effort to strategically position DHS to take advantage of the opportunities present by AI.

As previously noted, AI also presents risks and challenges for DHS in addition to significant opportunities. As recognized in EO 13960, questions around the governance of AI systems remains a persistent problem across academia, legal circles, and the AI user community. Goal 3 identifies a series of objectives to ensure that DHS implements the principles set out in EO 13960 to ensure that DHS AI systems are accurate, safe, understandable, and regularly monitored as part of a consistent approach to effective governance rather than through ad hoc DHS Component practices. Goal 3 also sets out strategic objectives to address other risks posed by reliance on AI systems across the homeland security enterprise and potential adversarial use of AI to target the homeland. Key objectives include engaging and educating key stakeholders about AI risk, supporting the development of community and international AI standards, and mitigating identified adversarial use AI risks to the homeland.

Of course, a foundation for success in meeting these first three strategic goals will be the development of a professional AI workforce supporting the homeland security mission. Goal 4 therefore focuses on the need for DHS to improve its ability to attract and retain AI professionals across disciplines as it works to improve its science, technology, engineering, and mathematics (STEM) workforce. To keep pace with the development and application of new AI systems, the Department must develop and retain an expert AI workforce within our Science & Technology Directorate (S&T) and across other DHS Components. Reflective of a national need for AI professionals in government, private industry, and academia, DHS must partner with the academic and private sectors to become a destination of choice for highly educated and highly talented STEM experts. Similarly, DHS success must be founded on engagements that will build public trust in AI generally and in DHS's use of AI particularly. Goal 5 focuses therefore sets out objectives for DHS efforts to engage and increase such trust.



DHS faces both opportunities and challenges to achieve our vision to become a leader in the trustworthy use of AI. We recognize the need to prioritize and inspire creative and innovative thinking that leads to the rapid and safe adoption of new technologies. We also recognize the need to identify and mitigate possible risks from AI by investing in our capabilities to meet those challenges. Our efforts, guided by the principles in EO 13960, will contribute to an improvement of the public's trust in the Department's ability to use AI effectively and responsibly to safeguard our homeland, our citizens, and our values.

## Development and Implementation

The DHS Office of Strategy, Policy, and Plans (PLCY) led the development of this strategy in collaboration with all DHS Components. DHS will also issue a corresponding implementation plan to outline DHS Component roles, responsibilities, programs, and timelines for accomplishing these goals and objectives.

This strategy and the implementation plan will be used to harmonize and prioritize DHS AI planning, programming, budget, training, and execution activities. In addition, the Joint Requirements Council will utilize the strategy and implementation plan to support the review of capability gap analyses and requirements generated by relevant Components. PLCY will annually assess implementation of this strategy and provide a report to the Secretary. The report will include areas of success, opportunities for improvement, constraints impeding progress, and suggested adjustments to the strategy. DHS will review and assess the need to update this strategy in 2024, and periodically thereafter.



## GOAL ONE – ASSESS POTENTIAL IMPACT OF AI ON THE HOMELAND SECURITY ENTERPRISE

The rapid development of AI and its broad applicability present opportunities for DHS to improve our mission execution but also new risks to our mission execution and risks to the homeland from increased adversary capabilities. DHS will begin by improving its strategic cooperation with interagency partners, foreign partners, academia, and the private sector to continually assess AI's impact to the DHS mission. Improved strategic cooperation will effectively position the Department to make critical and informed assessments of the impact of AI technology across its missions, both positive and negative. Such cooperation will provide the Department with broad perspectives on the most up-to-date information on the research, development, and potential applications of AI.

AI-related opportunities and risks will evolve at the pace of technological development and the Department must be positioned to act quickly to leverage new technologies. Concurrently, DHS must also counter threats such as AI-enabled computer network intrusions, threats to critical infrastructure, deep-fakes, big data processing, and misinformation campaigns. Accomplishment of this goal will create a robust and diverse understanding of opportunities and risks to the Department's mission while advancing the principles of EO 13960 domestically and internationally.

### *Objective 1.1: Develop Knowledge of Technical Applications of AI*

DHS will partner with private sector entities, international partners, and academic institutions to survey and assess current research and publications related to AI technological developments with the goal of assessing AI impacts specific to the homeland security mission. This objective will position the Department to identify possible AI-associated opportunities and risks, and to develop mitigation measures to counter the risks.

### *Objective 1.2: Identify Opportunities for AI Use*

DHS will identify legacy systems, processes, and mission areas to which the addition of a purpose-driven and tailored AI solution would result in increased efficiencies, optimal use of resources, and general mission enhancement across the Department. Integration of AI will be guided by the principles in EO 13960 and in accordance with Objective 3.1 below.

### *Objective 1.3: Identify Critical Applications and Impact of AI on U.S. Critical Infrastructure*

DHS will leverage its existing authorities and relationships and examine the need for new relationships to engage the critical infrastructure community and academia to inform and study the current and future beneficial effects and risks of AI on U.S. critical infrastructure systems.



## GOAL TWO – INVEST IN DHS AI CAPABILITIES

DHS must take steps to ensure that we are able to take advantage of new AI systems and the opportunities they present. AI algorithms rely on advanced computational capabilities, secure data storage, and the infrastructure to move large amounts of data at high speeds. DHS will therefore survey existing capabilities, identify gaps, and invest in infrastructure and supporting technologies to support the large computational and data storage demands of future DHS AI systems.

Infrastructure investments shall be structured in a manner that is consistent with law and policy, particularly when such infrastructure uses, leverages, or maintains personally identifiable information (PII). Specifically, the Department will focus on high performance computing, secure cloud computing, special purpose processors, Graphics Processing Unit (GPU) capabilities, fast data connections, large data storage and computation capabilities, and improved data connections. This investment will serve as the foundation upon which future AI capabilities can be built and will give the Secretary the option to implement an AI solution to execute DHS missions.

DHS will also develop research and development priorities to determine whether and how to contribute limited departmental R&D funding to support AI systems most relevant to the homeland security enterprise. To support these objectives, DHS must relook at current and future budgets in light of the impact AI will have on DHS operations and activities. Accomplishment of this goal in the spirit of this strategy will position the Department as a Federal Government leader in AI-ready infrastructure and capabilities and contribute to public trust in the responsible use of AI consistent with Goal 5.

### *Objective 2.1: Survey Existing Computational and Data Storage Capabilities for Security and Storage Capacity*

DHS will assemble a cadre of internal and external computational capacity, data storage, and security experts across DHS Components to survey the current state of the Department's AI-ready infrastructure to make recommendations on how it can be improved. DHS Components will then consider the recommendations for their infrastructure investment consistent with Goal Three.

### *Objective 2.2: Develop Phased Plans to Upgrade Department Infrastructure*

Leveraging insights learned through Goal One and its Objectives, DHS will produce a phased plan to upgrade the Department's computational capability, data storage, and associated infrastructure in accordance with the identified risks and needs aligned with related efforts to deploy secure IT systems. The Department will prioritize areas of greatest need from both a technical and risk mitigation perspective.

*Objective 2.3: Evaluate and Invest in AI Research and Development*

The Department will consistently evaluate AI research and development and invest in technologies with the potential to enhance the homeland security mission.

*Objective 2.4: Develop the Department's Budget Requirements for AI*

Multiple think-tank, university, and government studies call for a substantial increase to the U.S. Government's overall spending on government AI systems and the funding of research and development to keep the United States in a leadership position in its development. DHS Components will evaluate their current and projected AI needs and produce a projected budget requirement for necessary infrastructure upgrades. Associated resource increases must include both systems and the teams needed to maximize value and provide oversight. Proposed infrastructure upgrades will also improve other mission functions bringing a broad benefit to the Department's capabilities.



## GOAL THREE – MITIGATE AI RISKS TO THE DEPARTMENT AND TO THE HOMELAND

AI will pose evolving risks to DHS activities as we start to incorporate AI systems into our operations. More broadly, AI increased use of AI will also introduce new and evolving risk to the homeland security enterprise as new technologies and applications are relied upon by our partners and stakeholders. This includes risks from adversarial efforts to compromise AI systems or to use AI technology to target the homeland.

Mindful of these risks DHS will produce a comprehensive risk outlook and develop mitigation measures to ensure the continued success of the homeland security mission. DHS will cultivate a robust set of partnerships across the AI and homeland security communities to inform and manage risks related to AI applications.

DHS will also develop the policies, practices, and expertise necessary to serve as a model for AI deployment to organizations that function in complex operating environments. Furthermore, DHS will ensure transparency to the extent practicable around AI determinations that impact individuals or entities so that the American public can understand decision making influenced by AI systems and inform regulatory and policy outcomes.

### *Objective 3.1: Develop a Process for Continual Evaluation of AI Risks*

The pace of research and development of AI technology and applications, as well as the rapid increase in computing power using classical and quantum architectures, demands a constant and continual evaluation of AI risks. Components operating AI systems to support their missions must also continually validate the performance of their system to monitor for and take action to mitigate risks posed by bias or other unintended outcomes. DHS will develop a process by leveraging expertise from inside the Department, private sector, and academia partners to produce a report for the Secretary outlining the current state and projected future risks of AI to the homeland.

### *Objective 3.2: Produce and Release Public AI Data Use Guidance*

The appropriate use of data for the training or operation of an AI system is central to the responsible integration of AI technology into the successful execution of our mission and mitigating certain risks associated with DHS reliance on AI systems. DHS will prioritize the production of specific guidance on the use of data by DHS Components for AI purposes consistent with applicable legal requirements, including requirements related to data use and privacy to avoid issues of efficiency, trustworthiness, and biased system outcomes.

This guidance will, to the extent practicable, be released publicly and will clearly state the circumstances under which different types of data will be used and what measures DHS Components will take to protect privacy and ensure the responsible and trustworthy use of AI by DHS. DHS guidance and strategic vision can serve as a model across the homeland security

enterprise for entities using AI systems including our state, local, tribal, and territorial (SLTT) partners, critical infrastructure, and law enforcement consistent with protection of privacy, individual rights, economic development, and national security.

*Objective 3.3: Design and Implement a Program to Curate Native Datasets for Optimum AI Use*

Human-curated training data for AI systems is critical to ensuring trust in the output and operation of AI systems used by DHS. DHS will ensure that the algorithmic training matches the intended outcome for the system. DHS Components holding data that is currently or may be used in the operation or training of an AI system will design and implement a program to curate their datasets to optimize the use of AI in such a manner that mitigates the potential for biases, which could negatively impact the integrity or reliability of the AI system or degrade public trust, in accordance with the guidance developed by DHS as part of Objective 2.1.

*Objective 3.4: Document Smart Engineering and Operating Practices*

DHS Components shall include in any proposal for the use of an AI system a mechanism by which the data used to train the system is, to the extent practicable, accessible by managers and operators of the system. The accessibility shall also extend to affected populations outside of certain military, foreign intelligence collection, or law enforcement activities. By doing so, DHS Components will address potential issues with system explainability and operational integrity.

Documentation of engineering and operating practices by DHS Components makes it more feasible to provide assurances of both ethical alignment and operational integrity of AI systems. Components shall also use privacy engineering concepts to the greatest possible extent and document the consideration of these concepts.

The Department will not obscure its use of AI to support operations but will be transparent to the extent practicable, to build public trust and to encourage public and private development of AI systems that are accountable to users, the public, oversight bodies and the legal system. In all cases in which AI is used, operational assessments of potential risk and harm, the magnitude of those risks and harms, the technical state of the art, and the potential benefits of the AI system must be substantiated to facilitate both explainability and transparency.

*Objective 3.5: Formalize AI Governance Processes at DHS*

DHS will establish a DHS-enterprise wide AI Coordination and Advisory Council (Council) composed of internal subject matter experts to monitor and support the adoption of AI technology by DHS Components. The Council will leverage external and internal experts to assist DHS Components in AI adoption; to study and share best practices from other agencies and the private sector; and to coordinate with internal DHS governance bodies, as appropriate to allow Components to develop unique AI requirements that align with their respective missions. This Council will also consider legal, compliance, classification, civil rights and civil liberties, and privacy implications and responsibilities related to DHS AI projects, methods and capabilities.

*Objective 3.6: Targeted Engagement and Education to Homeland Security Enterprise Partners*

DHS will prioritize the sharing of risk information associated with the use of AI with SLTT partners, critical infrastructure operators, and other partners in the homeland security enterprise. The Department will also educate these partners on good AI use practices and risk mitigation techniques to ensure the security and responsible use of AI across the enterprise.

*Objective 3.7: Community Standards*

DHS will support interagency and infrastructure community efforts to institute standards governing responsible use of AI technologies. Through development and communication of shared best practices and standards, infrastructure usage of AI can be more securely deployed.

*Objective 3.8: Development of International AI Standards*

DHS will seek consistent and ongoing international cooperation from a broad range of partners to focus on the development of mutually agreed upon principles on responsible stewardship of trustworthy AI; multi-stakeholder, consensus-driven global technical standards for interoperability; internationally comparable metrics to measure AI research, development, and deployment.

The increased perspective on AI challenges presented in other countries will allow DHS to anticipate potential risks and respond in an effective manner. Increased cooperation on AI with international partners will also allow the opportunity for the Department to advance the principles of this strategy with like-minded nations.

*Objective 3.9: Counter Adversarial Use of AI Against the Department and the Homeland*

DHS will take a strategic approach to mitigate and counter efforts by our adversaries to leverage AI technology against the Department and the homeland. DHS will maintain a broad and evolving picture of potential AI risks and AI capabilities of our adversaries to inform the Secretary of potential risks and take effective actions to counter malicious activity.



## GOAL FOUR – DEVELOP DHS AI WORKFORCE

Critical to any DHS implementation and maintenance of an AI system is a trained workforce, including AI policy and privacy professionals, that can manage augmentation of current systems with AI functionality, design and implement new AI augmented systems and understand how advances in the field can benefit the Department. DHS Components will prioritize the hiring and development of AI professionals to manage and maintain systems in a manner that preserves public trust in the stewardship of AI systems and associated data. The Department will also engage in outreach to institutions training AI professionals to offer specific internships and fellowships for AI students and professionals. Accomplishment of this goal in the spirit of this strategy will grow and mature an expert cadre of uniquely DHS AI professionals supporting the Department’s missions.

### *Objective 4.1: Identify Current AI Expertise and Gaps Across DHS Enterprise*

DHS will evaluate its current AI use footprint and compare its in-house AI experience with the expertise requirements to responsibly manage and operate its systems. DHS will further conduct a survey to identify current employees with AI expertise. These employees will be made available as an enterprise resource on an ad hoc basis for DHS Components lacking sufficient AI experience to evaluate the need for a new AI solution or to evaluate an existing AI system for its overall technical health and for compliance with this strategy, under the oversight of the AI Coordinating and Advisory Council. The gap that exists between the native DHS AI experience and the proliferation of AI use in the Department will inform actions by DHS to recruit and retain AI expertise.

### *Objective 4.2: Identify External AI Training Courses and Make Available to Workforce*

Components will identify external AI training courses consistent with their AI needs and make them available to the workforce. Components should make training funds available to cover the cost of appropriate AI courses. This training program will encourage the growth of an expert AI cadre native to DHS. Along with increased operational effectiveness, AI carries enterprise-level fiscal, legal, security, operational and political risk. Training will, therefore, also be made available to business users, oversight and advisory office staff, managers and leadership who support and have oversight on AI projects. Formal and informal internal training will also be made available, as appropriate.

*Objective 4.3: Partner with Academic and Private Sector to Develop a Public/Private Sector Fellowship Program*

Consistent with Goal 1, DHS will continue to develop partnerships with academic and private-sector research and technology entities and engage in an exchange program sending DHS employees on rotation to private firms and academic institutions and accepting private and academic researchers and technical experts on rotation to DHS. Fellows will be a key part of future DHS AI workforce and be required to report to their Component and the broader Department on the state of AI research and development and new potential applications impacting Departmental equities. This effort will keep DHS at the leading edge of AI development.



## GOAL FIVE – IMPROVE PUBLIC TRUST AND ENGAGEMENT

The trust of the American people is vital to the success of the implementation and responsible use of AI by the Department. AI technology is not widely understood and, as such, carries a negative connotation with many non-experts. Further, the public must trust that the Department is evaluating AI-enabled threat vectors to the homeland using well-informed experts. Public trust will strengthen the Department’s use of AI by ensuring the support of the citizens it intends to protect and guard against reputational impacts to the Department.

DHS will facilitate this trust by engaging strategically with the American public regarding AI. DHS will also ensure that DHS Components fully consider privacy and civil liberty implications when leveraging personal information; engage in a public comment period as appropriate; and consult with the AI Advisory and Coordination Council to ensure Departmental equities are addressed and to leverage expertise from other parts of DHS to support the project. DHS will also seek to engage the public on AI risks to the maximum extent practicable. Accomplishment of this goal will result in an educated and supportive American public who supports the Department’s transparent and accountable use of AI in accordance with this strategy and applicable Executive Branch guidance.

### *Objective 5.1: Develop strategic communications plan to support Communication to the Public on AI*

DHS will produce guidance for DHS-wide communication with the public on AI. This communication will apply Department-wide with exceptions for certain military, law enforcement, and intelligence activities and be oriented toward increasing public awareness of AI systems in general and communicating the Department’s position on responsible and trustworthy use of human-involved AI with examples of how that position is being carried out.

DHS will also ensure a public communications plan will be developed and implemented across the Department. The communication plan will make clear the importance of AI for protecting the homeland and will also serve to deter adversaries. This plan shall effectively and clearly identify: 1) the intended use, 2) data elements needed for algorithmic training and function, 3) parameters and levels of human oversight and decision making, 4) transparency regarding the collection, use, dissemination, disclosure, and protection of information, 5) benchmarking and subsequent auditing of the system performance, 6) accuracy of results and means of providing individuals redress for inaccuracies, improper use, or disclosure, and 7) decision making parameters performed by humans and by an algorithm.

### *Objective 5.2: Establish a Framework for Releasing AI System Information for Public Comment*

Future AI systems implemented by DHS will require a public release of system information with appropriate exceptions for certain sensitive military and intelligence systems, and some exceptions for law enforcement activities. DHS will produce a framework for releasing AI system information and a process for public comment.

*Objective 5.3: Communicate Identified AI-Related Risks*

In support of establishing public trust in AI, the Department will seek to communicate the identification of AI-related risks when practical and considering intelligence collection, law enforcement, and military equities.



## CONCLUSION

In an age of the reemergence of great power competition, the United States faces threats to its homeland, its citizens, and its values from more vectors than ever before. This fact is compounded by the rapid development of a new technological frontier in AI. AI provides numerous challenges for DHS including its responsible use of AI consistent with American values and securing the homeland in light of new and evolving risks. It is imperative that the Department not only to prevail over these challenges, but also lead in investment and integration of AI technology into accomplishing our mission.

The Department will evolve to ensure it is postured for continued mission success as AI opportunities and risks evolve with the further development of the technology. The Department will act in a way to understand the risks AI poses to its mission and embrace innovation to counter those and other new and emerging risks. The goals and objectives presented in this strategy will therefore position the Department to confront these challenges and to become a leader in the effective adoption, use, and governance of AI and the mitigation of AI-related risks.

The security of our homeland depends on the Department's response to this national imperative. With this strategy, the Department can move forward with concrete and tangible goals to answer the challenges and create a safer homeland while preserving American values.



# S&T ARTIFICIAL INTELLIGENCE & MACHINE LEARNING STRATEGIC PLAN



**Homeland  
Security**

Science and Technology

**AUGUST 2021**



# LETTER FROM SCIENCE & TECHNOLOGY LEADERSHIP

Science and technology innovations help our nation answer the threats of tomorrow and the needs of today. Leading research and development for the homeland security mission, the U.S. Department of Homeland Security (DHS) Science and Technology Directorate (S&T) has the power to bring the right partners together to solve mission-critical problems for the Department's frontline employees as they protect and secure our nation. By leveraging current and emerging scientific advancements, S&T supports immediate DHS Component operational gaps while preparing the Department to address future threats.

Artificial Intelligence is a revolutionary capability that presents substantial opportunities for DHS to more efficiently and effectively accomplish our mission to secure the homeland. Effective implementation will also require significant expertise, coordinated research and development, and targeted investments.

I am proud to introduce the S&T Artificial Intelligence/Machine Learning Strategic Plan, which lays out an actionable path for S&T to advise and assist the Department in harnessing the opportunities of Artificial Intelligence and Machine Learning (AI/ML). Through this strategy, S&T will build and apply expertise to help the Department fulfil the game-changing promise of AI/ML technologies while mitigating the inherent risks.

As we grow S&T's proficiencies in artificial intelligence and machine learning, we ensure S&T continues supporting the Department's vision to enhance its capability to safeguard the American people, our homeland, and our values through the responsible integration of artificial intelligence into the Department's activities.

Sincerely,



Kathryn Coulter Mitchell

Senior Official Performing the Duties of Under Secretary for Science and Technology



# TABLE OF CONTENTS

<b>I. EXECUTIVE SUMMARY.....</b>	<b>1</b>
<b>II. PURPOSE.....</b>	<b>2</b>
<b>III. INTRODUCTION .....</b>	<b>2</b>
A. AI/ML IN DHS MISSION CONTEXT .....	2
B. AI/ML IN S&T MISSION CONTEXT.....	2
C. S&T AI/ML VISION .....	3
D. DEFINITIONS OF AI/ML .....	3
E. STRATEGY DEVELOPMENT PROCESS .....	3
<b>IV. VALUES AND PRINCIPLES .....</b>	<b>4</b>
<b>V. GOALS.....</b>	<b>4</b>
GOAL 1: DRIVE NEXT-GENERATION AI/ML TECHNOLOGIES FOR CROSS-CUTTING HOMELAND SECURITY CAPABILITIES .....	6
GOAL 2: FACILITATE USE OF PROVEN AI/ML CAPABILITIES IN HOMELAND SECURITY MISSIONS .....	10
GOAL 3: BUILD AN INTERDISCIPLINARY AI/ML-TRAINED WORKFORCE .....	13
<b>VI. CONCLUSION .....</b>	<b>15</b>
<b>VII. APPENDICES .....</b>	<b>16</b>
A. ACRONYMS .....	16
B. REFERENCES.....	16
C. ENDNOTES .....	18



# I. EXECUTIVE SUMMARY

The U.S. Department of Homeland Security (DHS) Science and Technology Directorate (S&T) presents goals that will enable S&T to conduct Artificial Intelligence and Machine Learning (AI/ML) research, development, test, and evaluation activities to support DHS mission needs, and to advise stakeholders on developments in AI/ML and the associated opportunities and risks.

The S&T AI/ML Strategic Plan defines S&T's approach to effectively address the opportunities and challenges that AI/ML poses to the Department, the broader Homeland Security Enterprise, and the missions they serve. The S&T AI/ML Strategic Plan presents three goals:

## ***GOAL 1: Drive Next-Generation AI/ML Technologies for Cross-Cutting Homeland Security Capabilities***

S&T will make strategic investments in AI/ML research and development activities that meet critical DHS needs. S&T has identified three R&D objectives: Advance Trustworthy AI, Advance Human Machine Teaming, and Leverage AI/ML for Secure Cyberinfrastructure. Advancing Trustworthy AI is an interdisciplinary effort to research and provide actionable solutions for issues such as explainable AI, privacy protection, countering bias, and countering adversarial machine learning. S&T will research Human Machine Teaming, optimizing human and machine interactions while limiting their weaknesses. In the area of Secure Cyberinfrastructure, S&T will research capabilities that allow data sharing and processing across systems, effective management of AI/ML models, and AI/ML capabilities that enable threat detection and response.

## ***GOAL 2: Facilitate Use of Proven AI/ML Capabilities in Homeland Security Missions***

S&T will identify technically mature capabilities and match them to mission needs to facilitate understanding and adoption of existing AI/ML solutions by DHS Components and stakeholders. S&T will also advance capabilities that can be used by non specialists to curate and process large datasets, while advising the Department on the technical and policy infrastructure needed for AI/ML.

## ***GOAL 3: Build an Interdisciplinary AI/ML-Trained Workforce***

S&T will recruit experts and train current personnel to improve AI/ML competence across the S&T workforce in order to more effectively achieve S&T missions. Additionally, S&T will provide expert advice and recommendations for training opportunities to the broader DHS and Homeland Security Enterprise (HSE)<sup>i</sup> communities.

S&T's approach to AI/ML is informed by national guidance and the [DHS Artificial Intelligence Strategy](#). S&T leadership is committed to ensuring that AI/ML research, development, test, evaluation, and departmental applications comply with statutory and other legal requirements, and sustain privacy protections and civil rights and civil liberties for individuals. A subsequent S&T AI/ML Implementation Plan will detail how the S&T AI/ML Strategic Plan will be operationalized.



## II. PURPOSE

The S&T Artificial Intelligence and Machine Learning Strategic Plan establishes the S&T AI/ML vision, mission, goals, and objectives. The Strategic Plan identifies the focus areas for AI/ML that S&T will address to carry out its missions as the research and development arm and science and technology advisor to DHS Components, DHS Headquarters, and the Homeland Security Enterprise.

## III. INTRODUCTION

### A. AI/ML in DHS Mission Context

**DHS missions** include, for example, managing cyber and physical risks to critical infrastructure, securing American borders while facilitating lawful trade and travel, preventing and investigating criminal activity, and responding to natural and human-made disasters. AI/ML presents opportunities to more efficiently and effectively accomplish the many and varied missions of the DHS Components and the broader HSE. Examples of possible uses for AI/ML in these missions include processing large quantities of sensor data at border crossings, scanning cyber activity for anomalies, and modeling the effects of a natural disaster on critical infrastructure. AI/ML also introduces new risks, which DHS must be prepared to identify, assess, and mitigate. DHS must develop and field this new technology in a way that protects privacy, civil rights and civil liberties, and protects against bias, both to ensure effectiveness and to maintain public trust. DHS must also be prepared to respond effectively to issues when they occur.

### B. AI/ML in S&T Mission Context

S&T's mission is to safeguard the nation by answering the threats of tomorrow and the needs of today through science, technology, and innovation. Created by the Homeland Security Act of 2002, S&T conducts basic and applied research, development, demonstration, testing, and evaluation activities relevant to DHS and the HSE. The S&T AI/ML Strategic Plan meshes with and is supported by the goals articulated in the [\*\*2021 S&T Strategic Plan\*\*](#). As the research and development arm of DHS and trusted science and technology advisor to DHS Components and HSE stakeholders, S&T will conduct research to understand opportunities and risks associated with rapidly changing AI/ML technologies and impacts to DHS missions.

S&T will enable DHS and the broader HSE to effectively use AI/ML to carry out their missions of protecting the American people and the homeland, while operating with ethical standards and in accordance with the values of the American people. S&T leadership is committed to ensuring that AI/ML research, development, test, evaluation, and departmental applications comply with statutory and other legal requirements, sustain privacy protections, and maintain civil rights and civil liberties for individuals.



## C. S&T AI/ML Vision

S&T is the Department's trusted advisor for AI/ML, providing expert, independent, and objective technical guidance for research, acquisition, and implementation of AI/ML capabilities for critical homeland security missions in partnership with federal, industry, academia, and international partners. S&T anticipates how the increasing ubiquity of AI/ML in society, including the use of AI/ML by adversaries, may impact DHS and the HSE. Additionally, S&T informs and fosters DHS Components' assessment and potential acquisition of AI/ML capabilities. Finally, S&T possesses organizational capacity to understand the opportunities and risks in the rapidly changing field of AI/ML, in order to advance DHS missions.

## D. Definitions of AI/ML

### *Artificial Intelligence*

"Artificial Intelligence (AI) refers to automated, machine-based technologies with at least some capacity for self-governance that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments" (DHS Artificial Intelligence Strategy, December 3, 2020).

### *Machine Learning*

Machine Learning (ML) is a subset of AI. ML systems receive inputs in the form of training data, and then generate rules that produce outputs. In other words, ML systems "learn" from examples, provided in the form of training data, rather than receiving explicit programming from humans. In recent years, increasing availability of very large datasets, developments in available computing power, and other technical advances have made ML useful and promising for a variety of applications.

As AI/ML technologies advance, DHS must continually assess the opportunities and risks associated with their uses in order to ensure that DHS can effectively leverage emerging technologies to achieve its missions, while anticipating new dependencies associated with algorithmic decision-making. DHS will ensure that the use of AI/ML meets ethical standards and promotes and protects U.S. interests. S&T will advise DHS Components and partners on state-of-the-practice to effectively prevent, deter, or counter threats.

## E. Strategy Development Process

The S&T AI/ML Strategy Working Group, led by the Technology Centers Division and including a matrixed membership from across S&T, convened in Summer 2020 to identify and scope core areas of research and development activity that will serve to inform, educate, and improve S&T's program areas as well as the Department's activities impacted by this disruptive technology. Its assessments were informed by national guidance, DHS policy, and DHS Component outreach to coordinate and focus current and future activities in AI/ML within S&T.



The goals and objectives below were informed by a series of workshops held in October and November 2020 with operators, program managers, technologists, senior leaders, and decision-makers from across DHS. S&T's goals align with those outlined in the DHS AI Strategy and specify S&T's research and development priorities in AI/ML.

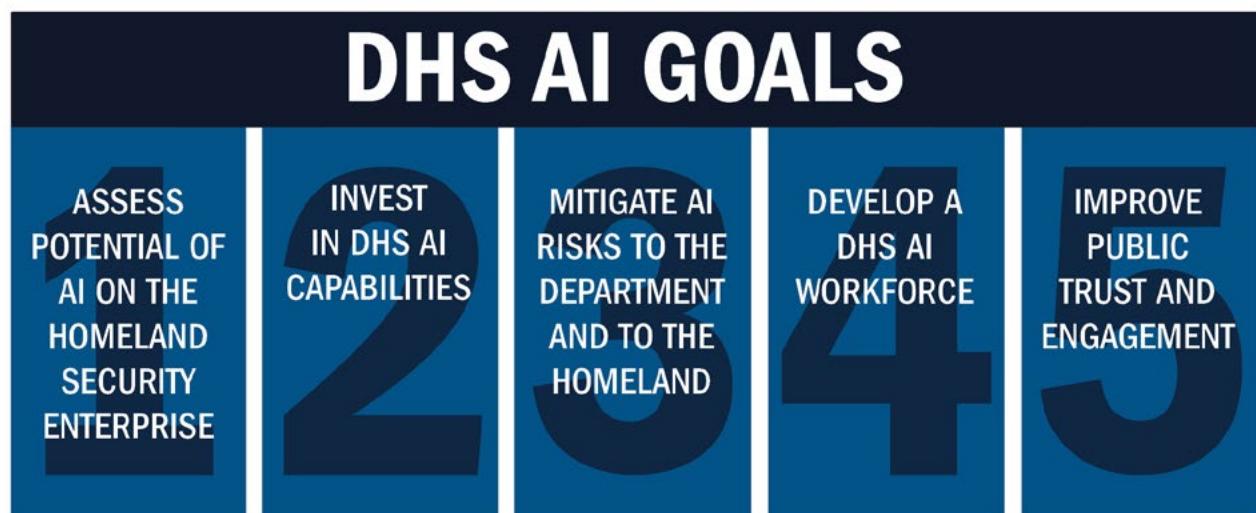
## IV. VALUES AND PRINCIPLES

The S&T AI/ML Strategic Plan aligns with the department-wide [DHS Artificial Intelligence Strategy](#) (December 3, 2020) and overarching [DHS Guiding Principles](#) by specifying how S&T will support and address the research and development challenges and opportunities that emerging AI/ML technologies pose to the Department. S&T is guided by the principles set forth in [Executive Order 13859 “Maintaining American Leadership in Artificial Intelligence”](#) (February 11, 2019) and [Executive Order 13960 “Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government”](#) (December 3, 2020), as well as the National Institute of Science and Technology report [U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools](#) (August 9, 2019). Further, S&T acts in accordance with the [AI principles](#) advanced by Organization for Economic Cooperation and Development (OECD) member countries and adopted by the G20.

These principles will inform the development of a subsequent S&T AI/ML Implementation Plan that describes a roadmap and governance approach for achieving the goals, objectives, and outcomes outlined in this S&T AI/ML Strategic Plan.

## V. GOALS

The DHS AI Strategy (December 3, 2020) sets out five goals to govern the Department's approach to integrating AI into the DHS mission in a responsible and trustworthy manner and successfully mitigating risks associated with AI across the HSE. The DHS AI goals are:





In addition to aligning with efforts across DHS and the HSE, S&T will continue to engage interagency, academic, industry, and international partners to offer evidence-based guidance for HSE mission and support challenges. S&T establishes three strategic goals to advance the organization's role in AI/ML at DHS:

## — S&T AI/ML GOALS —

**1**

### DRIVE

NEXT-GENERATION  
AI/ML TECHNOLOGIES  
FOR CROSS-CUTTING  
HOMELAND SECURITY  
CAPABILITIES

**2**

### FACILITATE

USE OF PROVEN AI/ML  
CAPABILITIES IN  
HOMELAND SECURITY  
MISSIONS

**3**

### BUILD

AN INTERDISCIPLINARY  
AI/ML TRAINED  
WORKFORCE





## Goal 1: Drive Next-Generation AI/ML Technologies for Cross-Cutting Homeland Security Capabilities

S&T partners with academia, industry, international and federal, state, and local partners to make research investments that leverage AI/ML breakthroughs for Homeland Security needs. S&T Strategic Goal One aligns with DHS AI Strategy Goal Two: *Invest in DHS AI Capabilities*. The priority areas for S&T's investments in AI/ML are Advancing Trustworthy AI, Human-Machine Teaming, and Secure Cyberinfrastructure. These priority areas were formulated and validated by the S&T AI/ML Working Group, based on alignment with existing and projected S&T and Component needs.

### **Objective 1A: Advance Trustworthy AI**

Advancing Trustworthy AI is a broad research area examining how to ensure confidence in AI/ML systems. This research area is critical for DHS and the HSE for multiple reasons:

- To enable DHS leaders and managers to effectively assess the performance of AI/ML systems against both technical and mission metrics, as well as meeting applicable legal and policy requirements;
  - To provide operators making critical decisions an appropriate level of trust and confidence in any AI/ML systems incorporated into their mission; and
  - To inspire trust in the general public towards AI/ML systems deployed by DHS.
  - The problems addressed in this area of research are not only technical, but are truly multi-faceted and require the full spectrum of the social sciences, as well as policy, legal, privacy, civil rights and civil liberties, and ethics research to develop governance approaches that build trust within DHS, the HSE, and broader society.
- There is considerable overlap among the critical areas included in Advancing Trustworthy AI.

**1A.i Advance Explainable AI:** S&T will support research in Explainable AI, through investments, collaborations, and knowledge-sharing, in order to promote research advances that would make responsible and trustworthy implementation of AI in DHS and HSE contexts possible and to facilitate legal and administrative processes for individuals who are affected by DHS activities that rely on AI/ML tools. Effective use of AI/ML requires explanations for how the systems work that are understandable and meaningful to DHS stakeholders, including oversight stakeholders. This requires technical research into how to effectively audit an AI/ML system to understand how it arrived at its outputs. Research into how humans understand and interpret the explanations is also necessary. This includes ensuring that the technical audit incorporates social, behavioral, and ethical considerations as well as legal, policy, and privacy requirements of DHS and HSE operations and that explanations are useful and meaningful to DHS and HSE personnel.

**1A.ii Build Privacy Protections within AI Capabilities:** Many AI/ML models are trained on and process vast amounts of data. Research on new approaches to privacy protections can benefit DHS by ensuring that the data is collected, used, maintained, and disseminated in a manner that adheres to privacy laws, regulations, and DHS policies, while maintaining public trust. S&T research will include how policies, business rules, and innovative



privacy enhancements can best ensure privacy for AI/ML capabilities. S&T will also research new and innovative privacy enhancements and protections, and how to train AI/ML models to process data in ways protect individuals' privacy. This research will inform guidance that S&T provides to Components.

**1A.iii Advance the Ability to Detect and Counter Bias in AI/ML Capabilities:** AI/ML models are trained on datasets that may possess or introduce forms of bias that can then be replicated or amplified by machine processes. This may include treating individuals in a manner inconsistent with Constitutional and legal guarantees of equality, or in some cases magnifying or exacerbating systemic social biases and creating disparate impacts. This could result in impermissible or adverse consequences for these groups, which is in fundamental contradiction to the core values and missions of DHS and the HSE: to protect the American people and uphold their values. Understanding how impermissible biases occur and detecting them prior to incorporation in an AI/ML solution is therefore a critical avenue of interdisciplinary research that S&T will undertake.

**1A.iv Ensure Trust in AI/ML Capabilities:** This area aligns with DHS AI Strategy Goal Five: *Improve Public Trust and Engagement*. It is important for those who interact with or are affected by AI/ML systems to have an appropriate level of trust. If the general public lacks confidence in AI/ML used by DHS and the HSE, then its use may undermine public trust and become a barrier to DHS and its Components performing its missions. S&T will pursue interdisciplinary research to assess how the public perceives AI/ML in homeland security applications and what methods and approaches will best build public trust. S&T believes such research can help to ensure that DHS and the HSE develops and uses AI/ML in ways that are in accord with community values and interests. Additionally, S&T will identify and leverage best practices—including those of industry—to ensure that potential implementations of AI/ML meet technical and ethical standards. If DHS does not incorporate technically and ethically sound technologies into its mission as industry and adversaries advance, DHS Component and HSE effectiveness may decline, which would negatively impact public perception.

**1A.v Counter Adversarial Uses of Machine Learning<sup>ii</sup>:** While AI/ML can bring enormous benefits, it also creates unique vectors of attack. An AI/ML model's universe is rooted in its training data. If that data is compromised, AI/ML systems can be trained in ways that result in negative outcomes. Alternately, adversaries can identify limitations in training data and take advantage of them. Additionally, the AI/ML model can also be attacked, stolen, spoofed, or modified. Research to understand, prevent, and counter adversarial machine learning is essential for developing trustworthy AI for DHS and the HSE.

## ***Objective 1B: Advance Human-Machine Teaming***

This objective is to conduct research into how humans and AI/ML can collaborate most effectively to carry out the missions of DHS and the HSE. AI/ML has very specific strengths and weaknesses – as do human beings. Understanding how AI/ML can most effectively augment human decision-making is critical to maximizing the potential benefits of AI/ML to DHS and the HSE. This area of research is interdisciplinary and requires the blending of technical research with social science.



**1B.i Optimize Human-in-the-Loop Architecture:** Given the sensitivity of DHS Component and HSE operations and the brittleness<sup>iii</sup> of AI/ML, constructing systems that fully engage human capabilities in order to get the best of both human cognition and computational processing is a critical priority. S&T will build on the extensive research in this area to develop solutions that will meet DHS mission needs.

**1B.ii Enable Collaboration Between Users and Heterogenous Architecture:** The explosive growth in sensors, processing, and data throughout society, particularly with the emergence of Internet of Things (IoT) and edge computing, creates new opportunities for DHS Components and the HSE to use data in an array of crisis and commonplace situations. This data, however, exists and will exist in varied architectures. S&T will conduct research into developing AI/ML that can – in real time – operate across varied architectures in order to leverage this data.

### ***Objective 1C: Leverage AI/ML for Secure Cyberinfrastructure***

Cyberinfrastructure “consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible.”<sup>iv</sup> Cyberinfrastructure undergirds critical infrastructure sectors, the security of which is a mission of DHS. The technological revolution that enables cyberinfrastructure also underpins AI/ML. Given the speed, scale, and computationally intensive demands of cyberinfrastructure, AI/ML is an essential tool in its security.

**1C.i Characterize Effective Model Lifetimes:** S&T will research the efficacy of AI/ML models, including how best to keep them up-to-date, in order to make recommendations and inform stakeholders on how to effectively manage AI/ML across the lifecycle.

**1C.ii Enable Rapid Threat Detection and Response:** Given the speed at which cyber threats can emerge, propagate, and evolve, strict reliance on human cognition is inadequate to process them in real-time. S&T will research, develop, test, and evaluate AI/ML capabilities that can identify and track emerging cyber threats in real time.

**1C.iii Enable Real-Time and Secure Shared Computations:** Much of the data and many of the systems critical to analyzing cyber threats and securing cyberinfrastructure have sensitivities such as security classification, PII, or proprietary information. S&T will research technical capabilities that allow sharing and processing of data across systems, while ensuring authorized access and use and not exposing sensitive information.

### ***Goal 1: Outcomes***

Some specific outcomes that will indicate progress in meeting DHS needs include:

- **Effective Model Performance:** Effective model performance relies upon well curated data sets. In most cases, S&T will endeavor to use real operational data to train AI/ML models, and to validate and verify model performance. For certain specific homeland security missions, however, collecting sufficient data to train an AI/ML model is not feasible. For these limited cases, achieving several of the objectives will require accurate, reliable models trained on synthetic data. Developing, testing, and evaluating models built with well-curated datasets will drive next-generational AI/ML technologies.

- **Security Mechanisms that Mitigate Reverse Engineering:** Understanding, preventing, and countering adversarial machine learning requires research to characterize threats and vulnerabilities in AI/ML systems. Security mechanisms that reduce risks will be a key indicator in countering adversarial machine learning. This outcome aligns with DHS AI Strategy Goal Three: Mitigating AI risks to the Department and the Homeland.
- **Human Focus Shifts to Cognitively Engaging Tasks:** One promise of AI/ML is to free humans from rote tasks in order to enable them to focus on more complex tasks that rely on human judgment. Properly implemented, AI/ML holds the potential to augment human intelligence in a variety of contexts in which well-characterized, rote tasks are currently executed by human analysts and operators. Developing performance metrics for human-machine teaming in particular DHS applications is an outcome that will enable effective development, testing, and evaluation.
- **Effective Metrics for Understanding Risk in Automated Systems:** Whether ascertaining the privacy risks of an AI/ML system or its potential for error in detecting threats, systematic measures of these risks are needed to determine if the system is meeting its requirements.





## Goal 2: Facilitate Use of Proven AI/ML Capabilities in Homeland Security Missions

This goal supports DHS AI Strategy Goal One: Assess Potential Impact of AI on the Homeland Security Enterprise and DHS AI Strategy Goal Two: Invest in DHS AI Capabilities. S&T enables DHS Components and HSE partners to assess and potentially adopt AI/ML in the near-term. S&T will work with Components to understand Component needs, and then will tailor AI/ML research to those needs. S&T's role includes identifying current technologies that can support HSE missions, supporting the Components in building the capacity to adopt AI/ML, developing test and security standards, and advising the HSE of the mission impacts of malicious or disruptive uses of AI/ML.

### ***Objective 2A: Identify, Evaluate, and Transition Existing AI/ML Capabilities for DHS Component and HSE Missions***

**2A.i Develop or Adopt Proven AI/ML Capabilities to Meet Component Needs:** In coordination with Components, S&T will work to understand Component needs and capabilities, and the policy and mission contexts within which an AI/ML system would operate, in order to tailor S&T research and development to fill capability needs.

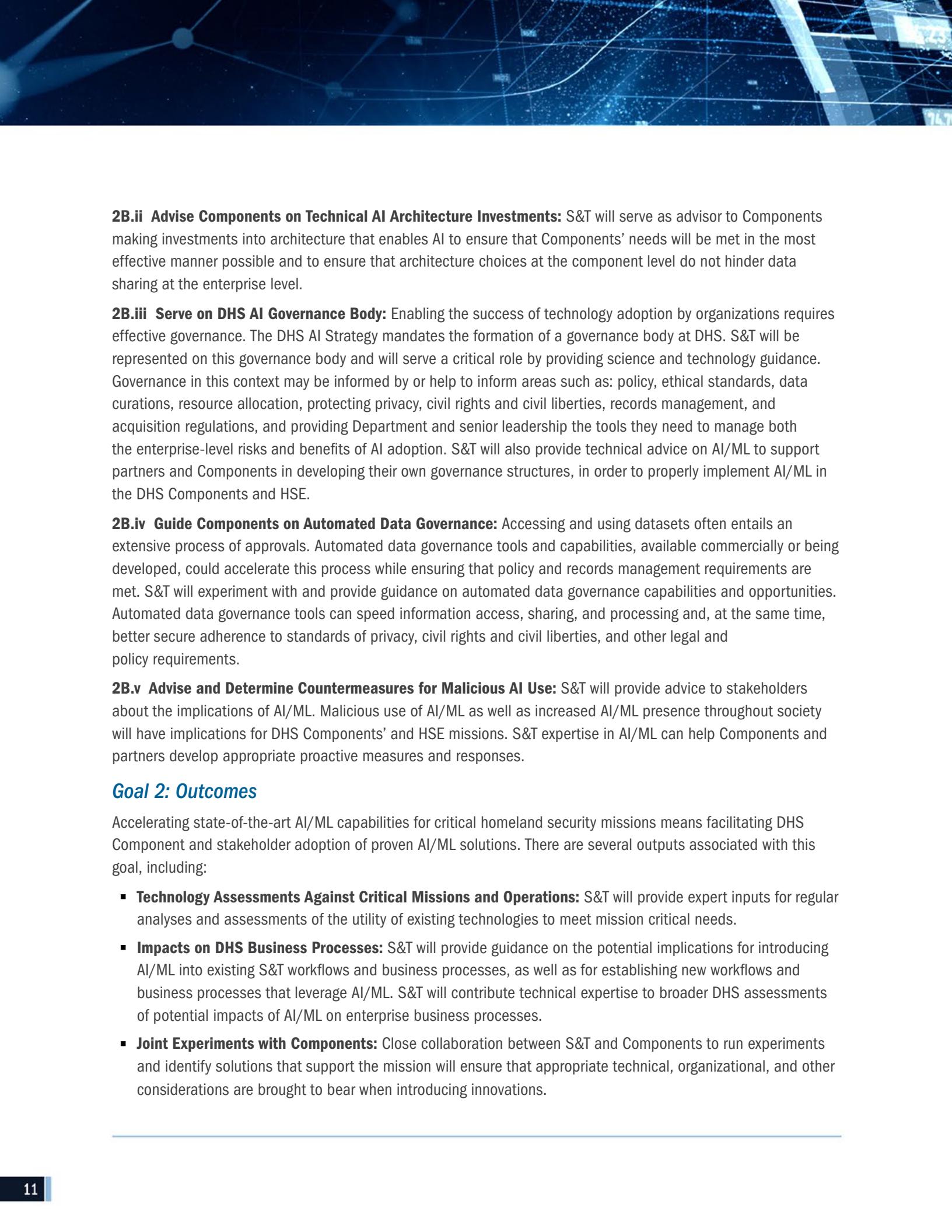
**2A.ii Identify, Evaluate, and Transition Capabilities and Inform Stakeholders:** If mature AI/ML capabilities exist, DHS must be in a position to evaluate their technical performance and potential mission impacts. This process includes matching the right technology to the appropriate mission, given the analytical maturity of the Component or partner and of their operators. This process is not only technical. Enabling the DHS Components and HSE to adopt AI/ML will need to draw upon the full spectrum of capabilities at S&T including social science, systems engineering, test and evaluation, human factors, and organizational analysis. Even if a capability is technically robust, determining if it fits the context of use, end-user requirements, and other strategic and tactical considerations is a process that necessitates the participation and judgment of diverse subject matter expertise.

**2A.iii Conduct Pilot Studies:** S&T will conduct pilot studies on promising technologies in order to rapidly experiment with AI/ML systems that may provide value to DHS Components. Such pilot studies will enable S&T to inform stakeholders about the functionalities of existing capabilities, and to potentially transition viable products to serve DHS Components' mission needs.

### ***Objective 2B: Enable AI/ML throughout DHS Components and the HSE***

S&T aims to enable innovation at every level of the HSE by identifying, evaluating, and advising on tools and capabilities that will enable Components and partners to better understand how AI/ML solutions can address mission challenges. This objective calls for a spectrum of technical investments while supporting Components in their organizational learning.

**2B.i Guide Components on Accessible AI/ML Tools:** As AI/ML advances, tools that enable its use by non-data scientists are becoming commercially available. These tools facilitate implementing AI/ML solutions in DHS and the HSE. S&T will provide guidance to Components on these capabilities, including self-service, data cleaning and prep capabilities, and accessible AI/ML tools. S&T will also inform stakeholders on potential governance issues associated with accessible AI/ML tools.



**2B.ii Advise Components on Technical AI Architecture Investments:** S&T will serve as advisor to Components making investments into architecture that enables AI to ensure that Components' needs will be met in the most effective manner possible and to ensure that architecture choices at the component level do not hinder data sharing at the enterprise level.

**2B.iii Serve on DHS AI Governance Body:** Enabling the success of technology adoption by organizations requires effective governance. The DHS AI Strategy mandates the formation of a governance body at DHS. S&T will be represented on this governance body and will serve a critical role by providing science and technology guidance. Governance in this context may be informed by or help to inform areas such as: policy, ethical standards, data curations, resource allocation, protecting privacy, civil rights and civil liberties, records management, and acquisition regulations, and providing Department and senior leadership the tools they need to manage both the enterprise-level risks and benefits of AI adoption. S&T will also provide technical advice on AI/ML to support partners and Components in developing their own governance structures, in order to properly implement AI/ML in the DHS Components and HSE.

**2B.iv Guide Components on Automated Data Governance:** Accessing and using datasets often entails an extensive process of approvals. Automated data governance tools and capabilities, available commercially or being developed, could accelerate this process while ensuring that policy and records management requirements are met. S&T will experiment with and provide guidance on automated data governance capabilities and opportunities. Automated data governance tools can speed information access, sharing, and processing and, at the same time, better secure adherence to standards of privacy, civil rights and civil liberties, and other legal and policy requirements.

**2B.v Advise and Determine Countermeasures for Malicious AI Use:** S&T will provide advice to stakeholders about the implications of AI/ML. Malicious use of AI/ML as well as increased AI/ML presence throughout society will have implications for DHS Components' and HSE missions. S&T expertise in AI/ML can help Components and partners develop appropriate proactive measures and responses.

## **Goal 2: Outcomes**

Accelerating state-of-the-art AI/ML capabilities for critical homeland security missions means facilitating DHS Component and stakeholder adoption of proven AI/ML solutions. There are several outputs associated with this goal, including:

- **Technology Assessments Against Critical Missions and Operations:** S&T will provide expert inputs for regular analyses and assessments of the utility of existing technologies to meet mission critical needs.
- **Impacts on DHS Business Processes:** S&T will provide guidance on the potential implications for introducing AI/ML into existing S&T workflows and business processes, as well as for establishing new workflows and business processes that leverage AI/ML. S&T will contribute technical expertise to broader DHS assessments of potential impacts of AI/ML on enterprise business processes.
- **Joint Experiments with Components:** Close collaboration between S&T and Components to run experiments and identify solutions that support the mission will ensure that appropriate technical, organizational, and other considerations are brought to bear when introducing innovations.

- **Well-Informed AI/ML Investments Across DHS:** S&T will generate knowledge products that inform the Components about the latest AI/ML tools, technologies, skills and techniques. Knowledge products will enable the Components to make well-informed investments in AI/ML. This outcome will also support DHS AI Strategy Objective 1.1: Develop Knowledge of Technical Applications of AI and DHS S&T AI/ML Strategy Objective 3B: Enabling Broader DHS AI/ML Competence.
- **Components use S&T Acquisition Guidance:** S&T will provide support to program offices throughout the acquisition process with regard to requirements development, analytic processes, use of standards, systems engineering, and technology readiness, and specific, actionable advice about AI/ML systems that can best meet Components' needs.
- **Engage Partners Via a Community of Practice:** For AI/ML to be adopted across DHS, S&T will need to be embedded in a Department-wide network of AI/ML experts and users who are in regular communication and sharing best practices and insights. S&T will also be represented on the GSA government-wide AI community of practice.
- **Outreach and Communications Guidance:** Building public trust in AI/ML, establishing appropriate use, and other effectively communicating with stakeholders are integral to using this technology in the Homeland Security domain. S&T, in conjunction with other headquarters elements, will play a central role in providing guidance to the Department on these outreach and communications issues.





## Goal 3: Build an Interdisciplinary AI/ML-Trained Workforce

Adopting AI/ML requires a workforce familiar and comfortable with the technology. To conduct AI/ML research and advise DHS and the HSE on AI/ML, S&T will lead the way by building a cadre of AI/ML experts with an array of disciplinary backgrounds and perspectives, including system architects, data scientists, engineers, computer scientists, social scientists, privacy professionals, ethicists, and policy analysts. S&T will also work to support DHS and the HSE as they build their own AI/ML enabled workforces. This goal aligns with DHS AI Strategy Goal Four: Develop DHS AI Workforce, as well as the S&T Strategic Plan 2021 goal to advance the S&T workforce to prepare for the future while safeguarding today. S&T will provide and communicate training opportunities in AI/ML, and will play a role in training the broader DHS workforce to promote better understanding about the opportunities and challenges of using AI/ML in DHS missions.

### ***Objective 3A: Build AI/ML S&T Workforce***

S&T will train and develop an interdisciplinary workforce, augmenting the expertise of the federal workforce with contractors, FFRDCs, national laboratories, Centers of Excellence, interns, and fellows. This requires more strategic recruitment of AI/ML talent as well as systematic efforts to retain and grow talent in order to build institutional knowledge.

**3A.i Recruit:** Given the competitive marketplace for AI/ML talent, S&T will work to establish and stabilize a pipeline for AI/ML talent to meet S&T and support DHS needs. One mechanism for building this pipeline is expanding DHS Fellowship programs that promote the development of AI/ML talent by offering faculty, post-docs, graduate students, and undergraduate students direct exposure to unique homeland security challenges. Use of the STEM (Scientific, Technical, Engineering, and Mathematics) hiring authority is another important recruitment avenue. Recruitment and hiring will be facilitated by integrating AI/ML competencies into an already established and approved job series identified by the Office of Personnel Management (OPM).

**3A.ii Retain:** S&T will improve how it communicates and delivers the value proposition of government service. S&T will explore hiring options, including offering competitive salary and opportunities and applying existing specialized DHS retention incentive plans and bonuses in order to recruit AI/ML talent. To retain talented AI/ML personnel, S&T must ensure that technical capabilities and policy frameworks are in place to allow talented experts to contribute to AI/ML research and development activities.

**3A.iii Develop:** S&T will create opportunities for highly skilled employees to develop professionally, including career path development that provides for AI/ML experts to seek new challenges within S&T, undertake details with other agencies, and participate in externships that enable the S&T workforce to keep abreast of AI/ML developments in industry, academia, and the non-profit sector.

**3A.iv Improve AI/ML Competence Across S&T Workforce:** S&T will provide training opportunities to its workforce to quickly impart some technical proficiency in AI/ML for all levels of familiarity.



## **Objective 3B: Enable Broader DHS AI/ML Competence**

For AI/ML to be broadly adopted across the HSE, the workforce must have the skills and knowledge to interact with the technology and to have informed discussions with internal and external stakeholders about it. S&T will play a central role in enabling the broader DHS workforce to understand and make use of AI/ML to fulfill their missions while adhering ethical standards and the principles of Executive Order 13960. S&T will do this by identifying and undertaking training activities and by developing criteria to evaluate technical expertise, in conjunction with DHS Office of the Chief Human Capital Officer (OCHCO) and Office of Personnel Management (OPM).

**3B.i Train the DHS Workforce:** Across the board, from executives leading agencies to frontline operators, and the array of personnel needed to support them, the DHS workforce needs familiarity with AI/ML. Given a baseline level of knowledge, those inclined will be able to innovate and see new opportunities to use AI/ML for their mission. S&T, as it undertakes its own culture shift, will carry out training and will identify AI/ML training that will enable the DHS workforce.

**3B.ii Support DHS Components in Evaluating Technical Expertise of Hires:** Just as S&T will need to expand its cadre of AI/ML professionals, Components will need their own cadre of professionals. Determining if a potential hire has the right technical skills for a position is crucial for ensuring that critical positions are properly staffed. S&T, leveraging its contacts in industry and academia, can support Components in evaluating technical expertise of candidates. New processes can be developed to assess, align, and improve the technical health of the Department. S&T, in conjunction with DHS OCHCO and OPM, will support the development of criteria for evaluating technical expertise.

## **Goal 3: Outcomes**

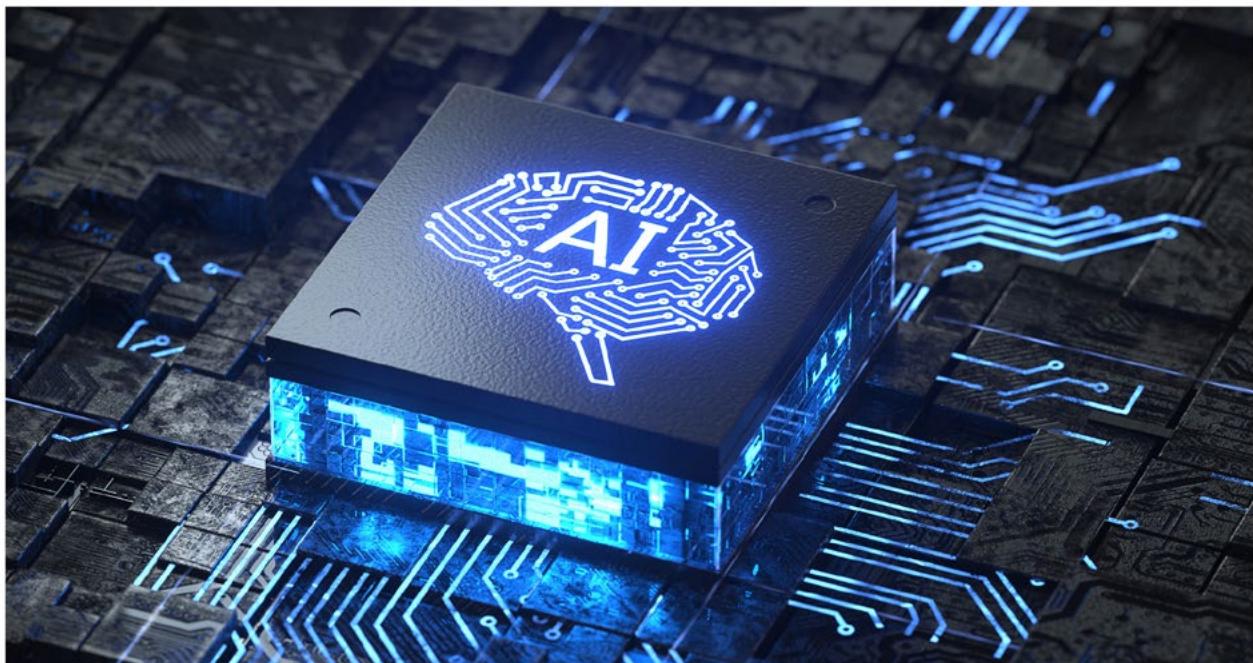
A key outcome of S&T efforts to build an interdisciplinary AI/ML-trained workforce is attracting, developing, and retaining talented AI/ML professionals. There are several other outcomes linked to building an AI/ML workforce, including those that pertain to human resources, training, and fulfillment of S&T's role as trusted science and technology advisor:

- **Provide S&T Expertise Department-wide:** Providing expert advice to DHS Components regarding AI/ML will indicate that S&T is fulfilling its role within the Department, and maintaining a strong positive reputation for providing technical and scientific advice. One important outcome will be that S&T continues to field requests for support from Components on AI/ML problems.
- **Hire Experts:** Hiring experts will ensure that S&T can continue to perform its vital mission of advising the Department.
- **Execute Externships:** Keeping abreast of developments in the rapidly changing AI/ML domain will require S&T experts to regularly spend time in academia and industry. An AI/ML externship program will enable S&T to facilitate these exchanges.
- **Host Interns:** S&T will continue to host interns with engineering, computer science, data science, artificial intelligence, and machine learning backgrounds. Internship programs are a pipeline for bringing AI/ML talent to S&T.

- **Offer Post-Internship Opportunities:** Developing opportunities for interns to remain engaged with homeland security topics will extend the talent pipeline by encouraging AI/ML professionals to continue to contribute to public service.
- **Offer S&T AI/ML Training and Enrichment Opportunities:** S&T will continue to offer seminars and training on the full spectrum of AI/ML applications and issues (including technical, policy, ethical, and societal dimensions) for all levels of knowledge and experience.

## VI. CONCLUSION

The S&T AI/ML Strategic Plan presents the vision, goals, objectives, and outcomes that constitute the S&T approach to the emerging opportunities and risks of AI/ML. For complex and rapidly evolving technologies such as AI/ML, building a robust portfolio of research and development activities and an interdisciplinary AI/ML-trained workforce in support of the DHS mission is essential. A subsequent S&T AI/ML Implementation Plan will detail how the S&T AI/ML Strategic Plan will be operationalized.





## VII. APPENDICES

### A. Acronyms

**AI** – Artificial Intelligence

**AI/ML** – Artificial Intelligence and Machine Learning

**DHS** – U.S. Department of Homeland Security

**FFRDC** – Federally Funded Research and Development Centers

**GSA** – General Services Administration

**HSE** – Homeland Security Enterprise

**IoT** – Internet of Things

**ML** – Machine Learning

**OCHCO** – DHS Office of the Chief Human Capital Officer

**OPM** – Office of Personnel Management

**PII** – Personally Identifiable Information

**R&D** – Research & Development

**S&T** – DHS Science and Technology Directorate

**STEM** – Science, Technology, Engineering, and Mathematics

**U.S.** – United States of America

### B. References

Allen, Greg, Chief of Strategy and Communications, Joint Artificial Intelligence Center (JAIC), Department of Defense. (April 2020) “Understanding AI Technology: A concise, practical, and readable overview of Artificial Intelligence and Machine Learning technology designed for non-technical managers, officers, and executives”.  
<https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf>

Department of Homeland Security, Guiding Principles  
<https://www.dhs.gov/guiding-principles>

Department of Homeland Security, Quadrennial Homeland Security Report, January 2010  
[https://www.dhs.gov/xlibrary/assets/qhsr\\_report.pdf](https://www.dhs.gov/xlibrary/assets/qhsr_report.pdf)

Department of Homeland Security Artificial Intelligence Strategy (December 3, 2020)  
[https://www.dhs.gov/sites/default/files/publications/dhs\\_ai\\_strategy.pdf](https://www.dhs.gov/sites/default/files/publications/dhs_ai_strategy.pdf)



Department of Homeland Security S&T Strategic Plan 2021

[https://www.dhs.gov/sites/default/files/publications/21\\_0121\\_st\\_strategic\\_plan\\_2021\\_final.pdf](https://www.dhs.gov/sites/default/files/publications/21_0121_st_strategic_plan_2021_final.pdf)

Executive Order, "Maintaining American Leadership in Artificial Intelligence" (EO 13859, Feb 11, 2019)

<https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence>

Executive Order, "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government" (EO 13960, Dec 3, 2020)

<https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government>

NIST, U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools (August 9, 2019)

[https://www.nist.gov/system/files/documents/2019/08/10/ai\\_standards\\_fedengagement\\_plan\\_9aug2019.pdf](https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf)

ODNI, Principles of Artificial Intelligence Ethics for the Intelligence Community

[https://admin.govexec.com/media/principles\\_of\\_ai\\_ethics\\_for\\_the\\_intelligence\\_community\\_\(1\).pdf](https://admin.govexec.com/media/principles_of_ai_ethics_for_the_intelligence_community_(1).pdf)

ODNI, Artificial Intelligence Ethics Framework for the Intelligence Community. June 2020

[https://admin.govexec.com/media/ai\\_ethics\\_framework\\_for\\_the\\_intelligence\\_community\\_1.0.pdf](https://admin.govexec.com/media/ai_ethics_framework_for_the_intelligence_community_1.0.pdf)

ODNI, The AIM Initiative: A Strategy for Augmenting Intelligence Using Machines. 2019.

<https://www.dni.gov/files/ODNI/documents/AIM-Strategy.pdf>

OECD Principles on AI

<http://www.oecd.org-going-digital/ai/principles/>

OSTP, Artificial Intelligence for the American People

<https://trumpwhitehouse.archives.gov/ai/>

Stewart, Craig A., Stephen Simms, Beth Plale, Matthew Link, David Y. Hancock, and Geoffrey C. Fox. "What is cyberinfrastructure?" In Proceedings of the 38th annual ACM SIGUCCS fall conference: navigation and discovery, pp. 37-44. 2010.

<http://dsc.soic.indiana.edu/publications/fp109a-stewart.pdf>



## C. Endnotes

<sup>i</sup>“The “Homeland Security Enterprise” refers to the collective efforts and shared responsibilities of Federal, State, local, tribal, territorial, nongovernmental, and private-sector partners as well as individuals, families, and communities—to maintain critical homeland security capabilities.” (Department of Homeland Security, Quadrennial Homeland Security Report, January 2010, pp. 12-13)

<sup>ii</sup>Adversarial Machine Learning refers to the range of techniques that can be used to manipulate an AI/ML model.

<sup>iii</sup>In the context of AI/ML systems, “brittleness” refers to the inability of algorithms to function well beyond the set of originally-defined conditions or parameters in which they were developed. AI/ML systems that effectively solve a problem under narrow constraints may not generalize or apply to other contexts.

<sup>iv</sup>Stewart, Craig A., Stephen Simms, Beth Plale, Matthew Link, David Y. Hancock, and Geoffrey C. Fox. “What is cyberinfrastructure?” In Proceedings of the 38th annual ACM SIGUCCS fall conference: navigation and discovery, pp. 37-44. 2010.



### ONLINE

[www.dhs.gov/cyber-research](http://www.dhs.gov/cyber-research)



### FACEBOOK

[Facebook.com/dhsscitech](https://Facebook.com/dhsscitech)



### EMAIL

SandT-Cyber-Liaison@hq.dhs.gov



### YOUTUBE

[www.youtube.com/dhsscitech](https://www.youtube.com/dhsscitech)



### TWITTER

@dhsscitech



### PERISCOPE

@dhsscitech



### LINKEDIN

[www.linkedin.com/company/dhsscitech](https://www.linkedin.com/company/dhsscitech)

# PROCUREMENT AS POLICY: ADMINISTRATIVE PROCESS FOR MACHINE LEARNING

*Deirdre K. Mulligan<sup>†</sup> & Kenneth A. Bamberger<sup>††</sup>*

## ABSTRACT

At every level of government, officials contract for technical systems that employ machine learning—systems that perform tasks without using explicit instructions, relying on patterns and inference instead. These systems frequently displace discretion previously exercised by policymakers or individual front-end government employees with an opaque logic that bears no resemblance to the reasoning processes of agency personnel. However, because agencies acquire these systems through government procurement processes, they and the public have little input into—or even knowledge about—their design or how well that design aligns with public goals and values.

This Article explains the ways that the decisions about goals, values, risk, and certainty, along with the elimination of case-by-case discretion, inherent in machine-learning system design create policies—not just once when they are designed, but over time as they adapt and change. When the adoption of these systems is governed by procurement, the policies they embed receive little or no agency or outside expertise beyond that provided by the vendor. Design decisions are left to private third-party developers. There is no public participation, no reasoned deliberation, and no factual record, which abdicates Government responsibility for policymaking.

This Article then argues for a move from a procurement mindset to policymaking mindset. When policy decisions are made through system design, processes suitable for substantive administrative determinations should be used: processes that foster deliberation reflecting both technocratic demands for reason and rationality informed by expertise, and democratic demands for public participation and political accountability. Specifically, the Article proposes administrative law as the framework to guide the adoption of machine learning governance, describing specific ways that the policy choices embedded in machine-learning system design fail the prohibition against arbitrary and capricious agency actions

---

DOI: <https://doi.org/10.15779/Z38RN30793>

© 2019 Deirdre K. Mulligan & Kenneth A. Bamberger.

<sup>†</sup> Associate Professor, School of Information, University of California, Berkeley; Faculty Director, Berkeley Center for Law and Technology.

<sup>††</sup> The Rosalinde and Arthur Gilbert Foundation Professor of Law, University of California, Berkeley; Faculty Director, Berkeley Center for Law and Technology. Much appreciation to Nitin Kohli for his expert input and close reading, Margot Kaminski for her detailed comments on an earlier draft, participants at the ACM FAT\* conference, members of the UC Berkeley Algorithmic Fairness and Opacity working group, and participants in the Simons Institute for the Theory of Computer Science Summer 2019 Cluster on Fairness for insightful and helpful comments and discussions; to the Berkeley Center for Law and Technology for its support of this project; and to Sanjana Parikh and Miranda Rutherford for their superb editing and research contributions to this Article. Research for this Article has been supported by generous funding from the US NSF INSPIRE SES1537324.

absent a reasoned decision-making process that both enlists the expertise necessary for reasoned deliberation about, and justification for, such choices, and makes visible the political choices being made.

Finally, this Article sketches models for machine-learning adoption processes that satisfy the prohibition against arbitrary and capricious agency actions. It explores processes by which agencies might garner technical expertise and overcome problems of system opacity, satisfying administrative law's technocratic demand for reasoned expert deliberation. It further proposes both institutional and engineering design solutions to the challenge of policymaking opacity, offering process paradigms to ensure the "political visibility" required for public input and political oversight. In doing so, it also proposes the importance of using "contestable design"—design that exposes value-laden features and parameters and provides for iterative human involvement in system evolution and deployment. Together, these institutional and design approaches further both administrative law's technocratic and democratic mandates.

## TABLE OF CONTENTS

I.	INTRODUCTION .....	784
II.	THE PROCUREMENT MINDSET: A MISMATCH FOR MACHINE LEARNING ADOPTION.....	791
A.	THE ALGORITHMIC TURN IN GOVERNANCE.....	791
B.	CHALLENGES OF ALGORITHMIC GOVERNANCE: VALUES IN TECHNOLOGY DESIGN.....	794
C.	EXAMPLES: POLICY IN SYSTEM DESIGN.....	798
1.	<i>Optimization Embeds Policy</i> .....	798
2.	<i>Decisions About Target Variables Embed Policy</i> .....	801
3.	<i>The Choice of Model Embeds Policy</i> .....	802
4.	<i>Choosing Data on Which to Train a Model Embeds Policy</i> .....	803
5.	<i>Decisions About Human-System Interactions Embed Policy</i> .....	805
III.	BRINGING MACHINE-LEARNING SYSTEM DESIGN WITHIN ADMINISTRATIVE LAW.....	808
A.	ADMINISTRATIVE PROCESS FOR MACHINE LEARNING DESIGN.....	808
B.	A FRAMEWORK FOR REASONED DECISION MAKING ABOUT MACHINE LEARNING DESIGN .....	811
1.	<i>Determining What System Choices Should Require Reasoned Decision Making</i> .....	812
a)	Design Choices that Limit Future Agency Discretion....	813
b)	Normative Choices Between “Methods of Implementation” .....	817
c)	Application to Machine Learning Systems.....	819
2.	<i>Designing Agency Decision Making: Reflecting the Technocratic and Democratic Requirements of Administrative Law</i> .....	820
a)	Technocratic Elements in Reasoned Decision Making About Machine Learning Systems.....	820
i)	<i>Citron’s Concerns: Displacement of Expert Agency Judgment</i> .....	821
ii)	<i>Updating Concerns: How Machine Learning Displaces Rational Expert Agency Decision Making</i> .....	822
a.	<i>Element 1: Delegating “Logic-Making” to Machines</i> .....	822
b.	<i>Element 2: Constraints on Policymaking Evolution</i> .....	824
iii)	<i>The Challenge: Reintroducing Expert Justification for Agency Decisions</i> .....	825
b)	Democratic Elements in Reasoned Decision Making About Machine Learning Systems .....	828
IV.	BUILDING ADMINISTRATIVE PROCESS FOR MACHINE LEARNING .....	829
A.	INFORMING AGENCY DELIBERATION WITH TECHNICAL EXPERTISE .....	831
1.	<i>Reviewing Piecemeal Efforts</i> .....	831

2.	<i>A Paradigm for Expert Decision Making</i> .....	836
a)	The Institutional Paradigm: USDS and the 18F “Skunk Works” .....	837
b)	Models to Inform the Centralized Process .....	840
B.	INFUSING AGENCY DELIBERATION WITH POLITICAL VISIBILITY ....	842
1.	<i>Impact Assessments: Bridging Technocracy and Democracy in Agency Deliberation</i> .....	842
2.	<i>Other Political Visibility-Enhancing Processes</i> .....	845
a)	Fostering Ongoing Public Engagement Through Agenda-Setting.....	847
b)	Fostering Public Engagement on Specific Systems .....	848
3.	<i>Contestable Design</i> .....	850
a)	Design Should Expose Built-in Values.....	852
b)	Design Should Trigger Human Engagement.....	854
c)	Design Should Promote Contests About Social and Political Values.....	856
V.	CONCLUSION .....	857

## I. INTRODUCTION

The U.S. Solicitor General’s 2017 arguments opposing Supreme Court review of *Loomis v. Wisconsin*,<sup>1</sup> a case presenting the constitutionality of the use of risk assessment software—software that uses statistical models to predict the likelihood of an individual failing to appear at trial or engaging in future criminal activity—in sentencing, may have prevailed in convincing the Justices to deny the petition for certiorari.<sup>2</sup> The Solicitor General conceded that one of the issues raised in the case—“the extent to which actuarial assessments considered at sentencing” may take gender into account—is a serious constitutional question.<sup>3</sup> Yet he argued that Mr. Loomis’s challenge to the use of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system used by the State of Wisconsin in sentencing was “not a suitable vehicle” for Supreme Court review because “it is unclear *how* COMPAS accounts for gender.”<sup>4</sup>

---

1. See Brief for the United States as Amicus Curiae, *Loomis v. Wisconsin*, 138 S. Ct. 2290 (2017) (No. 16-6387), <https://www.scotusblog.com/wp-content/uploads/2017/05/16-6387-CVSG-Loomis-AC-Pet.pdf> [<https://perma.cc/L98E-8AVH>]; see also *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016). The Wisconsin Supreme Court case generated petition for certiorari. *Id.*

2. See *Order List*: 582 U.S., SUP. CT. U.S. 5 (June 26, 2017), [https://www.supreme court.gov/orders/courtorders/062617zor\\_8759.pdf](https://www.supreme court.gov/orders/courtorders/062617zor_8759.pdf) [<https://perma.cc/X85J-PGRK>].

3. Brief for the United States, *supra* note 1, at 19.

4. *Id.*

Yet, however persuasive this argument might have been in the context of Supreme Court case management, the implications of this concession are shocking as a matter of policy. At no time during the challenge, which was appealed all the way to the Wisconsin Supreme Court, could the courts even determine how constitutionally relevant variables were used in the system's analysis.<sup>5</sup> More significantly, it is unclear whether the government ever deliberated about—or was even fully aware of—the way gender was used during the procurement of this system, or its application in the sentencing over thousands of cases.<sup>6</sup> The state asserted that it used “the same COMPAS risk assessment on both men and women, but then compares each offender to a ‘norming’ group of his or her own gender.”<sup>7</sup> In the end, however, all evidence suggests that the State of Wisconsin left the decision of how gender was to be used at the discretion of the software vendor.

While deeply troubling, this phenomenon is widespread. At every level of government, officials purchase, or contract for use of, technology systems that employ machine learning—systems that perform tasks without using explicit instructions, relying on patterns and inference instead. These systems

---

5. This is particularly striking because regardless of how gender is used, the decision would not constitute a trivial detail, as under the Due Process Clause, a sentencing court may not consider as “aggravating” factors characteristics of the defendant “that are constitutionally impermissible or totally irrelevant to the sentencing process, such as for example race, religion, or political affiliation.” *Zant v. Stephens*, 462 U.S. 862, 885 (1983). The Supreme Court of Wisconsin too prohibits the use of gender as a sentencing factor. *See State v. Harris*, 786 N.W.2d 409, 416 (Wis. 2010).

6. The court record does not document any evidence of such deliberation, and we could find no evidence of such deliberation elsewhere. In fact, there are indications that the state had not even adopted high level guidelines for the design of tools. SUZANNE TALLARICO ET AL., NAT'L CTR. FOR STATE COURTS, EFFECTIVE JUSTICE STRATEGIES IN WISCONSIN: A REPORT OF FINDINGS AND RECOMMENDATIONS, 122 (2012), <https://www.wicourts.gov/courts/programs/docs/ejsreport.pdf> [<https://perma.cc/L78K-VSRT>] (suggesting that draft standards developed by a national coordinating network, which require risk tools to be “equivalently predictive for racial, ethnic and gender sub-groups represented in the Drug Court population,” “could serve as a model for standards *should the state of Wisconsin wish to develop them*”) (emphasis added). It is, moreover, difficult to assess what courts are doing to consider the embedded policies in these tools, even with substantial effort. *See generally* Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103, 137–38 (2018) (reporting that only one of sixteen courts provided any information about a risk assessment tool (not COMPAS) in response to public records acts, with most claiming to be exempt).

7. *Loomis*, 881 N.W.2d at 765. The Practitioner’s Guide provided by Northpointe does not mention norming. Wisconsin may be referring to either what Northpointe calls “normative subgroups,” which include (1) male prison/parole, (2) male jail, (3) male probation, (4) male composite, (5) female prison/parole, (6) female jail, (7) female probation, and (8) female composite. *Practitioner’s Guide to COMPAS Core*, NORTHPOINTE 1, 11–12 (Mar. 19, 2015), [http://www.northpointeinc.com/files/technical\\_documents/Practitioners-Guide-COMPAS-Core-\\_031915.pdf](http://www.northpointeinc.com/files/technical_documents/Practitioners-Guide-COMPAS-Core-_031915.pdf) [<https://perma.cc/775A-6GMH>].

frequently displace discretion previously held by either policymakers charged with ordering that discretion, or individual front-end government employees on whose judgment governments previously relied, with an opaque logic that bears no resemblance to the bounded and rational reasoning processes of agency personnel, but rather by patterns that machines induce by observing human actions.<sup>8</sup>

However, research reveals that government agencies purchasing and using these systems most often have no input into—or even knowledge about—their design or how well that design aligns with public goals and values. They know nothing about the ways that the system models the phenomena it seeks to predict, the selection and curation of training data, or the use of that data—including (as in the *Loomis* case) whether and how to use data that relate to membership in a protected class. And agencies have no input into the system’s analytic technique, treatment of risk or uncertainty, preferences for false positives or false negatives, or confidence thresholds. In short, governments play no role in setting important policy.

Indeed, in a recent study by Robert Brauneis and Ellen Goodman involving open records requests seeking information about six algorithmic programs used by forty-two different agencies in twenty-three states, only one jurisdiction provided the algorithm and details about its development.<sup>9</sup> In most instances, by contrast, agency documents revealed that they did not have access to the algorithm, the model’s design, or the processes through which the algorithm was generated or adjusted.<sup>10</sup> Indeed, most government bodies did not even have a “record of what problems the models were supposed to address, and what the metrics of success were.”<sup>11</sup>

Algorithmic systems generally, and those that design and sell them, are increasingly subject to criticism for inattention to context and culture, the values baked into their design, and the biases they embed.<sup>12</sup> Yet government

---

8. See *infra* Section III.B.1 (discussing decision making by machine learning systems).

9. Brauneis & Goodman, *supra* note 6, at 137 (“[O]nly one of the jurisdictions, Allegheny County, was able to furnish both the actual predictive algorithms it used (including a complete list of factors and the weight each factor is given) and substantial detail about how they were developed.”).

10. *Id.*

11. *Id.* at 152.

12. See Mary Flanagan et al., *Embodying Values in Technology: Theory and Practice*, in INFORMATION TECHNOLOGY AND MORAL PHILOSOPHY 322, 322–47 (Jeroen van den Hoven & John Weckert eds., 2008) (arguing that technology can embody values by design and developing a framework for identifying moral and political values in such technology); Deirdre K. Mulligan & Kenneth A. Bamberger, *Saving Governance-By-Design*, 106 CALIF. L. REV. 697, 708–13 (2018) (discussing the science and technology studies as well as computer science and legal literatures on “Values in Design”); Lucas D. Introna & Helen Nissenbaum, *Shaping the Web: Why the Politics of Search Engines Matters*, 16 INFO. SOC’Y 169, 169–85 (2000) (discussing

agencies seeking to automate tasks left to their discretion seem persistently tone deaf to the need for greater agency and public participation in shaping technology systems. Across the country there is a smattering of public efforts to assess the policies embedded in algorithmic systems, but these are exceptions. A January 2019 Request for Proposal (RFP) issued by the Program Support Center of the U.S. Department of Health and Human Services sought a contractor who could in turn coordinate the procurement of Intelligent Automation/Artificial Intelligence (IAAI) on behalf of a range of agencies.<sup>13</sup> In the words of the proposal, “[t]his contract is the next logical step to integrating IAAI technologies into all phases of government operations.”<sup>14</sup> This RFP reflects the dominant mindset of agencies: It positions machine learning systems as machinery used to support some well-defined function, rather than new methods of arranging how an institution makes sense of and executes on its mission, which is often tied to an empiricist epistemology where prediction, rather than causation, is a sufficient justification for action.<sup>15</sup>

The marked absence of a public sector culture of algorithmic responsibility reflects a “procurement” mindset that is deeply embedded in the law of public administration. Technology systems are acquired from third-party vendors with whom government agencies enter into contracts for goods or services. Public procurement is governed by an extensive body of regulation intended to promote certain bureaucratic values—including price, fairness in the bidding process, innovation, and competition<sup>16</sup>—and elaborates methods of challenging contracting decisions on these elements. This body of regulation generally limits standing to challenge contracting decisions to jilted commercial

---

biases in the creation of search indexes and search results); James H. Moor, *What is Computer Ethics?*, 16 METAPHILOSOPHY 266, 266–75 (1985) (discussing the ethical implications of invisible abuse, emergent bias due to designers’ values, and bias rooted in complexity within computer systems).

13. See Aaron Boyd, *HHS Contract Will Offer AI Tech, Support to All of Government*, NEXTGOV.COM (Jan. 10, 2019), <https://www.nextgov.com/emerging-tech/2019/01/hhs-contract-will-offer-ai-tech-support-all-government/154078/> [https://perma.cc/W8NH-CYHY].

14. *Solicitation/Contract/Order for Commercial Items: Solicitation Number 19-233-SOL-00098*, U.S. DEP’T HEALTH & HUM. SERVS. 9 (Jan. 10, 2019) <https://www.fbo.gov/utils/view?id=39d0a0ce8bfe09391b9fee07833274de> [https://perma.cc/6DEC-L5WQ] [hereinafter *Solicitation Number 19-233-SOL-00098*].

15. Rob Kitchin, *Big Data, New Epistemologies and Paradigm Shifts*, BIG DATA & SOC’Y 3–5 (2014), <https://doi.org/10.1177/2053951714528481> [https://perma.cc/3N7Q-3LYG] (describing and critiquing Big Data “empiricism, wherein the volume of data, accompanied by techniques that can reveal their inherent truth, enables data to speak for themselves free of theory”).

16. See generally Steven L. Schooner, *Desiderata: Objectives for a System of Government Contract Law*, 11 PUB. PROCUREMENT L. REV. 103 (2002) (summarizing nine goals identified for government procurement systems: competition, integrity, transparency, efficiency, customer satisfaction, best value, wealth distribution, risk avoidance, and uniformity).

competitors. Both public contracting and decision making about agency management are largely exempted from administrative procedures that govern decisions of policy<sup>17</sup>—procedures intended to promote a different set of public values: substantive expertise, transparency, participation and political oversight, and reasoned decision making. Thus, current agency perception and practice leave the policies that algorithms embed obscured, unarticulated, and unvetted.

This Article makes the case that because choices in the design, adoption, and use of machine learning systems often make substantive policy, design, adoption, and use should be approached with a different mindset—a “policymaking” mindset—and should reflect the frameworks for legitimate policymaking embodied in administrative law.

Designing algorithmic and machine learning systems involves decisions about goals, values, risk and certainty, and a choice to place constraints on future agency discretion. If these systems employ adaptive machine learning capabilities, their design choices make policy—not just once when they are designed, but over time as they adapt and change. When the adoption of those systems is governed by procurement, the policies they embed receive little or no agency or outside expertise beyond that provided by the vendor: no public participation, no reasoned deliberation, and no factual record. Design decisions are left to private third-party developers. Government responsibility for policymaking is abdicated.

An important body of scholarship has explored the possibilities and shortcomings inherent in algorithmic systems,<sup>18</sup> suggested ways in which individual government determinations based on algorithmic systems might be challenged,<sup>19</sup> and proposed methods for increasing transparency and

---

17. See, e.g., 5 U.S.C. §§ 553(a)(2)–(3) (2012) (containing the Administrative Procedure Act’s exemption of matters relating to “agency management” or to “public property, loans, grants, benefits, or contracts” from the section’s general requirements of notice-and-comment rulemaking).

18. See, e.g., FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015) [hereinafter BLACK BOX]; Jane Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1 (2018); Kenneth A. Bamberger, *Technologies of Compliance: Risk and Regulation in a Digital Age*, 88 TEX. L. REV. 669, 724 (2010); Peter A. Winn, *Judicial Information Management in an Electronic Age: Old Standards, New Challenges*, 3 FED. CTS. L. REV. 135 (2009); Guy Stuart, *Databases, Felons, and Voting: Bias and Partisanship of the Florida Felons List in the 2000 Elections*, 119 POL. SCI. Q. 453 (2004); Kate Crawford, *The Hidden Biases in Big Data*, HARV. BUS. REV. (Apr. 1, 2013), <https://hbr.org/2013/04/the-hidden-biases-in-big-data> [https://perma.cc/MH6U-28M2]; Julia Angwin et al., *Machine Bias: There’s Software Used Across the County to Predict Future Criminals. And it’s Biased Against Blacks*, PROPUBLICA (May 23, 2016), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [https://perma.cc/WK73-BW9S].

19. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1252 (2008).

accountability.<sup>20</sup> Fewer researchers have extended these insights to accommodate the pressing challenges of machine learning,<sup>21</sup> and even fewer have explored what moving technology systems acquisition and design from a “procurement” mindset to a “policymaking” mindset would mean in terms of technical design, administrative process, participation, and deliberation.<sup>22</sup>

This Article begins to fill that gap. It argues that, in contexts in which policy decisions are likely to be made through procurement, process suitable for substantive administrative determinations should be used: process ensuring the type of deliberation that safeguards fundamental administrative law values. Such processes must satisfy administrative law’s *technocratic* demands that policy decisions be the product of reasoned justifications informed by expertise—

---

20. See Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 633 (2017) (suggesting a “technological toolkit to verify that automated decisions comply with key standards of legal fairness”); Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. TELECOMM. & HIGH TECH. L. 235, 235–36 (2011); Katherine Fink, *Opening the Government’s Black Boxes: Freedom of Information and Algorithmic Accountability*, INFO., COMM. & SOC’Y 1–19 (May 30, 2017), <https://doi.org/10.1080/1369118X.2017.1330418> [<https://perma.cc/ATP4-KRZ8>] (reviewing current state of law and practice with respect to whether algorithms would be considered “records” under the Freedom of Information Act and reviewing agency bases for withholding algorithms and source code under FOIA requests); see also Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 121 (2019) (arguing that recently introduced provisions protecting employees against trade secret actions could immunize whistleblowers policing algorithms from within firms).

21. See, e.g., Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1 (2019); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147 (2017); Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399 (2017).

22. Margot E. Kaminski, *Binary Governance: Lessons from the GDPR’s Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 6, 26–30 (2019) (proposing a regulatory toolkit to govern the use of algorithms in the private sector, including “substantive rulemaking mechanisms, such as the use of safe harbors and private sector codes of conduct, and accountability mechanisms, such as the use of oversight boards and audits”); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 110 (2017) (calling for “criminal justice expertise and political process accountability” to be brought into the design of recidivism risk tools); Andrew D. Selbst, *Disparate impact in big data policing*, 52 GA. L. REV. 109, 109 (2017) (recommending police be required to complete “algorithmic impact statements” before adopting predictive policing technology); Catherine Crump, *Surveillance Policy Making By Procurement*, 90 WASH. L. REV. 1595 (2016) (proposing steps to strengthen democratic input); Dillon Reisman et al., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*, AI NOW INST. (Apr. 2018), <https://ainowinstitute.org/aiareport2018.pdf> [<https://perma.cc/N6W6-JRHQ>]. Danielle Citron’s work examining a prior generation of expert systems has provided foundational analysis for thinking about ways that administrative law concerns about delegation and process might be translated to the technological context. Citron, *supra* note 19, at 1252.

elements grounded in the rule of law.<sup>23</sup> And they must reflect *democratic* requirements of public involvement and political accountability. The Article thus makes the case that the policies designed into machine learning systems adopted by government agencies must be surfaced and deliberated about through new processes and brought fully within an administrative law mindset. Governance through technology cannot be allowed to quietly route around the processes that ground agency action's legitimacy.

Part II describes the ways that the integration of machine learning into governance has been viewed as a matter of procurement and the failures of that approach. Government agencies have relied on private vendors for the design of algorithmic systems, largely exacerbating the challenges of governing through technology by abdicating government's role in shaping important design choices. It then explores five examples of ways in which system design embeds policy decisions to make the case that machine-learning system adoption should often instead be understood as policymaking.

Part III examines administrative law as an alternative framework for the adoption of machine learning in governance. Describing the specific ways in which machine learning systems displace administrative discretion and human logic, this Part argues that the policy choices embedded in system design fail the prohibition against arbitrary and capricious agency actions absent a reasoned decision-making process that enlists the expertise necessary for reasoned deliberation, provides justifications for such choices, makes visible the political choices being made, and permits iterative human oversight and input. This Part focuses on changing the system-adoption process, arguing that design choices should occur through a decision-making process that reflects the technocratic and democratic goals of administrative law.

Finally, Part IV envisions what models for machine learning adoption processes that satisfy the prohibition against arbitrary and capricious actions might look like. It first explores processes by which agencies might overcome the problems of system opacity and their own lack of technical expertise, satisfying administrative law's technocratic demand for reasoned expert deliberation. Specifically, we urge the reliance on centers of expertise—on the model of the U.S. Digital Services Team (USDS) and the 18F “skunk works” team first developed by the Obama Administration—that develop and provide shared technical knowledge in ways that address expertise gaps across agencies, while providing a systemic approach to the use of technology in government activity.

---

23. Kevin M. Stack, *An Administrative Jurisprudence: The Rule of Law in the Administrative State*, 115 COLUM. L. REV. 1985, 1989 (2015) (grounding reasoned justification as a rule-of-law requirement).

Part IV explores both institutional and engineering design solutions to the challenge of policymaking opacity, offering process paradigms to ensure the “political visibility” required for public input and political oversight as well as proposing the importance of using “contestable design.” Contestable systems foster user engagement by exposing value-laden features and parameters, and provide for iterative human involvement in system evolution and deployment in a way that would foster agency staff’s awareness and participation as policies embedded in systems evolve dynamically. Together, these institutional and design approaches further administrative law’s democratic mandate.

Where machine learning systems “learn” and “exercise discretion” in ways that are not guided by reasoned human decision-making inputs, and then make substantive policy that alters the legal rights and responsibilities of individuals, policymaking fails the touchstone obligation that agency actions not be “arbitrary and capricious.” It shirks the requirement that decisions reflect reason, facts, context, and the factors mandated by Congress in the relevant organic statute, while avoiding elements extraneous to the legislative command.<sup>24</sup> The current adoption of such systems through procurement processes threatens the very premises for administrative delegation to agencies and deference to their decisions, such as expertise, reasoning, flexibility, and accountability. We outline a path to realign these powerful new tools with democratic ideals.

## II. THE PROCUREMENT MINDSET: A MISMATCH FOR MACHINE LEARNING ADOPTION

### A. THE ALGORITHMIC TURN IN GOVERNANCE

The State of Wisconsin’s decision to purchase the COMPAS system from the Northpointe Corporation reflects an accelerating public administration trend. The increasing availability of machine-learning products and services has ushered in reliance on algorithmic decision-support and decision-making systems throughout all levels of government.<sup>25</sup> Agencies increasingly recognize the promise and power of systems employing artificial intelligence and machine learning for augmenting human administrative capacity. On the one hand, they more accurately and effectively analyze and learn from data while identifying and managing risk;<sup>26</sup> on the other, they “reduc[e] repetitive

---

24. Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 42–43 (1983).

25. Mulligan & Bamberger, *supra* note 12.

26. Coglianese & Lehr, *Transparency*, *supra* note 21, at 6 (describing how privately developed algorithms produce “unparalleled accuracy” compared to other statistical methods and human judgment).

administrative tasks,” thus freeing government “employees to focus their time and human capacity on higher value activities and decisions.”<sup>27</sup>

Artificial intelligence and algorithmic systems are being employed, on the one hand, to better automate processes like data management and procurement, and on the other, to automate decision making across a range of substantive contexts. Federal agencies have employed technology systems for tasks ranging from determining the level of different veterans’ disabilities for purposes of compensation<sup>28</sup> to identifying fraud in a variety of public benefits programs.<sup>29</sup> States and localities rely on a wide range of analytic systems “to generate predictive models to guide the allocation of public services”;<sup>30</sup> to govern individual determinations such as teacher evaluations, bonuses, and terminations,<sup>31</sup> and to identify the risk that children are victims of abuse or

---

27. *Solicitation Number 19-233-SOL-00098*, *supra* note 14, at 7 (quoting OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, OMB BULL. NO. M-18-23, SHIFTING FROM LOW-VALUE TO HIGH-VALUE WORK (2018)).

28. See Bob Brewin, *Goodbye paper: VA Installs Automated Claims System in All Regional Offices*, NEXTGOV.COM (Jan. 10, 2019), <http://www.nextgov.com/health/2013/06/goodbye-paper-va-installs-automated-claims-system-all-regional-offices/65030/> [<https://perma.cc/7D9D-JC6P>]; Marion-Florentino Cuéllar, *The Surprising Use of Automation by Regulatory Agencies*, REG. REV. (Dec. 20, 2016), <https://www.theregreview.org/2016/12/20/cuellar-surprising-use-of-automation-agencies/> [<https://perma.cc/HUL3-Y6NJ>] (“[T]he software took over this responsibility for determining levels of disability from Department ‘raters’—human beings charged with determining a claimant’s entitlements.”).

29. See Leon Erlanger, *The Tech HHS, SEC, SSA and Other Agencies Use to Ferret Out Cheaters and Crooks*, FEDTECHMAGAZINE.COM (May 10, 2017), <https://fedtechmagazine.com/article/2017/05/tech-hhs-sec-ssa-and-other-agencies-use-ferret-out-cheaters-and-crooks> [<https://perma.cc/QC55-7HGA>] (discussing the ALERT program, intended to identify suspicious transactions under the SNAP (“Supplemental Nutrition Assistance Program”) food stamps program, and the Social Security Administration’s application to disability benefits); see also Coglianese & Lehr, *Regulating by Robot*, *supra* note 21, at 1162–67 (discussing initiatives within the Post Office, the Environmental Protection Agency, the Internal Revenue Service, the Federal Aviation Administration, and the Food and Drug Administration).

30. Brauneis & Goodman, *supra* note 6, at 107; see Coglianese & Lehr, *Regulating by Robot*, *supra* note 21, at 1161 (providing examples); Kimberly A. Houser & Debra Sanders, *The Use of Big Data Analytics by the IRS: Efficient Solutions or the End of Privacy as We Know It?*, 19 VAND. J. ENT. & TECH. L. 817 (2017) (discussing different ways the Internal Revenue Service uses data mining to solve the problem of tax noncompliance).

31. See Marissa Cummings, *Federal Lawsuit Settled Between Houston’s Teacher Union and HISD*, HOUS. PUB. MEDIA (Oct. 10, 2017), <https://www.houstonpublicmedia.org/articles/news/2017/10/10/241724/federal-lawsuit-settled-between-houstons-teacher-union-and-hisd/> [<https://perma.cc/JA6A-G3MW>] (discussing artificial intelligence system used by the City of Houston, the “Education Value-Added Assessment System (EVAAS), which made decisions about teacher evaluations, bonuses, and terminations based on variables including student’s performance on prior standardized tests); see also *Settlement Agreement*, AFT.ORG (Oct. 2, 2017), [https://www.aft.org/sites/default/files/settlementagreement\\_houston\\_100717.pdf](https://www.aft.org/sites/default/files/settlementagreement_houston_100717.pdf)

neglect.<sup>32</sup> Advocates have recently determined that local police departments are relying on Amazon Web Services facial-recognition product Rekognition to assist in identifying suspects,<sup>33</sup> and the FBI has announced its intention to do so as well.<sup>34</sup> Axon, a company producing body cameras, plans to introduce real-time facial recognition software into the products it provides to law enforcement;<sup>35</sup> the data obtained would be used by police departments, but retained and analyzed by Axon on Axon's own cloud services.<sup>36</sup> The Department of Homeland Security, moreover, has proposed procurement for an “extreme vetting” machine learning system that seeks to make “determinations via automation” as to whether an individual seeking a visa for entry to the United States will be a “positively contributing member of society,” will “contribute to the national interests,” or “intends to commit criminal or terrorist acts.”<sup>37</sup>

Recent work by regulation scholars Cary Coglianese and David Lehr has drawn an important roadmap of the ways in which machine learning’s capacity might enable widespread application within the administrative state. In the future they envision such systems will permit the conduct of adjudications “by algorithm” and rulemakings “by robot” without any human involvement, facilitating, for example, full automation of the antitrust review process, or real-time dynamic evolution of the Securities and Exchange Commission’s rules governing market transactions.<sup>38</sup> Regardless of whether this future is desirable, it emphasizes the broad and deep policy implications of these technical systems.

---

[<https://perma.cc/SC7W-XZEC>] (agreeing that teachers would no longer be terminated based primarily on their EVAAS score).

32. Allegheny County Department of Human Services, *Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions*, ALLEGHENY COUNTY ANALYTICS (May 1, 2019), <https://www.allegenycountyanalytics.us/index.php/2019/05/01/developing-predictive-risk-models-support-child-maltreatment-hotline-screening-decisions/> [<https://perma.cc/YPX4-J3P8>].

33. Nick Wingate, *Amazon Pushes Facial Recognition to Police. Critics See Surveillance Risk*, N.Y. TIMES (May 22, 2018), <https://www.nytimes.com/2018/05/22/technology/amazon-facial-recognition.html> [<https://perma.cc/WSF9-Q7ZK>] (discussing facial recognition use in Orlando and Washington State).

34. Frank Konkel, *The FBI is Trying Amazon’s Facial-Recognition Software*, NEXTGOV.COM (Jan. 3, 2019), <https://www.nextgov.com/emerging-tech/2019/01/fbi-trying-amazons-facial-recognition-software/153888/> [<https://perma.cc/BZL7-EX6T>].

35. Ian Wren & Scott Simon, *Body Camera Maker Weighs Adding Facial Recognition Technology*, NPR (May 12, 2018), <https://www.npr.org/2018/05/12/610632088/what-artificial-intelligence-can-do-for-local-cops> [<https://perma.cc/4JUQ-NCFK>].

36. *Id.*

37. *Extreme Vetting Initiative: Statement of Objectives (SOO)*, FEDBIZOPPS.GOV (June 12, 2017), <https://www.fbo.gov/utils/view?id=533b20bf028d2289633d786dc45822f1> [<https://perma.cc/UZQ9-9N8W>].

38. Coglianese & Lehr, *Regulating by Robot*, *supra* note 21, at 1171–84.

## B. CHALLENGES OF ALGORITHMIC GOVERNANCE: VALUES IN TECHNOLOGY DESIGN

An extensive body of research suggests the difficulties inherent in attempts to use technology systems in government decision making. As a basic matter, translating legal values into design requirements is difficult. While legal policy “tempers rule-based mandates with context-specific judgment that allows for interpretive flexibility and ongoing dispute about the appropriateness of rules, [some] computer code operates by means of on-off rules,”<sup>39</sup> and all software systems and models require up-front decisions about what data they can assess. The social and technical environment, in which regulatory norms and the norms of coders who actually design technology are “translated” into code, exacerbates divergences between “law in the books” and “law in emerging technology.”<sup>40</sup> Thus, even if technology accurately captures intended legal variables in its design, it may not clearly reflect the choice of how to respond to outputs “in a normative sense.”<sup>41</sup> As a result, depending on who is involved in the process of translation, technical solutions for enabling, enforcing, or restricting rights and values can result in unintended consequences—consequences that privilege certain stakeholders and values at the expense of others.<sup>42</sup>

Moreover, algorithmic decision-making systems are biased. They make classification decisions based on selected data that may be inadequate or unrepresentative, improperly cleaned or interpreted, and reflect historical and ongoing structures of discrimination.<sup>43</sup> For example, machine learning

---

39. Mulligan & Bamberger, *supra* note 12, at 710.

40. Mireille Hildebrandt & Bert-Jaap Koops, *D7.9: A Vision of Ambient Law*, FUTURE IDENTITY INFO. SOC’Y 22 (Oct. 4, 2007), [http://www.fidis.net/fileadmin/fidis/deliverables/fidis-wp7-d7.9\\_A\\_Vision\\_of\\_Ambient\\_Law.pdf](http://www.fidis.net/fileadmin/fidis/deliverables/fidis-wp7-d7.9_A_Vision_of_Ambient_Law.pdf) [<https://perma.cc/T47H-Y3CV>].

41. Noëmi Manders-Huys, *What Values in Design? The Challenge of Incorporating Moral Values into Design*, 17 SCI. ENG. ETHICS 271, 279 (2011) (describing what she calls “The Naturalistic Fallacy”).

42. See Alvin M. Weinberg, *Can Technology Replace Social Engineering?*, in TECH. & FUTURE 28, 34 (Albert H. Teich ed., 11th ed. 2009); Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245, 1252–69 (2016) (describing how the processes of developing and adopting technical systems in the criminal justice system, largely driven by law enforcement, produced a hyper focus on the elimination of false negatives); Eaglin, *supra* note 22, at 101–04 (describing how normative policy judgments are delegated to the developers of actuarial risk assessment tools whose incentives and preferences can produce tools in conflict with public laws and policies); see also Carsten Orwat & Roland Bless, *Values and Networks—Steps Toward Exploring Their Relationships*, 46 COMPUTER COMM. REV. 25, 28 (2016) (discussing how technology choices can shift “costs or other burdens to parties not involved in decisionmaking”).

43. Tal Z. Zarsky, *Transparent Predictions*, U. ILL. L. REV. 4 (2013); Nicholas Diakopoulos, *Algorithmic Accountability: Journalistic Investigation of Computational Power Structures*, 3 DIGITAL JOURNALISM 398 (2015); Rachel Courtland, *Bias Detectives: The Researchers Striving to Make Algorithms Fair*, NATURE (2018).

algorithms trained from human-tagged data inadvertently learn to reflect biases of the human taggers.<sup>44</sup> Two years ago, academics conducted studies showing that human annotators of data used in systems exhibit core human biases that end up decreasing the accuracy of the system at large.<sup>45</sup> On a more general scale, Lisa Gitelman in her book *Raw Data is an Oxymoron* notes that no data is free from certain bias since it is all “cooked” at some point by software, whether it be by end-users on an online platform or by back-end algorithms.<sup>46</sup> Even if the training data is adequate, tagged correctly, and minimizes inherent bias, predictive algorithms can still insert inaccuracy into a given system. Predictive algorithms are essentially autonomous profiling by a machine-learning system.<sup>47</sup> While the aim of predictive algorithms is to identify correlations and make predictions about behavior at the individual level, the system uses groups or profiles to do so. In some systems, these groups may be constantly changing as the algorithm identifies more salient patterns. This redefinition sometimes creates profiling algorithms that correlate bias in outputs.<sup>48</sup> The system’s algorithms in turn learn from such data, generating their own biases through the features they identify and the weights they place on them.

These concerns are paramount in machine learning systems that learn and adapt while in use. Such systems blur the line between implementation and policymaking.<sup>49</sup> To the extent technical systems generally are perceived as mere tools straightforwardly implementing policy choices determined elsewhere, the use of systems that change over time surely cannot fit within this fiction. In this context especially, designing for values is complicated, as the data and models, as well as the policy implications, shift at design, configuration, and run time.<sup>50</sup>

These fundamental concerns—about whether system design reflects desired values, embeds bias, or produces inaccurate results—are exacerbated by the opacity of algorithmic systems. Scholars have identified a number of

---

44. Diakopoulos, *supra* note 43, at 398.

45. Ishan Misra et al., *Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels*, 2016 IEEE CONF. ON COMPUTER VISION & PATTERN RECOGNITION (CVPR) 2930, 2930 (2016).

46. LISA GITELMAN, RAW DATA IS AN OXYMORON 2 (2013).

47. Mireille Hildebrandt & Bert-Jaap Koops, *The Challenge of Ambient Law and Legal Protection in the Profiling Era*, 73 MOD. L. REV. 428 (2010).

48. Tal Z. Zarsky, *supra* note 43, at 4.

49. Citron makes a related but distinct point that automated systems “blur the line between adjudication and rulemaking, confounding the procedural protections governing both systems.” Citron, *supra* note 19, at 1278.

50. David D. Clark et al., *Tussle in Cyberspace: Defining Tomorrow’s Internet*, 13 IEEE/ACM TRANSACTIONS NETWORKING 462, 466 (2005).

ways that these systems operate as black boxes, inscrutable from the outside:<sup>51</sup> (1) corporate secrecy, by which the design details are kept secret by private developers;<sup>52</sup> (2) technical illiteracy—the impenetrable nature of system rules to non-engineers even where they are shared; and (3) the inability of humans, even those who design and deploy machine learning systems, to understand the dynamic models learned by complex machine learning systems.

Each of these levels of opacity plague government agencies seeking to employ machine learning in governance, which most often lack the technical expertise to design or assess algorithmic systems on their own. The resulting concerns are aggravated in the context of algorithmic systems used for public governance—rather than in the private sector—as the innards of the software system that privatizes public functions are typically shielded from public scrutiny.<sup>53</sup>

Government procurement procedures are often extensive and time consuming. They are focused on promoting management goals, such as competition, integrity, transparency, efficiency, customer satisfaction, best value, wealth distribution, risk avoidance, and uniformity.<sup>54</sup> They are not focused on restricting the privatization of public functions, or providing oversight over delegated policymaking. Recent attempts to promote agency

---

51. Jenna Burrell, *How the machine ‘thinks’: Understanding opacity in machine learning algorithms*, BIG DATA & SOC’Y 3.1 (2016).

52. See Brauneis & Goodman, *supra* note 6, at 38–44 (reporting on cities’ use of trade secrecy to limit responses to Public Record Act requests for information about algorithms); *id.* at 44–47 (reporting on cities’ resisting Public Record Act requests about algorithms due to concerns about gaming or circumvention and other concerns); Rebecca Wexler, *When a Computer Program Keeps You In Jail*, N.Y. TIMES (June 13, 2017), <https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html> [<https://perma.cc/47VB-R2JU>] (discussing trade secrecy limitations on access to the algorithms used in the COMPAS system at issue in the *Loomis* case); Danielle Keats Citron, *Open Code Governance*, 2008 U. CHI. L. F. 355, 357 (2008) (“Because these systems’ software is proprietary, the source code—the programmers’ instructions to the computer—is secret.”).

53. Fink, *supra* note 20, at 1–19 (reviewing current state of law and practice with respect to whether algorithms would be considered “records” under the Freedom of Information Act (FOIA), and reviewing agency bases for withholding algorithms and source code under FOIA requests and finding exemptions claimed under national security, privacy, law enforcement investigations as well as trade secrecy exemptions).

54. Schooner, *supra* note 16, at 104. They are, moreover, notoriously burdensome and slow, especially in the context of a dynamic information technology landscape. The Federal Acquisition Regulation (FAR), 48 C.F.R. §§ 1–53, for example, which sets forth the federal procurement process, establishes best practices, procedures, and requirements for agencies, and provides standard clauses and forms. 48 C.F.R. §§ 52–53. In addition, the FAR expressly authorizes agency heads to issue agency-specific procurement regulations implementing or supplementing the FAR, meaning that agency procurement varies greatly from agency to agency. 48 C.F.R. § 1.3; see FEDERAL CHIEF INFORMATION OFFICER COUNSEL, STATE OF FEDERAL INFORMATION TECHNOLOGY 36 (2017).

technology modernization by focusing on empowering Chief Information Officers in the approval, certification, and ongoing oversight of IT systems<sup>55</sup> have not addressed the need for agency participation in system design. In many instances, the technology is a commercial off-the-shelf product purchased after a period of market research and a general solicitation.<sup>56</sup> So agencies use procurement processes to acquire these complex technical systems much as they do to purchase other goods, despite these systems' widespread implications for governance and policy.<sup>57</sup>

This “procurement mindset” reflects a number of phenomena. As an initial matter, much of the adoption of machine learning systems through procurement no doubt comes from the perspective that these systems simply supply a new process or practice for fulfilling the agency’s mission. Indeed, some such systems look more administrative in nature.

Second, procurement of off-the-shelf products often reflects a realistic assessment of agency capacity, in light of the opacity and complexity of algorithmic systems.<sup>58</sup> On the one hand, private developers keep much of the relevant code secret. On the other hand, agency staff frequently have few technical skills, so they can neither assess technology design shared with them nor participate in design themselves.

Finally, even if government bodies realize that there are important decisions embedded in systems, agencies may believe that those decisions do not constitute “policy” in the way that law traditionally understands it. Under the Federal Administrative Procedure Act, for example, matters related to agency management and contracts are both exempt from procedures that govern the adoption of policy.<sup>59</sup> Accordingly, the Internal Revenue Service (IRS), one of the few agencies to publicly address agency participation in system design—and one that relies heavily on algorithmic systems—has stated publicly that its use of “decision analytics” and “data and predictive modeling”

---

55. Federal Information Technology Acquisition Reform Act § 101(a), 40 U.S.C. § 11319 (2012); OFFICE OF PERSONNEL MGMT., OFFICE OF THE CHIEF INFO. OFFICER, FITARA COMMON BASELINE IMPLEMENTATION PLAN: FISCAL YEAR 2016 11 (2016), <https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/opm-fitara-common-baseline-implementation-plan.pdf> [<https://perma.cc/4HG3-FC64>].

56. See 48 C.F.R. § 14.101 (setting forth the “negotiated contract” process that would generally be used in the federal system for acquiring complex software involving machine-learning algorithms).

57. See, e.g., Houser & Sanders, *supra* note 30, at 865–66 (2017) (discussing findings that underlying databases used by IRS algorithms “seriously” lacked supporting documentation, implicating their accuracy).

58. Burrell, *supra* note 51.

59. See 5 U.S.C. § 553(a)(2) (2012) (excepting such matters from the requirements of informal rulemaking).

constitutes “internal enforcement policy” that does not require public feedback during its development.<sup>60</sup>

### C. EXAMPLES: POLICY IN SYSTEM DESIGN

Decisions about how to design these systems<sup>61</sup>—as well as how they are configured and how agency staff interact with them—touch on, and at times embed decisions about, traditional substantive policy questions. This Section identifies five of these types of determinations<sup>62</sup> and provides illustrative examples in which abdicating policy questions has led to real-world failures.

#### 1. *Optimization Embeds Policy*

The choice of task for which a machine learning system is designed to optimize rests on, among other things, a set of assumptions about human behavior and social structure. Using such systems to govern implicates not only questions about whether the assumptions are well-founded generally, but also about how widely-applicable they are: do they reflect all individuals, groups, or situations rather than just some? These assumptions also determine the factors that will be most salient in the choices made by algorithmic systems—the governance metrics—in much the same way traditional policy decisions identify which factors should guide any administrative decision.

The governing power of assumptions is reflected in an algorithmic system used by the United States Department of Agriculture (USDA) in the food stamp program. The system used transaction records created by the “electronic benefit transfer,” or debit cards, issued by the government to monitor stores for evidence of fraud. In 2002, based on a determination by that system, the agency disqualified several grocery stores serving predominantly Muslim East

---

60. TAXPAYER ADVOCATE SERV., 2010 ANNUAL REPORT TO CONGRESS: IRS POLICY IMPLEMENTATION THROUGH SYSTEMS PROGRAMMING LACKS TRANSPARENCY AND PRECLUDES ADEQUATE REVIEW 80 (2010), [https://www.irs.gov/pub/irs-utl/2010arcmsp5\\_policythroughprogramming.pdf](https://www.irs.gov/pub/irs-utl/2010arcmsp5_policythroughprogramming.pdf) [<https://perma.cc/3PDD-D369>] [hereinafter TAXPAYER ADVOCATE SERV.].

61. Several scholars have explicated the design processes of computer systems to reveal a broader range of interventions to address legal and policy concerns. See David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 669–700 (2017) (arguing that machine learning’s “playing-with-the-data stages” demands attention from legal scholars as the stages of problem definition, data collection, data cleaning, adjustment based on summary statistics, decisions about the portion of a data set to use for training versus testing, and model selection and training present both distinct opportunities and risks to accuracy, explainability, and discrimination, among others); Kroll, *supra* note 20 (describing how computational methods can be used throughout the design of computer systems generally to ensure procedural regularity).

62. This is an illustrative, not exhaustive, set of examples. Each of the “playing-with-the-data stages” described by Lehr and Ohm require choices that, depending upon the substantive context, may embed what administrative law would consider substantive policies.

African communities from accepting federal food stamps—a decision with significant effect, as these stores supported the religious dietary needs of families that relied on the federal program.<sup>63</sup>

Citing the system's conclusion, the agency informed the relevant shops that “a careful analysis revealed . . . transactions that establish a clear and repetitive pattern of unusual, irregular and/or inexplicable activity for your type of firm,” and on that basis eliminated them from the program.<sup>64</sup> The suspicious transactions at issue included large purchases made minutes apart, transactions for even-dollar amounts—described as unusual for food purchases—and instances in which a few households made several unusually large purchases in a single month, which is, the USDA letter stated, “not consistent with the conditions in your store for store physical size, stock of eligible items, lack of counter space and the lack of carts and baskets.”<sup>65</sup>

The model that the system deployed to identify fraud rests on a particular assumption: fraud will manifest in certain behaviors shared across groups. Yet consideration of demographically specific spending patterns would have identified that assumption’s flaws. Culture affects food purchasing habits in profound ways, rendering a one-size-fits-all model inappropriate. The purchasing patterns identified as anomalous may be normal for a subset of the population. Rather than being an indication of fraud, the patterns reflect how religion, nationality, economics, food preparation, and ordering behavior influence the purchasing behavior of a specific community.

In fact, as reporter Florangela Davila explained in an analysis of the USDA’s action,<sup>66</sup> East African immigrant women often shop in groups of two or three, which explains why transactions from the same household often occur in pairs and threes. Because East African immigrants often lack transportation, they tend to make large consecutive purchases in fewer shopping trips. It is customary to make larger purchases of Halal meat in one trip to the market, and even to buy an entire goat, spending as much as \$150 at a time, to be frozen and eaten over weeks or a month. And a habit of ordering meat by the dollar amount, rather than the pound, produces the supposedly anomalous large number of even-dollar purchases at the relevant stores.<sup>67</sup>

---

63. Florangela Davila, *USDA disqualifies three Somalian markets from accepting federal food stamps*, SEATTLE TIMES (Apr. 10, 2002), <http://community.seattletimes.nwsource.com/archive/?date=20020410&slug=somalis10m> [https://perma.cc/Q2GZ-W6BZ].

64. Chris McGann, *Somali Merchant Waits And Hopes*, SEATTLE POST-INTELLIGENCER REP. (July 1, 2002), <https://www.seattlepi.com/news/article/Somali-merchant-waits-and-hopes-1090433.php> [https://perma.cc/6KXW-W8Z9].

65. *Id.*

66. Davila, *supra* note 63.

67. *Id.*

As this episode demonstrates, because algorithms optimize over large sets of data, distinct patterns in small subpopulations are obscured by design. The decision of whether to generate one model to identify fraud across all users of electronic benefits, or separate models that attend to variations in subpopulations, is a policy decision.

The recidivism risk system at issue in *Loomis* also reflects the ways assumptions set policy. Loomis's expert witness set out several questions along these lines related to the design of the COMPAS assessment. He noted that the "Court does not know how the COMPAS compares that individual's history with the population that it's comparing them with. The Court doesn't even know whether that population is a Wisconsin population, a New York population, [or] a California population," and argued that "it is critical that it be validated for use in the jurisdiction that is planning to use it."<sup>68</sup> In fact, a report provided to the State of Wisconsin had emphasized the need for such local validation, but the state had not performed one prior to using the tool.<sup>69</sup>

A California study recommending rejection of the same COMPAS system used by Wisconsin discussed the problem of validating for relevant populations extensively, asking: "Will the results generalize to other samples?"<sup>70</sup> The study concluded that "it is unclear" whether the formulas will generalize from this New York sample of probationers to other samples of offenders.<sup>71</sup> To the extent that the predictors of recidivism differ across groups, these formulas may not work in some of the California Department of Corrections and Rehabilitation's (CDCR) primary populations of interest (e.g., inmates, parolees). Given how actuarial formulas are derived and issues

---

68. *State v. Loomis*, 881 N.W.2d 749, 756–57 (Wis. 2016).

69. At the time this tool was used to make decisions about Loomis, Wisconsin had not undertaken a local validation, despite determining its necessity. *Loomis*, 881 N.W.2d at 762 ("Wisconsin has not yet completed a statistical validation study of COMPAS for a Wisconsin population."); Suzanne Tallarico et al., *supra* note 6, at 22–23. While the Wisconsin Supreme Court allowed COMPAS risk assessments to be used at sentencing, it circumscribed its use by requiring Presentencing Investigation Reports that contained them to inform the sentencing court that the "risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed" along with information about the secrecy of the model, the need for monitoring and updating, and research raising concerns about disproportionate classification of minority offenders as high-risk. *Loomis*, 881 N.W.2d at 764. Whether judges understand the implications of the lack of local validation is unknown.

70. Jennifer L. Skeem & Jennifer Eno Loudon, *Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)*, U.C. IRVINE, at 22–23 (2007), <http://ucicorrections.seweb.uci.edu/files/2013/06/CDCR-Skeem-EnoLouden-COMPASeval-SECONDREVISION-final-Dec-28-07.pdf> [<https://perma.cc/EMT7-7CJY>] (prepared for the California Department of Corrections and Rehabilitation).

71. *Id.*

of over-fitting, it is necessary to cross-validate actuarial formulas with a sample of individuals from the population of interest.<sup>72</sup>

These examples illustrate the importance of identifying the fault lines between populations for a given task. As in the assistance fraud case above, relevant subpopulations may not always be evident up front but rather only be discovered using exploratory machine learning approaches. Further, even if subpopulations are identified, the question remains whether or not they should or may be subject to different models. These decisions regarding whether and how to segment populations for different models is a core question of policy.

## 2. Decisions About Target Variables Embed Policy

In social science, a key element of research design is identifying how to construct an experiment that will test the phenomena of interest. The term “construct validity” is used to ask whether the observations—choices of both instrumentation and data—will actually capture the phenomena of interest. In machine learning systems, the same question also arises. For example, when creating a risk assessment tool, one must determine how to operationalize the risk of recidivism.<sup>73</sup> The decision about how to optimize the target variable has sweeping policy implications.

In the recidivism risk context, an agency might like to measure actual recidivism but lacks the data to do so: we simply do not have ground truth to know whether any individual will commit a crime after release. Because recidivism itself cannot be measured, re-arrest is used as an outcome variable in the model.

As many have pointed out, however, arrests are both poor and biased proxies for actual recidivism rates.<sup>74</sup> First, incomplete observations mean we do not know all outcomes. Second, re-arrest rates reflect policing patterns, which historically police communities of color at higher rates than white communities. Thus optimizing for arrests in place of recidivism creates systems that overrepresent populations in ways that play past discrimination forward *by design*.<sup>75</sup> A study conducted for the CDCR rejected a recidivism risk

---

72. *Id.*

73. *Id.* at 5 (discussing construct validity: “it must measure the criminogenic needs it purports to measure; for example, it should relate coherently to other measures of needs and capture change in risk state over time”).

74. Eaglin, *supra* note 22, at 94–95 (discussing inherent bias in selecting re-arrest as the measure of recidivism, given disproportionate police scrutiny of minority communities); Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 CRIM. JUST. & BEHAV. 185, 196–97 (2019) (describing research that finds “people of color, especially Black people, are more likely to be arrested than Whites for the exact same behavior,” which makes arrest a racially tainted proxy for recidivism).

75. Eckhouse, *supra* note 74, at 197. Eckhouse stated:

tool on these grounds, stating that there is no evidence “that it assesses the criminogenic needs it purports to assess”<sup>76</sup> and concluding that the tool “reliably assess(es) something that looks like criminogenic needs and recidivism risk” but “there is little evidence that this is what . . . [it] actually assesses.”<sup>77</sup>

Given that many of the phenomena we use models to measure—such as risk—cannot be truly observed, the proxies we select to measure them reflect policy choices about how best to measure and predict the phenomena.

### 3. The Choice of Model Embeds Policy

Designers of machine learning systems must choose a modeling framework.<sup>78</sup> Consequently, the choice of framework embeds a theory of how or why a phenomenon is occurring in the world.<sup>79</sup> For example, PredPol,<sup>80</sup> the predictive policing system adopted off the shelf, and without public process, by dozens of police departments across the United States (including Los Angeles and the University of California, Berkeley),<sup>81</sup> uses a “seismological” model to describe how crimes propagate throughout a region.<sup>82</sup> The motivation for using a seismological approach is that an original crime has ripple effects that lead to other crimes, much as an earthquake can lead to

---

When both the data used to produce the risk-assessment instrument and the data used to evaluate it come from the criminal justice system, quantitative risk assessments merely launder that bias. . . . [T]he legitimating process of quantitative assessment converts unequal data-generating processes into apparently objective data, without removing the fundamental problems.

*Id.*

76. Skeem & Loudon, *supra* note 70, at 6.

77. *Id.*

78. See Lehr & Ohm, *supra* note 61, at 688–95 (distinguishing between different general classes of models—random forests, neural networks, etc.—highlighting six considerations that can influence model selection, and explaining that models can be chosen from pre-configured options bundled into software and services, or be modifications made to an existing model).

79. This Section benefited from the analysis of Nitin Kohli. Memo and conversation are on file with author.

80. PREDPOL, <https://www.predpol.com/> [<https://perma.cc/Q9N4-9CSP>] (last visited Sept. 30, 2019).

81. Caroline Haskins, *Dozens of Cities Have Secretly Experimented With Predictive Policing Software*, VICE (Feb. 6, 2019, 7:00 AM), [https://www.vice.com/en\\_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software](https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software) [<https://perma.cc/N384-4HJW>].

82. The analysis of the PredPol algorithm that follows is based on information provided on their website as well as academic papers (also referenced by PredPol) to provide insight into the domain specific methods. *Predictive Policing: Guidance on Where and When to Patrol*, PREDPOL, <https://www.predpol.com/how-predictive-policing-works/> [<https://perma.cc/5LCY-6N25>] (last visited Sept. 30, 2019).

aftershocks that propagate through space and time. This model, known as Epidemic Type Aftershock Sequence (ETAS), decomposes crime into two components: background events and aftershock events. If the model is valid, then this decomposition will allow the system to predict when and where similar crimes are likely to occur, given information of recent criminal activity.

The ETAS model transplants assumptions valid in scientific geological models to the criminal context. In doing so, it implicitly constructs a particular theory of how crimes ripple through time and space. Yet the approach threatens to trade off modeling simplicity with real-world boundary conditions: the behavior of earthquakes cannot accurately predict key elements in modeling background and aftershock effects of crime, such as whether all violations have the same value—by producing the same sort of aftershocks—or whether the kind or location of crime factors into the modeling. In the seismological context, moreover, it is appropriate to assume that aftershock effects are less common after longer periods of time, less common at locations far away from the source, and uni-directional—radiating outward. But aftershock crimes need not obey the same constraints.

The assumptions embedded in the model become more problematic in conjunction with the known limitations of data about crime discussed above. Historical crime data is a lower bound on the actual representation of crime, raising important issues of selection bias and generalizability. More importantly, any bias in crime reporting patterns—for example, social stigmas related to reporting some crimes and risks of reporting that vary by context—further reduces PredPol’s knowledge about “aftershock events” of the observed crime, let alone all crime in general. Simply put, the social laws that govern the spread and reporting of crime do not obey the physical laws that govern the spread and visibility of earthquakes.

Choosing a model is a significant decision of policy. Doing it well requires an understanding of how the model relates to relevant domain- or field-specific theories of the phenomena of interest, a careful examination of any properties a model inherits from its domain of initial development, and an examination of the way model choices might introduce bias.

#### *4. Choosing Data on Which to Train a Model Embeds Policy*

Machine learning systems generate algorithms based on sample data, known as “training data,” in order to make predictions or decisions without being explicitly programmed to perform those tasks. They are then shaped through feedback gleaned by the system’s observations of additional data.

Thus, the choice of the data on which to train a model will have profound implications for the model's outputs.<sup>83</sup>

An analysis of the COMPAS system provides an excellent example of the rigorous way in which data must be interrogated to determine if they are likely to produce an accurate model for a given population. The authors of the report explain:

[T]he COMPAS data are not representative of the California Department of Corrections and Rehabilitation inmates because among other things, eight groups of inmates with potentially greater needs, including those with mental health classifications and those targeted for the substance abuse programs, were excluded from the sample . . . [moreover,] it is not clear how these offenders compare to offenders in other states. Moreover, the data are largely based on offenders' self-report, and there is no protection against reporting bias, including exaggeration or minimization of needs.<sup>84</sup>

The selection and use of protected attributes like race and gender within a dataset used to train machine learning models is a particularly significant policy decision. While it may be that some uses of gender could advance justice, that does not mean that such use would survive an equal protection clause challenge. As the U.S. Supreme Court has said, the "Equal Protection Clause [is] not to be rendered inapplicable by statistically measured but loose-fitting generalities."<sup>85</sup> Even where the use of gender may serve a just purpose—as the State of Wisconsin claimed in *Loomis*—the Court has upheld disparate treatment based on gender only where it seeks to level the playing field. As the Court established in *Mississippi University for Women v. Hogan*: "In limited circumstances, a gender-based classification favoring one sex can be justified if it intentionally and directly assists members of the sex that is disproportionately burdened."<sup>86</sup> The proposition that including gender as a factor might enhance the predictive accuracy of a model or that using it to normalize results improves predictive accuracy for both men and women

---

83. Many scholars have offered excellent explanations and examples of data-related bias; these four are particularly rich and powerful: Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677–94 (2016); Kate Crawford, *The Hidden Biases in Big Data*, HARV. BUS. REV. (Apr. 1, 2013), <https://hbr.org/2013/04/the-hidden-biases-in-big-data> [<https://perma.cc/WC3X-CBJD>]; and ROB KITCHIN, *Conceptualising Data*, in THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES & THEIR CONSEQUENCES 1–26 (2014).

84. Skeem & Loudon, *supra* note 70, at 24 (citing JEFFREY LIN, PAROLEE NEEDS IN CALIFORNIA: A DESCRIPTIVE ANALYSIS OF 2006 COMPAS DATA, THE CENTER FOR EVIDENCE-BASED CORRECTIONS (2007)).

85. *Craig v. Boren*, 429 U.S. 190, 209 (1976).

86. *Miss. Univ. for Women v. Hogan*, 458 U.S. 718, 728 (1982).

however, is, on its own, a legally insufficient reason for choosing whether or how to use it. This use cuts to the heart of equal protection law; as the Court notes, “proving broad sociological propositions by statistics is a dubious business, and one that inevitably is in tension with the normative philosophy that underlies the Equal Protection Clause.”<sup>87</sup>

Thus, the question of how protected attributes are used in a statistical model, whether with pen and paper or by software algorithm, is a question of great political and legal importance. The COMPAS manual describes a system that uses gender in several ways. It presents sixteen “common categories or prototypical offending and behavior patterns that often reappear in criminal justice populations” for use in treatment planning which are segmented along gender lines.<sup>88</sup> It provides users with the ability to consider scale scores in reference to the scale distributions of eight normative subgroups that again are broken down along gender lines.<sup>89</sup> These choices about how to use data cut to the heart of commitments to equal protection and are surely substantive policy.

##### *5. Decisions About Human-System Interactions Embed Policy*

Last but not least, the interfaces and policies that structure interactions between agency staff and machine learning systems shape policy outcomes. The ways humans and machines are bound together through interfaces, processes, and policies in “automation policy knots”<sup>90</sup> shape their impact. A few examples illustrate their importance.

First, as noted above, local police departments, and now the FBI, are relying on Amazon Rekognition to assist in identifying suspects.<sup>91</sup> Like many other software products, Rekognition has preconfigured defaults. The default “confidence threshold” for the face-matching is 80%. Leaving the default confidence threshold as such, ACLU researchers found that it incorrectly matched twenty-eight members of Congress with arrestees in the database—a 5% error rate among legislators—with a disproportionate number of false positives for African-American and Latino members. While Amazon’s system

---

87. *Craig*, 429 U.S. at 204.

88. NORTHPOINTE, *supra* note 7, at 48–49.

89. *Id.* at 11 (the current normative subgroups for comparison are “(1) male prison/parole, (2) male jail, (3) male probation, (4) male composite, (5) female prison/parole, (6) female jail, (7) female probation, and (8) female composite”).

90. Meg Leta Jones, *The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles*, 18 VAND. J. ENT. & TECH. L. 77 (2015) (revealing how legal approaches that ignore the complex relations between humans and machines fail to protect the values they were drafted to protect); Steven J. Jackson et al., *The Policy Knot: Re-integrating Policy, Practice and Design in CSCW Studies of Social Computing*, PROC. CSCW ’14 1 (Feb. 15, 2014) (coining the term policy knot: “practices and design impact and are impacted by structures and processes in the realm of policy”).

91. Wingate, *supra* note 33.

documentation contains some language recommending law enforcement to use a confidence threshold of 99%,<sup>92</sup> the out-of-the-box default does not appear to have any particular relation or relevance to the domains in which it is being used. More importantly, the choice of threshold implicitly makes a policy decision about the tradeoffs between false positive and false negatives.<sup>93</sup> Such choices are paradigmatic questions of policy—they do not have answers in data but reflect instead value judgments that should reflect the goals set for the agency, refined through policy processes, not the designer's preference.<sup>94</sup>

A second example comes again from the recidivism risk domain and involves the policies governing decision maker behavior in the face of system determinations. Specifically, jurisdictions can establish policies that make departures from the recommendation of a recidivism risk system procedurally and potentially politically costly, placing a thumb on the scale of outcomes. For example, a judge may be required to justify a decision to deviate from a risk score on the record.<sup>95</sup> Such provisions not only raise the time required to exercise discretion but also may make judges vulnerable when they release an individual who later commits a crime, against the system's recommendation, thus contributing to a culture of deference to machine reasoning even when the law might prefer human judgment. Or as in the *Loomis* case, judges may be burdened with detailed information about the system's construction—its lack of local validation, for example—but lack the expertise to understand their practical meaning. This information mandate places a human more in the loop

---

92. *Amazon Rekognition Developers Guide*, AMAZON 143 (2019), <https://docs.aws.amazon.com/rekognition/latest/dg/rekognition-dg.pdf> [<https://perma.cc/P8TH-RL5W>]. The guide states:

All machine learning systems are probabilistic. You should use your judgment in setting the right similarity threshold, depending on your use case. For example, if you're looking to build a photos app to identify similar-looking family members, you might choose a lower threshold (such as 80%). On the other hand, for many law enforcement use cases, we recommend using a high threshold value of 99% or above to reduce accidental misidentification.

*Id.*

93. See Eckhouse, *supra* note 74, at 194–95 (describing the lack of guidance on how to translate a risk score produced by a recidivism risk tool into categories to support judges or other decision-makers, and the way that categorizations can inflate perceptions of risk if they deviate from mental models about how a five-part scale, for example, would relate to overall risk); see also Eaglin, *supra* note 22, at 85–87 (discussing how translation between numerical output of algorithmic system and risk is subjective policy choice).

94. See, e.g., Felicitas Kraemer et al., *Is There an Ethics of Algorithms?*, 13 ETHICS & INFO. TECH. 251 (2011) (describing the ways that value judgments regarding false positives and false negatives govern the choice between different rational design decisions, and the setting of thresholds); Eaglin, *supra* note 22, at 88.

95. See, e.g., N.J. STAT. ANN. § 2A:162–23 (West 2017).

but it is unclear how this particular actor—a legal professional—at this particular juncture can protect the values of due process and equality put at risk by ill-chosen design.<sup>96</sup>

Therefore consequential decisions about people's lives are delegated in a variety of ways to machine learning systems that governments buy, or more typically contract to use—generally in an off-the-shelf manner. In effect, governments are outsourcing decisions of policy—sometimes life-changing ones—to algorithmic systems with little understanding of the assumptions those systems embed, the logics on which they rely, the data on which they were trained, or any of the other information necessary to understand whether or not they adequately and appropriately perform the reasoning tasks being handed to them.

Moreover, in the most shocking instances, institutions within the justice system have procured and used recidivism risk systems without understanding the embedded definitions of fairness, the confidence thresholds, the limitations presented by choices of training data sets, or the systems' use of protected characteristics. Individuals whose lives are being altered by these black box decision-making systems are not the only ones who do not understand how these systems work. In an unprecedented dereliction of oversight, government agencies at all levels are, perhaps unwittingly, turning over key policy questions to privately developed algorithmic systems.<sup>97</sup>

At various points over the past fifty years, policymakers have recognized the substantive nature of decisions that can be masked by procurement, and have suggested alternative models to ensure administrative processes of the type usually accorded traditional types of policy decisions. In 1969, for example, the Administrative Council of the United States recommended that, consistent with the goal of “assur[ing] that Federal agencies will have the benefit of the information and opinion that can be supplied by persons whom regulations will affect,” the exemption from notice-and-comment rulemaking procedures for matters relating to “public property, loans, grants, benefits, or

---

96. See Jones, *supra* note 90, at 90–100 (describing how laws designed to achieve a goal by removing or inserting a human in the loop without thoroughly considering how the knot of policies, processes, and design work in practice—taking a socio-technical systems view—often fail, and advocating a set of Fair Automation Practice Principles to guide the construction of human-machine collaborations).

97. Mulligan & Bamberger, *supra* note 12, at 741 (“Public power is too often exercised in private, by private parties, or without nonpartisan or nonpolitical sources of expertise. The substance and political nature of choices fixed by technology is thus obscured, which enfeebles citizen awareness and involvement, diminishes ex post accountability, and yields unintended outcomes.”).

contracts” be discontinued,<sup>98</sup> and several federal agencies, at different points in time, required such procedures for procurement decisions.<sup>99</sup> More recently the IRS Taxpayer Advocate advocated (unsuccessfully) for subjection of IRS “policy guidance embedded in [automated] systems”<sup>100</sup>—which are neither reviewed internally nor published—to the “same stringent vetting and review process as written instructions or policies.”<sup>101</sup> Those written policies undergo a formal clearance, subject to public scrutiny, by staff of the Taxpayer Advocate Service, who review proposed guidance for conflicts with existing policies and procedures, for technical accuracy, and to identify policies or procedures that may harm taxpayers, and offer solutions and alternatives to alleviate these burdens.

Each of these experiments suggests a shifting mindset for structuring the adoption of algorithmic systems—from procurement to administrative process. The next Part takes up these suggestions; those aspects of machine learning systems that touch on substantive aspects of the relationship between the citizen and the state must be viewed as policy and should be brought within the framework that maintains and constrains the exercise of agency power.

### III. BRINGING MACHINE-LEARNING SYSTEM DESIGN WITHIN ADMINISTRATIVE LAW

#### A. ADMINISTRATIVE PROCESS FOR MACHINE LEARNING DESIGN

Identifying the ways that the design of machine learning systems can embed value decisions reveals the ways that the adoption of machine learning systems through procurement can render policymaking invisible. Design choices set policy without input from agency employees, stakeholders, or other experts. The models, assumptions, metrics, and, at times, even the data that

---

98. Admin. Conference of the U.S., Recommendation number: 69-8, Elimination of Certain Exemptions from the APA Rulemaking Requirements (Oct. 22, 1969); 38 Fed. Reg. 19784 (July 23, 1973).

99. See, e.g., 36 Fed. Reg. 13804 (July 24, 1971); Revocation of Statement of Policy on Public Participation in Rulemaking, 78 Fed. Reg. 64194 (Oct. 28, 2013) <https://www.federalregister.gov/documents/2013/10/28/2013-25321/revocation-of-statement-of-policy-on-public-participation-in-rulemaking> [<https://perma.cc/7ZTG-NWHM>] (showing the Department of Agriculture’s history of forty-two years of notices and comments); 29 C.F.R. § 2.7 (1979) (promulgated at 36 Fed. Reg. 12976 (July 10, 1971)). The C.F.R. section states:

It is the policy of the Secretary of Labor that in applying the rule making provisions of the APA the exemption therein for rules relating to public property, loans, grants, benefits or contracts shall not be relied upon as a reason for not complying with the notice and public participation requirements thereof.

*Id.*

100. TAXPAYER ADVOCATE SERV., *supra* note 60.

101. *Id.* at 78.

drive such systems, are largely opaque and unknown to government officials who acquire them and the public they govern.

When such systems embed policies, the current method of adoption lacks all hallmarks of legitimate governance. Administrative actors are excused from reasoning, analysis, and the requirement that they justify policy choices. They bring no expertise to bear. They elicit no public participation or input. Their decisions evade judicial review and political oversight. Scholarship has largely failed to address this phenomenon of lawless governance. To be sure, a robust literature has focused on the challenge of system opacity, proposing algorithmic “transparency” as a means to address the ways opacity can obscure bias, error, and outcomes that diverge from public goals.<sup>102</sup> Proposals for transparency have focused on open sourcing a given system’s software code and releasing it to the public for inspection; mandating disclosure of system methodology;<sup>103</sup> disclosing the sources of any data used;<sup>104</sup> requiring audit trails that record the facts and rules supporting administrative decisions when they are based on automated systems; mandating that hearing officials explain in detail their reliance on an automated system’s decision;<sup>105</sup> and notifying those affected when algorithmic systems are used.<sup>106</sup>

Such transparency mechanisms, in turn, are intended to facilitate accountability.<sup>107</sup> Openness about the algorithms that drive technological

---

102. See generally Charles Vincent & Jean Camp, *Looking to the Internet for Models of Governance*, 6 ETHICS & INFO. TECH. 161, 161 (2004) (explaining that automated processes remove transparency); Paul Schwartz, *Data Processing and Government Administration: The Failure of the American Legal Response to the Computer*, 43 HASTINGS L.J. 1321, 1343–74 (1992) (setting forth an influential paradigm for addressing data-driven governance, which includes making data processing systems transparent; granting limited procedural and substantive rights to the data subject; and creating independent governmental monitoring of data processing systems).

103. PASQUALE, *supra* note 18, at 14–15; Giovanni Buttarelli, *Towards A New Digital Ethics: Data, Dignity, And Technology*, EUR. DATA PROTECTION SUPERVISOR 2 (Sept. 11, 2015); Rob Kitchin, *Thinking Critically About and Researching Algorithms*, 20 INFO. COMM. & SOC’Y 14 (2017); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. U. L. REV. 1, 21 (2014); Nicholas Thompson et al., *Emmanuel Macron Talks to WIRED About France’s AI Strategy*, WIRED (Mar. 31, 2018, 06:00 AM), <https://www.wired.com/story/emmanuel-macron-talks-to-wired-about-frances-ai-strategy/> [<https://perma.cc/X7HH-4K6K>].

104. PASQUALE, *supra* note 18, at 14.

105. Citron, *supra* note 19, at 1310–12.

106. See Citron & Pasquale, *supra* note 103, at 21; see also Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 125–28 (2014) (advocating a right to “procedural data due process” to address the harms of predictive systems).

107. Kroll et al., *supra* note 20, at 657 (describing transparency as “[a] native solution to the problem of verifying procedural regularity” and describing its utility and limits); Fink, *supra* note 20, at 1453–56 (explaining limits of transparency due to current state of law and practice with respect to whether algorithms would be considered “records” under the Freedom of

systems government agencies use permits public analysis and critique,<sup>108</sup> and an assessment of the fairness of their use. It allows software audits<sup>109</sup> that identify correct, and incorrect, inputs and outputs, back-testing of those input and outputs to assure the system is executing its intended goals,<sup>110</sup> and testing of software on specific scenarios with pre-determined outcomes. It can allow individuals to contest, inspect, and adjudicate problems with data or decisions made by a system, facilitating challenges to government determinations based on algorithmic systems. Such measures facilitate mechanisms to “vindicate the norms of due process” and administrative decision making even when decisions are automated.<sup>111</sup> This allows individuals to plead extenuating circumstances that software cannot anticipate<sup>112</sup> and accords the subjects of automated decisions the right to inspect, correct, and dispute inaccurate data.<sup>113</sup>

Yet, while critics have debated the limits of these approaches,<sup>114</sup> the debate has focused largely on the use and effectiveness of transparency,

---

Information Act (FOIA) and agency bases for withholding algorithms and source code under FOIA requests); Pasquale, *supra* note 20, at 235–36.

108. Citron, *supra* note 19, at 1311–12.

109. See Citron & Pasquale, *supra* note 103, at 20–22 (advocating for transparency requirements for data and calculations and placing scoring systems used in the context of employment, insurance, and health care under licensing and audit requirements); see also Crawford & Schultz, *supra* note 106, at 122–23.

110. PASQUALE, *supra* note 18, at 14–15; Diakopoulos, *supra* note 44, at 399–402; Citron & Pasquale, *supra* note 103102, at 21–22.

111. Citron, *supra* note 19, at 1301.

112. *Id.* at 1304.

113. PASQUALE, *supra* note 18, at 145.

114. Kroll et al., *supra* note 20, at 657–58 (explaining that while “full or partial transparency can be a helpful tool for governance in many cases . . . transparency alone is not sufficient to provide accountability in all cases”); see generally Katherine Noyes, *The FTC Is Worried About Algorithmic Transparency, and You Should Be Too*, PC WORLD (Apr. 9, 2015, 08:36 AM), <https://www.pcworld.com/article/2908372/the-ftc-is-worried-about-algorithmic-transparency-and-you-should-be-too.html> [<https://perma.cc/7KHT-GHZ7>]; Christian Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, UNIV. MICH. (May 22, 2014), <http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf> [<https://perma.cc/TJ3Y-2UZK>] (presenting at “Data and Discrimination: Converting Critical Concerns into Productive Inquiry,” a preconference at the 64th Annual Meeting of the International Communication Association). Critiques include the fact that open sourcing a given machine learning system’s neural network does not necessarily mean an outside third party will verify how the system determined a given output. See Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability*, NEW MEDIA & SOC’Y 973, 983–84 (2016); Jakko Kemper & Daan Koklman, *Transparent to Whom? No Algorithmic Accountability Without a Critical Audience*, INFO. COMM. & SOC’Y (2018); Braunes & Goodman, *supra* note 6, at 137–38 (pointing out the difficulty of understanding complex AI systems and the shortcomings of

whistleblowers, ex post challenges, and oversight. The ex post focus positions accountability after critical design decisions have been made. And while new scholarship has begun to focus on the process of machine learning system design,<sup>115</sup> this literature has not explored the full potential of administrative law to remedy the abdication of government agencies' involvement in design questions, even when they implicate issues that we usually regard as involving traditional substantive policy questions.

Administrative law maps another direction. It suggests that, when the design of machine learning systems embeds policy, policymakers should be required to engage in reasoned decision making. To be meaningful, given the character of the decisions involved in machine learning design, that deliberation must address an understanding, informed by both technical and domain expertise, of the methodologies adopted and the value choices behind them, and provide justifications for those choices' resolution. Administrative law, moreover, provides guidance about what types of concerns should trigger such requirements, and how, given the characteristics of machine learning, those concerns translate to the particular context of system design.

#### B. A FRAMEWORK FOR REASONED DECISION MAKING ABOUT MACHINE LEARNING DESIGN

The administrative state's legitimacy is premised on the foundational principle that decisions of substance must not be arbitrary or capricious.<sup>116</sup> Rather, those decisions must be the product of a contemporaneous process of reasoned decision making.<sup>117</sup> Requiring such process vindicates core public law values: it ensures, on the one hand, that technical expertise has been brought to bear on a decision; and on the other, that the decisional visibility necessary to permit public accountability exists.<sup>118</sup> Together, a transparent reasoning process prohibits an agency from "simply asserting its preference."<sup>119</sup>

---

knowing inputs and outputs of a given system as the basis for adequate oversight); Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181, 194–96 (2016). For the impediment posed to transparency by trade secret law, see Brauneis & Goodman, *supra* note 6, at 153–57; David S. Levine, *Secrecy and Unaccountability: Trade Secrets in Our Public Infrastructure*, 59 FLA. L. REV. 135, 180 (2007).

115. See *supra* note 22; see also Katyal, *supra* note 20, at 54.

116. Courts may "hold unlawful and set aside [an] agency action" they deem to be "arbitrary [or] capricious." 5 U.S.C. § 706(2)(A).

117. SEC v. Chenery Corp. (Chenery II), 332 U.S. 194, 196 (1947) (holding that courts may uphold an agency's action only for reasons on which the agency relied when it acted); *see generally* Kevin M. Stack, *The Constitutional Foundations of Chenery*, 116 YALE L.J. 952 (2007) (grounding the *Chenery* norm in the Constitution).

118. Cass R. Sunstein, *From Technocrat to Democrat*, 128 HARV. L. REV. 488 (2014) (discussing the technocratic and democratic directions in administrative law).

119. *Id.* at 496.

Specifically, an agency must produce a record that enables courts “to see what major issues of policy were ventilated,” and “why the agency reacted to them as it did.”<sup>120</sup> Thus the agency must have engaged in reasoned analysis about relevant factors consistent with the record before it, and they may not have considered irrelevant factors or decided without sufficient evidence. An agency falls short where there is no record of “examin[ing] the relevant data” or “articulat[ing] a satisfactory explanation for its action including a ‘rational connection between the facts found and the choice made.’ ”<sup>121</sup>

By the terms of this standard, the complete abdication of any agency role in considering the important policy choices inherent in a machine learning system’s design would be an abject failure. This Section explores the alternative, using the arbitrary and capricious paradigm to identify the types of machine learning systems, and system elements, whose design should be guided by reasoned and transparent decision making, and what such decision making would require in the machine learning context to survive legal challenge.

### *1. Determining What System Choices Should Require Reasoned Decision Making*

Government agencies increasingly rely on artificial intelligence across their operations. Many functions—from monitoring IT system security to managing government supply lines and procurement—involve largely management support, and therefore may not implicate the types of policy decisions that should trigger the type of decisional record discussed above. This raises a threshold challenge in distinguishing systems that are inward-facing from those that create public-facing policy of the type that agencies should deliberate about and ventilate in a public manner.

Administrative law has dealt with comparable distinctions in a range of contexts and offers some insights into where and how we might draw lines about when a machine learning system is engaged in policymaking of concern to us, and when it is not. Specifically, jurisprudence has identified important indicia of contexts in which administrative choices trigger concerns necessitating a reasoned and transparent decision making process, and the creation of a record sufficient for judicial review: whether the agency action in question limits future agency discretion in deciding issues of legal consequence, and whether the action reflects a normative choice about implementation. Each of these inquiries offer useful insight for the question

---

120. Auto. Parts & Accessories Ass’n v. Boyd, 407 F.2d 330, 338 (D.C. Cir. 1968).

121. Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 43 (1983) (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

of which machine learning systems' design, and which system elements' design, should be treated as making policy.

a) Design Choices that Limit Future Agency Discretion

In a variety of contexts, courts have identified the constraining effect of an administrative decision on the future substantive discretion of agencies or their staff as a baseline determinant of whether agency decisions will be subject to judicial review, and therefore to analysis under the arbitrary and capricious standard. When current decisions hem in choices about the law's application going forward, they reflect binding policy choices and thus may be reached openly, explicitly, and through reasoned analysis.

Even in contexts in which executive discretion is broad—as it is in internal agency management—such factors argue for requiring reasoned decision making. Thus, while agencies have largely unreviewable discretion regarding enforcement decisions,<sup>122</sup> judicial oversight is appropriate when an agency adopts a “general enforcement policy” that “delineat[es] the boundary between enforcement and non-enforcement.”<sup>123</sup> Such actions limit agency discretion going forward, with implications for “a broad class of parties.”<sup>124</sup> In such contexts, in contrast to individual decisions to forgo enforcement, an agency is expected to present a clearer and more easily reviewable statement of its reasons for acting.<sup>125</sup>

Related concerns govern the determination of whether an agency action is “final,” which is a second Administrative Procedure Act (APA) prerequisite for judicial review.<sup>126</sup> To satisfy this requirement, an agency action must not simply mark the “consummation” of an agency’s “decision-making process”—a standard satisfied by many nonbinding or advisory decisions, even when they are made informally.<sup>127</sup> The decision must also “be one by which rights or

---

122. See *Heckler v. Chaney*, 470 U.S. 821, 832 (1985). The court opinion stated:

[W]e recognize that an agency's refusal to institute proceedings shares to some extent the characteristics of the decision of a prosecutor in the Executive Branch not to indict—a decision which has long been regarded as the special province of the Executive Branch, inasmuch as it is the Executive who is charged by the Constitution to “take Care that the Laws be faithfully executed.”

*Id.*; APA excludes from review “agency action . . . committed to agency discretion by law.” 5 U.S.C. § 701; *see also* *Citizens to Preserve Overton Park, Inc. v. Volpe*, 401 U.S. 402, 410 (1971) (holding that the “committed to agency discretion” exception to judicial review is “very narrow” and “is applicable in those rare instances where ‘statutes are drawn in such broad terms that in a given case there is no law to apply’”).

123. *Crowley Caribbean Trans. v. Pena*, 37 F.3d 671, 676–77 (D.C. Cir. 1994).

124. *Id.*

125. *Id.*

126. The APA extends judicial review only to “final agency action.” 5 U.S.C. § 704.

127. *Bennett v. Spear*, 520 U.S. 154, 177–78 (1997).

obligations have been determined, or from which legal consequences will flow.”<sup>128</sup> The Supreme Court has recently counseled a “pragmatic” approach to the interpretation of this standard, focusing on the prospective limits it places on agency discretion as a key component of the “legal consequences” test.<sup>129</sup> Lower courts have already taken such a pragmatic approach—looking at whether, as a practical matter, a purportedly non-binding agency decision effectively guides future agency decisions and constrains agency discretion, such as if “an agency act[s] ‘as if a document issued at headquarters is controlling in the field.’ ”<sup>130</sup>

The distinctions drawn with respect to finality track those governing whether agency actions must satisfy the notice and comment procedures prescribed by § 553 of the APA. While reasoned decision making sufficient for system design and adoption decisions to survive arbitrary and capricious review can certainly occur through a range of administrative processes beyond informal rulemaking, this jurisprudence offers an informative framework in which courts have thought carefully about which agency actions should trigger more robust process, reflecting reasoned deliberation, participation, expertise, and judicial review.

In this context, courts have developed extensive doctrine regarding what types of agency actions are “non-legislative” and therefore exempt from such process requirements, as compared to those that are “substantive” and therefore must satisfy them. Such exempt actions (involving, for example, internal agency procedure, agency management, or guidance to regulated parties) do not carry the “force of law” in that they do not make substantive changes to the legal rights and obligations of regulated individuals. As understood by case law, agency guidance statements are those “issued by an agency to advise the public prospectively of the manner in which the agency proposes to exercise a discretionary power.”<sup>131</sup> These statements provide agencies with the opportunity to announce their “tentative intentions for the future” in a non-binding manner. An agency articulation, then, that “genuinely leaves the agency and its decision makers free to exercise discretion” raises few process concerns.<sup>132</sup> The agency may adopt it with little process, and it is not, in and of itself, reviewable by courts.

By contrast, courts are also sensitive to the concern that agencies are circumventing the need for decision-making process when they make

---

128. *Id.*

129. *U.S. Army Corps of Eng’rs v. Hawkes Co.*, 136 S. Ct. 1807, 1814–15 (2016); see William Funk, *Final Agency Action After Hawkes*, 11 N.Y.U.J.L. & LIBERTY 285 (2017).

130. Appalachian Power Co. v. EPA, 208 F.3d 1015, 1021 (D.C. Cir. 2000).

131. Am. Bus. Ass’n v. United States, 627 F.2d 525, 529 (D.C. Cir. 1980) (internal citation omitted).

132. *Id.*

substantive policy in a manner purported to govern only internal agency procedure or provide only informal guidance. As a result, courts sometimes find that notice-and-comment is necessary, even when the agency statement in question does not seem in and of itself to have any binding legal effect on regulated entities. This seems especially so when the relevant statutes and legislative rules give the agency wide discretion, but the challenged agency statement indicates that agency personnel will in reality exercise that discretion only in narrowly defined circumstances.<sup>133</sup> In those situations, courts have found that the agency action is “practically” (although not formally) binding. Because of the severe constraints that the agency’s “informal” action imposed on agency discretion, the agency should have engaged in the full notice-and-comment rulemaking procedure.

Tracking these standards, existing jurisprudence regarding the setting of formulae and numerical cutoffs, and the choices regarding underlying methodology, provides useful guidance for identifying aspects of machine-learning systems that set discretion-constraining policy.

*Pickus v. United States Board of Parole*,<sup>134</sup> a case arising in the challenge to an agency’s decision to adopt a formula informally (without a notice and comment process), describes well the ways in which such adoption can set future policy by limiting agency discretion going forward. In *Pickus*, the D.C. Circuit considered a challenge to two rounds of Parole Board “guidelines” that set formulae by which parole would be determined. The court rejected the Board’s contention that, under the APA, the issuance of such guidelines lacked legal force because they were merely “general statements of policy, interpretative rules,” or “rules relating to agency organization, practice or procedure.”<sup>135</sup>

In so doing, the court focused on the practical implications on agency decision-making discretion, and the subsequent legal consequences. As the court described, the first set of guidelines “consist of nine general categories of factors, broken down into a total of thirty-two sub-categories, often fairly specific.” Therefore,

---

133. Gen. Elec. v. EPA, 360 F.3d 188 (D.C. Cir. 2004); Cmty. Nutrition Inst. v. Young, 818 F.2d 943 (D.C. Cir. 1987).

134. *Pickus v. United States Board of Parole*, 507 F.2d 1107 (D.C. Cir. 1974). In a later case, *Prows v. United States Dep’t of Justice*, 704 F. Supp. 272 (D.C. Cir. 1988), a Program Statement from the Federal Bureau of Prisons declaring that inmates had to deposit at least 50% of their payment from prison jobs to “legitimate financial obligations” was struck down. Analogizing the rule to the guidelines in *Pickus*, the court found the Statement “has been interpreted by defendants in a ‘formula like’ manner,” without any discretion and therefore wasn’t an interpretative rule nor a policy statement and should have proceeded through notice and comment. *Prows*, 704 F. Supp. at 277.

135. *Pickus*, 507 F.2d at 1112 (D.C. Cir. 1974) (citing 5 U.S.C. § 553(a)(2) and providing exemptions).

[a]lthough they provide no formula for parole determination, they cannot help but focus the decisionmaker's attention on the Board-approved criteria. They thus narrow his field of vision, minimizing the influence of other factors and encouraging decisive reliance upon factors whose significance might have been differently articulated had [more formal decision-making processes] been followed.<sup>136</sup>

Because of this narrowing of decision-making focus, the court held, the guidelines “were of a kind calculated to have a substantial effect on ultimate parole decisions.”

The second agency action, styled an “announcement,” consisted of a “complex, detailed table which purport[ed] to state the range of months which the Board [would] require an inmate to serve depending upon the severity of his offense (six classifications) and his ‘salient factor score’ (four classifications).”<sup>137</sup> The score, the court continued,

is computed using only those criteria, and the quantitative input of each is specified as well. Computation of the score is a purely mechanical operation. Third, the chart sets a narrow range of months of imprisonment that will be required for a given category of offense and a given salient factor score. This is not to suggest that these determinants are either unfair or undesirable, but merely that they have significant consequences.<sup>138</sup>

Thus, the court concluded, both policies defining parole selection criteria “are substantive agency action,” and “the interested public should have an opportunity to participate, and the agency should be fully informed, before rules having such substantial impact are promulgated.”<sup>139</sup>

Moreover, in *Community Nutrition Institute v. Young*,<sup>140</sup> the D.C. Circuit determined that FDA “action levels”—the allowable levels of unavoidable contaminants in food, and again a precise number—while purportedly without the “force of law,” practically bound third parties and should have gone through the notice-and-comment procedure required for legislative rules. Pursuant to its statutory mandate to limit the amount of “poisonous or deleterious substances” in food,<sup>141</sup> the FDA established “action levels”—which the FDA characterized as guidance statements—that set permissible levels of unavoidable contaminants such as aflatoxins in food. Producers who exceed action levels are subject to enforcement proceedings. The FDA claimed

---

136. *Id.* at 1111–13.

137. *Id.* at 1110–11.

138. *Id.* at 1113.

139. *Id.*

140. Cmty. Nutrition Inst. v. Young, 818 F.2d 943 (D.C. Cir. 1987).

141. 21 U.S.C. § 346.

that action levels were “nonbinding statements of agency enforcement policy,” but the court found that setting precise numerical limits cabined the FDA’s enforcement discretion, effectively binding the FDA and therefore affecting the rights of regulated parties.<sup>142</sup>

b) Normative Choices Between “Methods of Implementation”

Judge Richard Posner, writing for the Seventh Circuit in *Hoctor v. U.S. Department of Agriculture*,<sup>143</sup> has articulated the way that numerically-based line-drawing can often reflect a particularly unconstrained form of normative policymaking—which, when it does, enhances the need for more robust process. *Hoctor* involved a challenge to an informal USDA internal memorandum fixing a specific requirement for the height of perimeter fences used to contain “dangerous animals.” While the background regulation in force for a number of years had required fencing “appropriate” for the animals involved, the memorandum sought uniformly to require eight-foot fences. The court, however rejected the agency’s attempt to arrive at a numerical standard with little decisionmaking process, which, in the court’s mind, undermined the decision’s democratic legitimacy.<sup>144</sup>

Generally, Judge Posner emphasizes the policy-making nature of administrative decisions that “translate[ ] a general norm into a number”<sup>145</sup>—a phenomenon, he notes, that arises “especially in scientific and other technical areas, where quantitative criteria are common.”<sup>146</sup> Moreover, he describes, the “flatter” (or more specific) the ultimate line drawn by the agency, “the harder it is to conceive of it as merely spelling out what is in some sense latent in a statute or regulation”<sup>147</sup> and the more it represents a choice among “methods of implementation.”<sup>148</sup> Such choices are legislative in nature, and should be treated as such.

Jurisprudence reviewing agency decision making under the arbitrary and capricious standard reflects these insights about numerical or formula-based agency implementation choices, and provides important foundations for identifying which elements of machine learning systems must satisfy the arbitrary-and-capricious metric in their adoption. Indeed, courts have explicitly

---

142. *Cmty. Nutrition Inst.*, 818 F.2d at 946–48 (“[T]his type of cabining of an agency’s prosecutorial discretion can in fact rise to the level of a substantive, legislative rule.”). That is exactly what has happened here.

143. *Hoctor v. U.S. Dep’t of Agric.*, 82 F.3d 165, 171 (7th Cir. 1996).

144. SUSAN ROSE-ACKERMAN, STEFANIE EGIDY & JAMES FOWKES, *Due Process of Law-making: The United States, South Africa, Germany and the European Union* 91 (2015).

145. *Hoctor*, 82 F.3d at 171.

146. *Id.*

147. *Id.*

148. *Id.* at 170.

held that agencies must engage in reasoned analysis in choosing methods of implementation reflecting many of the very type of decisions inherent in machine learning design described in Part II.

In assessing risk, courts have held, agency decision makers must actively consider the decision whether to err in the direction of false negatives or false positives, and provide reasons for their choice.<sup>149</sup> Similarly, agencies must justify the assumptions behind their use of specific models when determining costs and designing impact statements<sup>150</sup> and provide information to justify the methodology behind models that they use for risk prediction.<sup>151</sup> They must take steps to confirm the validity of their chosen models<sup>152</sup> and, in deciding whether to use a particular scientific methodology, both demonstrate its reliability and transparently discuss its shortcomings. With respect to data, agencies must provide information on its source.<sup>153</sup>

What reasoned deliberation entails is set out across a range of procedural contexts—from cost-benefit analysis to environmental impact assessments—and in a range of substantive policy contexts. The failure to identify, disclose, engage with, and justify the consequent policy choices within models closely correlated to machine learning systems—statistical and economic models, for example—has been held to constitute a “complete lack of explanation for an important step in the agency’s analysis.”<sup>154</sup> And absent efforts to surface and affirmatively explain the assumptions underlying decision-making models, they may remain “fatally unexplained” and unappreciated.<sup>155</sup>

---

149. See Int’l Union, United Mine Workers of Am. v. Fed. Mine Safety & Health Admin., 920 F.2d 960, 962–66 (D.C. Cir. 1990) (remanding to the agency for “more reasoned decision making” on the issue of whether carbon monoxide monitors provide enough protection for workers, after it engaged only in an analysis of false negatives, without discussion of false positives, “ignor[ing] this problem altogether”).

150. Nat. Res. Def. Council, Inc. v. Herrington, 768 F.2d 1355, 1412–19 (D.C. Cir. 1985) (analyzing the Department of Energy’s use of a real annual discount rate of 10% when determining life cycle costs and the net present value of savings from appliance energy efficiency standards).

151. Owner-Operator Indep. Drivers Ass’n, Inc. v. Fed. Motor Carrier Safety Admin., 494 F.3d 188, 199–204 (D.C. Cir. 2007) (holding that an agency must disclose the methodology of the agency’s operator-fatigue model, a crash-risk analysis that was a central component of the justification for the final rule).

152. Ecology Ctr. v. Austin, 430 F.3d 1057 (9th Cir. 2005). The Forest Service used a model to conclude that treating old-growth forest through salvage logging was beneficial to dependent species but did not confirm their hypothesis by any on-the-ground analysis.

153. Nat. Res. Def. Council, 768 F.2d at 1412–19.

154. Owner-Operator Indep. Drivers Ass’n, Inc., 494 F.3d at 204.

155. Natural Res. Def. Council, 768 F.2d at 1414–19 (analyzing the Department of Energy’s use of a real annual discount rate of 10% when determining life cycle costs and the net present value of savings from appliance energy efficiency standards).

### c) Application to Machine Learning Systems

Decisions about the design of a machine learning system—particularly one modeling fairness—constrain agency discretion much like the formulae in *Pickus*, and the action levels in *CNI*. These cases underscore the ways in which precise numerical limits or formulae have anchoring effects that constrain agency action, and the consequent importance of robust process in their adoption. Machine learning systems are rife with similar issues, such as cutoffs determining who is high, medium, or low risk in recidivism risk systems, or the thresholds in the Amazon Rekognition service described in Part II.

There is, moreover, often no unique connection between the cutoffs or thresholds chosen and a statutory mandate or technological requirement, as in *Hocter*. Rather—like the agency actions in the arbitrary-and-capricious cases involving which risk models to adopt, whether to prefer false negatives and positives, what data to use, and which scientific methodology to employ—those decisions reflect normative choices between methods of implementation. In the machine learning context, these might also include the unit of analysis (the algorithm, the algorithmic system, or the overall system of justice);<sup>156</sup> how fairness is measured—whether it is by group-level demographic parity, equal positive predictive values, equal negative predictive values, accuracy equity, individual fairness metrics such as equal thresholds, or a devised similarity metric—and the related question of whether and how to use attributes related to protected classes such as in the *Loomis* case.

When the design of a machine learning system deprives an agency and its staff of future substantive discretion,<sup>157</sup> especially through numerical or methodological choices that reflect normative judgments on implementation rather than ones required directly by statute or technical or scientific knowledge, the design choices embedded in machine learning systems should not be reached in an arbitrary or capricious manner. Thus if a record lacks evidence of agency deliberation or reveals deliberations that demonstrate one of the other indicia of arbitrariness, an agency's reliance on the system should be subject to legal challenge.

---

156. Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, in PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 59, 60–61 (2019) (describing the “framing trap”—the tendency to analyze fairness at the level of inputs and outputs of the model rather than at the level of socio-technical system in which the machine learning system is embedded).

157. Am. Bus. Ass'n v. United States, 627 F.2d 525, 529 (D.C. Cir. 1980) (asking “whether a purported policy statement genuinely leaves the agency and its decision-makers free to exercise discretion”).

## 2. *Designing Agency Decision Making: Reflecting the Technocratic and Democratic Requirements of Administrative Law*

Where the policy decisions embedded in system design supplant administrative discretion, what would it mean, in the language of the arbitrary and capricious jurisprudence, that on the one hand “the agency should be fully informed”<sup>158</sup> and provide a justification for its choice based on a “rational connection” with the “facts found,”<sup>159</sup> and on the other, that decisions should be open to public engagement and political accountability? These elements reflect dual (and sometimes competing) impulses—technocratic and democratic—animating the law of administrative process.<sup>160</sup>

As to the first, to engage in reasoned deliberation, agency staff must address their lack of technical knowledge, enlist additional expertise to “inform” themselves sufficiently, and provide reasons justifying the resolution of four questions specific to machine learning systems. Those questions include: (1) for what a system is optimizing; (2) what determinations are being made about the choice and treatment of data; (3) what assumptions and limitations are implied by the choice of model; and (4) what interfaces and policies structure agency staff’s interactions with machine learning systems—the human-machine loop. Importantly, as to the second question, meaningful processes must address the opacity of value choices made through design by ensuring “political visibility,”<sup>161</sup> to surface the fact that technical choices involve a policy judgment. In this context, transparent decision making involves not simply making algorithms transparent, but making policy visible.<sup>162</sup>

### a) Technocratic Elements in Reasoned Decision Making About Machine Learning Systems

A comparison of machine learning systems with prior automation—generally so-called “expert systems”—helps identify particular aspects of machine learning system design decisions that displace traditional modes of expert administrative decision making. Danielle Citron’s 2008 work on *Technological Due Process*<sup>163</sup> provided a foundational analysis of the ways that administrative automation based on a prior generation of expert systems

---

158. Pickus v. United States Board of Parole, 507 F.2d 1107, 1113 (D.C. Cir. 1974).

159. Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 43 (1983) (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

160. See generally Sunstein, *supra* note 118 (discussing the technocratic and democratic strains in administrative law).

161. Mulligan & Bamberger, *supra* note 12, at 776–80.

162. See, e.g., Eaglin, *supra* note 22, at 88 (noting how recidivism risk tools make it “difficult to ascertain . . . policy decisions”).

163. Citron, *supra* note 19.

transformed the technological decision-making landscape in ways that matter for policymaking norms.<sup>164</sup> It is further instructive in highlighting the ways machine learning has both compounded and redirected the displacement of expert human judgment, a challenge with which agencies must grapple when adopting such systems.

i) Citron's Concerns: Displacement of Expert Agency Judgment

Citron identified a related set of objections to earlier attempts to automate agency processes. First, she described how “[a]utomated systems inherently apply rules because software predetermines an outcome for a set of facts.”<sup>165</sup> This, in turn, displaces the ongoing exercise of human judgment, which is better reflected in standards. She thus concludes that “[d]ecisions best addressed with standards should not be automated.”<sup>166</sup> Citron further drew on the “rules versus standards” debate to emphasize the distinction between automated systems, which implement rules and favor consistency, and human decision-making systems, which favor “situation-specific discretion.”<sup>167</sup>

Second, Citron raised the related question of *who* sets the rules that displace the standards-like exercise of human judgment. Her concern involved the displacement of expert agency decision making by the choices of engineers who design technical systems.<sup>168</sup> In particular, she was concerned that engineers’ interpretations and biases, and their general preference for tractable binary questions, distort decision making.

Finally, Citron expressed concern regarding the lack of record-keeping and transparency about the rules automated systems apply. Absent such a digital trail, the ability to seek redress or accountability is limited. To enable individual due process and support overall accountability, Citron advocates that systems

---

164. For an excellent and accessible discussion of expert systems and what lessons from their development suggest about the discussion for explainability and interpretability in machine learning, see generally David C. Brock, *Learning from Artificial Intelligence's Previous Awakenings: The History of Expert Systems*, 39 AI MAG. 3 (2018).

165. Citron, *supra* note 19, at 1303.

166. *Id.* at 1304.

167. *Id.* at 1303; see Bamberger, *supra* note 18, at 676 (“Computer code [in contrast to human judgment] operates by means of on-off rules, while the analytics it employs seek to quantify the immeasurable with great precision.”) (internal quotation marks omitted).

168. Citron, *supra* note 19, at 1261 (“Code writers also interpret policy when they translate it from human language to computer code. Distortions in policy have been attributed to the fact that programmers lack ‘policy knowledge.’ ”); *id.* at 1262 (“Changes in policy made during its translation into code may also stem from the bias of the programmer. . . . Policy could be distorted by a code writer’s preference for binary questions, which can be easily translated into code.”).

be built to produce “audit trails that record the facts and rules supporting their decisions.”<sup>169</sup>

ii) Updating Concerns: How Machine Learning Displaces Rational Expert Agency Decision Making

While Citron’s conception of what is inherent in automation may have been largely accurate with respect to the automated systems used by government at the time (prior to her 2008 publication date), the rote application of predetermined rules she documents is an inapt description of the machine learning systems coming into government use today. Machine learning systems do not apply predetermined rules to sets of facts, but rather develop probabilistic models that optimize for a particular goal. They are then allowed to learn in the field, generate new rules on the fly, and iteratively update them.

In this way, like earlier expert systems, machine learning systems too displace agency reasoning and expertise, and constrain future agency discretion. However, the displacement takes new forms, stems from additional sources, and requires distinct responses. The risk of displacement no longer stems from the explicit reasoning of engineers translating agency rules into code, but rather arises from the “logic” the model machine learning systems derive from training data reflecting past agency actions. The assumptions and policy choices built into the machine learning model used to generate the predictive model, as well as policy choices in the application of the predictive model, rather than engineer-coded rules, are the key hidden constraints.

a. Element 1: Delegating “Logic-Making” to Machines

Today’s machine learning systems, then, delegate “logic-making” to algorithms. Unlike expert systems that Citron rightly identified as displacing nuanced and fact-specific agency staff decisions with the rote application of predetermined rules as coded by engineers, machine learning systems *construct their own logics* from training data. Machine learning systems skip the process of codifying an agency’s decision-making process, and instead rely on the machine learning model to learn a classifier—its own machine logic—from a set of training data that reflects past agency actions. While the machine logic captured in the classifier could be considered more analogous to the intuition and instinct associated with agency experts,<sup>170</sup> importantly, it in no way reflects the logic of agency decision makers. In fact, it answers without the

---

169. *Id.* at 1305.

170. Of course, human intuition is produced by neurological processes and machines’ through computational processes. While machine learning abounds with terms that evoke the brain, only some machine learning systems attempt to mirror cognitive processes.

causal reasoning associated with logic, or as one scholar notes, “they don’t ‘think’ in any colloquial sense of the word—they only answer.”<sup>171</sup>

While many consequential decisions are made by the engineers, the decision about how to model agency judgments is not explicitly constructed by engineers through rules—abstract or specific—but rather learned by algorithms through analysis of data traces reflecting agency decision making. In theory, one might surmise that because machine learning models are trained on data that represents the past decisions and related outcomes of the agency—or “similar” ones—they might naturally align more closely with the judgment of agency experts and, by design, provide less room for interference or usurpation of judgment by engineers. If that were so, perhaps machine learning systems should raise *less* concern about displacement of human expert judgment than earlier automated systems.

Unfortunately, this surmise breaks down under more careful scrutiny. The training data reflects patterns reflected in professional decisions, but not professionals’ *decision-making processes*. This is an important distinction. It means that a machine learning model’s “logic” may well reflect the actions and outcomes of professional decision making (the outputs) but bear little resemblance to the rationales and justifications behind those decisions.

Significantly, the “reasoning” of complex machine learning systems often bears no resemblance to human logic and is impossible to discern.<sup>172</sup> The divergence in intuition is intrinsic because machines and humans “see” in different ways. For example, machine learning systems can identify complex patterns and scan across massive data sets. Humans, by contrast, can identify things they’ve seen (such as faces) despite a wide range of subtle and relatively extreme perturbations (changes to hair style, plastic surgery, aging, etc.). The different intuitions developed by human and machine systems may therefore produce similar outputs in some instances, but not in others, or similar outputs but for very different reasons.

Thus, the machine has learned its own logic based on the training data: it has not learned to mirror the agency’s logic, only to predict outcomes of it. Like two students producing correct answers to a math problem, unless they “show their work,” it would be wrong to assume they used the same method, let alone that either used the right method or appropriately applied it. The

---

171. Jonathan Zittrain, *The Hidden Costs of Automated Thinking*, NEW YORKER (July 23, 2019), <https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking> [https://perma.cc/D9YK-JYHZ].

172. For example, machine learning image recognition systems are famous for appearing to perform well on a task but actually relying on a simplistic and poorly-chosen heuristic. In addition, as Kaminsky describes, algorithms lack the contextual understandings of acceptable bases for decision making and the common sense of humans. Kaminsky, *supra* note 22, at 14–15.

design process of machine learning systems does not explicitly transfer expert reasoning and therefore does not create the pattern of displacement found in expert systems. Yet because machine learning classifiers are developed by studying the outcomes of agency logics rather than the logic itself, it creates potentially more troubling displacement effects.

b. Element 2: Constraints on Policymaking Evolution

Machine learning systems develop *probabilistic models that optimize for a particular goal*—and then, where they are allowed, update them as they learn in the field. Rather than the automated rules that concerned Citron, the constraints imposed on agency discretion in machine learning systems are found in choices about what a system is optimizing for and how the goal is operationalized going forward within the system.

Once deployed, the logic of the model—whether fixed, or allowed to learn over time—remains constrained by the assumptions and choices made during design. In contrast, the judgment of agency professionals and staff may evolve over time, sometimes on a gentle slope, but at other times diverging swiftly in response to new research, new political winds, or other internal or external jolts. While a machine learning model may learn new ways to optimize for the goal established, it is tethered to the beliefs and biases that are fixed in the model, as well as the assumptions and ingested data used during development. As a result, machine learning systems can instill patterns of racism, debunked science, or other faulty or unjust reasoning that may be captured in the training data or optimization choices.

Even if the policies embedded in a model are fully aligned with agency decision making at the time of its initial deployment, if the system is not updated to reflect changing agency understanding of sound judgment and agency practice, machine learning systems can constrain agency discretion in particularly problematic ways.

Finally, the extent to which a model developed on one data set can be safely used on another is an immensely important policy question. It is well documented that models trained on one data set can perform catastrophically poorly on a data set that many might assume to be similar by some set of metrics.<sup>173</sup> For example, models trained on newswire copy perform poorly on texts from other domains.<sup>174</sup> Even for discrete Natural Language Processing

---

173. Selbst et al., *supra* note 156, at 4–5 (calling this the “portability trap” and tying it to the quest for abstractions and tools that can be reused across contexts).

174. See David Bamman, *Natural Language Processing for the Long Tail*, DIGITAL HUMAN. 2 (2017) (“[T]he performance of an out-of-the-box part-of-speech tagger can, at worse, be half that of its performance on contemporary newswire. On average, differences in style amount to a drop in performance of approximately 10–20 absolute percentage points across tasks.”).

(NLP) tasks, such as identifying words as nouns, verbs, adjectives, etc.—called part-of-speech or grammatical tagging or word-category disambiguation—which lay people might consider simple and transferable across corpora, models trained on news articles perform quite poorly on literary works.<sup>175</sup>

iii) The Challenge: Reintroducing Expert Justification for Agency Decisions

The way in which machine learning systems generate decisions without decision-making processes challenges administrative law's fundamental mandate of reasoning. To be legitimate, reliance on machine learning in governance requires processes that reintroduce appropriate expertise in providing justifications for administrative choices.

The requirement of “justification” regarding a system’s design and its subsequent choices is critical. Justification is distinct from two elements identified by computer scientists related to system accountability: interpretability—properties or qualities or techniques related to a system that help humans understand the relationship between inputs and outputs<sup>176</sup>—and explainability: the ability to explain the operation of a machine learning system in human terms.<sup>177</sup> Explainability provides the reasoning behind the relationship between inputs and outputs interpretability reveals.<sup>178</sup>

While both interpretability and explainability might be helpful, they are not sufficient to satisfy administrative legitimacy.<sup>179</sup> Explaining an algorithm’s operation without providing informed justifications for the choices reflected

---

175. See *id.* (summarizing research investigating the disparity between training data and test data for several NLP tasks).

176. Finale Doshi-Velez & Been Kim, *Towards a Rigorous Science of Interpretable Machine Learning*, ARXIV (Mar. 2, 2017), <https://arxiv.org/pdf/1702.08608.pdf> [<https://perma.cc/6CVL-DWRK>].

177. Explanations can describe the operation of a model in general (so-called “global” explanations) or for a particular mechanism in the model used to relate inputs and outputs (so-called “local” explanations). Upol Ehsan et al., *Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations*, ARXIV (Dec. 19, 2017), <https://arxiv.org/pdf/1702.07826.pdf> [<https://perma.cc/GK5X-MXC5>].

178. Both explainability and interpretability are areas of debate and research among computer scientists and the multiple disciplines within the broader “fairness, accountability and transparency” research community. For a discussion of these terms and others within and across relevant disciplines, see Nitin Kohli et al., *Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems*, CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY (2018).

179. For a discussion of the relationship between explanations and justifications in criminal law, and probable cause in particular, see Kiel Brennan-Marquez, “*Plausible Cause*”: *Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249, 1288 (2017) (“Apart from safeguarding constitutional values, explanations also vindicate rule-of-law principles. A key tenet of legality, separating lawful authority from ultra vires conduct, is the idea that not all explanations qualify as justifications.”).

in that operation fails the “arbitrary and capricious” threshold. Instead, design choices that embed policy choices must reflect reason, a rational connection to the facts, context, and the factors mandated by Congress in the relevant organic statute, while avoiding elements extraneous to the legislative command. And such justification, in turn, requires the application of a range of forms of expertise, including technical knowledge about machine learning and algorithm design, as well as statistics, domain expertise, and specialized fields such as those represented in the Fairness, Accountability, and Transparency in Machine Learning (FAT\*) and ethics, law and sociology communities, whose members investigate the social and political consequences of algorithmic systems.

As an initial matter, when systems are adopted by governments, agencies must be able to enlist sufficient expertise at the design phase to permit knowledgeable exploration of technical design and data choices that embed policy. As discussed above, decisions about system goals (what is optimized for), how to operationalize the goal into a target variable for the system to optimize for, and what modeling frameworks to use, all require expert input because they are fundamental policy decisions. In addition, determinations about the data—its selection, curation, cleaning, and similarity to the data on which it will be used—and about the triggers for updating or replacing it are all essential policy questions with which agencies must grapple explicitly. Decisions about the use and inclusion of data about protected traits warrant particular scrutiny. Precise numerical limits such as cut-offs or thresholds—particularly those that cabin discretion—must be the product of reasoned agency decision making.

Additionally, consistent with the case law’s emphasis on agency discretion, agencies must comprehend and address the impact of a system on future agency choices. Traditionally, agency staff are able to adjust to new informational inputs as a situation requires.<sup>180</sup> They can selectively pull data in and out of the decision-making frame based on case-specific, situational knowledge. Machine learning—like other automated systems—can constrain the ability to flexibly alter the data brought to bear on a decision in response to the particular problem or person presented.<sup>181</sup> While machine learning systems can process tens of thousands of data points, they can only consider the data predetermined to be relevant. Setting bounds on what can be considered—ensuring, for example, that information about race, gender, age, or other protected attributes does not infiltrate agency decision making—may

---

180. Cf. Kaminsky, *supra* note 22, at 13–14 (describing how moving from a human to an automated decision can eliminate “cultural knowledge about what is or is not an appropriate decisional heuristic in a particular case”).

181. See Citron, *supra* note 19, at 1304 (explaining that policies allowing “individuals to plead extenuating circumstances that software cannot anticipate” should not be automated).

align with a simplistic notion of fairness. But using such simple categories has been found to frequently be at odds with justice, the goal it purportedly serves.

Even where systems are billed as “decision support,” ostensibly allowing decision makers to consider other information, automation bias may lead to overreliance on machine outputs.<sup>182</sup> Without efforts—policy, system design, and accountability frameworks—to foster questioning, agency staff may come to defer to machine outputs, particularly over time. In doing so, systems may elevate ideals of procedural fairness at the cost of substantively just and right outcomes. Angele Christin’s research documents that automation bias may not always result and suggests that this tension between different visions of fairness may be a point of resistance. She found different kinds of resistance and tinkering with recidivism risk tools in the justice system—some of which appears to be grounded in battles over competing conceptions of fairness, its relation to justice, and the role that discretion, rather than rigidity, plays in advancing the latter.<sup>183</sup> The risks posed by automation bias nevertheless loom large when relevant professional, regional, or site-specific experts are not consulted during system development,<sup>184</sup> or when the systems are acquired as commercial off-the-shelf products rather than collaboratively developed or tailored for the conditions and context of use.

Because of this limited input and the ways these systems constrain agency staff’s ability to expand or narrow the data used to render a decision, and to shift their reasoning over time, machine learning systems risk upsetting context-specific, domain-specific, and evolving judgments—key rationales for agency existence. For these reasons, the interfaces and policies that structure agency staff’s interactions with machine learning systems must be the subject of agency deliberation and involve reasoned application of expertise about the human-machine loop. This includes agency policies of the type Citron

---

182. See Kate Goddard et al., *Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators*, 19 J. AM. MED. INFORMATICS ASS’N 121 (2011) (reviewing literature on automation bias in health care clinical decision-support systems).

183. Angèle Christin, *Algorithms in practice: Comparing web journalism and criminal justice*, BIG DATA & SOC’Y 1, 9 (July 16, 2017) (discussing a senior judge’s perspective on recidivism risk tools: “I don’t look at the numbers. There are things you can’t quantify . . . [y]ou can take the same case, with the same defendant, the same criminal record, the same judge, the same attorney, the same prosecutor, and get two different decisions in different courts. Or you can take the same case, with the same defendant, the same judge, etc., at a two-week interval and have completely different decision. Is that justice? I think it is” and finding probation officers similarly resisting rigidity by tinkering with the criteria to obtain the score they thought adequate for a given defendant).

184. For example, criminal justice risk-assessment tools, which have been around for decades and are often simply logistic regressions, are almost uniformly created outside of the jurisdictions in which they are deployed. There are fewer than sixty tools used across the entire United States. Angèle Christin et al., *Courts and predictive algorithms*, DATA & CIV. RTS.: A NEW ERA POLICING & JUST. (Oct. 27, 2015).

recommends—such as training on automation bias and requiring explanations of the facts and findings produced by automated systems on which agency staff rely<sup>185</sup>—as well as decisions about system interfaces, such as whether to communicate uncertainty and, if so, how to do so.

b) Democratic Elements in Reasoned Decision Making About Machine Learning Systems

In addition to gathering the expertise necessary to understand, explain, and justify these design choices, the arbitrary and capricious jurisprudence points to deeper issues about what meaningful deliberation would require in the machine learning context. Specifically, its emphasis on the public disclosure of the decisions made and the assumptions behind them reflects the reality that “[m]odels and proxies are built on numerous assumptions, often based in scientific principles but also laden with value judgments.”<sup>186</sup> As political scientist Sheila Jasanoff describes, “there is growing awareness that science cannot answer all of our questions about risk and that both scientific and value judgments are involved in the processes of risk assessment and risk management.”<sup>187</sup>

Agencies cannot create a meaningful record of pertinent “issues of policy” involved in machine learning system design and “why the agency reacted to them as it did”<sup>188</sup>—indeed they cannot be transparent to the public, if they fail to disclose both information about the code and its underlying models, limits, defaults, assumptions, training data, and the very fact that they engaged in a policy judgment and how those judgments were resolved. Decisional transparency must involve not only openness about design but also publicity about the very existence and political nature of value questions being resolved through design processes.

Thus “political visibility,”<sup>189</sup> rather than algorithmic transparency, is the essential characteristic of legitimate processes for adopting complex algorithmic systems. Administrative legitimacy is predicated on the explicit public articulation of value choices under consideration and transparent deliberation about their resolution.<sup>190</sup> When values are embedded in design

---

185. Citron, *supra* note 19, at 1306–07.

186. Sara A. Clark, *Taking a Hard Look at Agency Science: Can the Courts Ever Succeed*, 36 ECOLOGY L.Q. 317, 331 (2009).

187. Sheila Jasanoff, *Cultural Aspects of Risk Assessment in Britain and the United States*, in THE SOCIAL AND CULTURAL CONSTRUCTION OF RISK 359, 359 (B. B. Johnson & V. T. Covello eds., 1987).

188. Auto. Parts & Accessories Ass’n v. Boyd, 407 F.2d 330, 338 (1968).

189. Mulligan & Bamberger, *supra* note 12, at 251.

190. See, e.g., *Boyd*, 407 F.2d at 338 (D.C. Cir. 1968) (noting that an agency rulemaking record must make visible “what major issues of policy were ventilated” and “why the agency reacted to them as it did”).

choices they are “less visible as law, not only because it can be surreptitiously embedded into settings or equipment but also because its enforcement is less public.”<sup>191</sup> The regulative features of technology design can appear “constitutive”—non-normative and part of the natural state of things.<sup>192</sup> If they are not explicitly surfaced (as they often are not), the policy decisions built into machine learning systems “fade into the background and hide the political nature of [their] design.”<sup>193</sup> Value trade-offs, unrecognized as governance, remain unaddressed at the design stage, hindering both robust consideration of substantive policy and ex post oversight.

#### IV. BUILDING ADMINISTRATIVE PROCESS FOR MACHINE LEARNING

Reasoned decision making about machine learning system adoption requires both deep subject-matter expertise, a key grounding for delegating the power to implement and enforce laws to agencies,<sup>194</sup> and processes ensuring that policies embedded in system design appear and remain politically salient to agency employees as well as to the public and the political branches. Unconsidered resolution of policy issues—including those impacting protected classes—constitutes the epitome of arbitrary and capricious decision making and an abdication of policymaking responsibility at the heart of administrative legitimacy,<sup>195</sup> which displaces expert agency judgment with algorithmic output. Furthermore, this disappearance of values can unintentionally lead agencies that are heavily dependent on machine learning systems to ossify policies that no longer serve the agency’s interests and goals.

These pitfalls have particular resonance when policy is driven by machine learning system design, where the metrics by which legal rights and obligations are fixed in individual cases are dynamic, and where forms of localization are necessary for performance, including on values such as fairness. Learning systems “learn.” Their analytics and algorithms evolve and change according

---

191. Lee Tien, *Architectural Regulation and the Evolution of Social Norms*, 7 YALE J.L. & TECH. 1, 22 (2004).

192. Mireille Hildebrandt, *Legal and Technological Normativity: More (and Less) than Twin Sisters*, 12 TECHNE 169, 179 (2008).

193. Mulligan & Bamberger, *supra* note 12, at 778.

194. Cass R. Sunstein, *Constitutionalism After the New Deal*, 101 HARV. L. REV. 421, 442–44 (1987) (discussing the “New Deal belief in the importance of technical expertise” as a justification for accordinng agencies “a large measure of autonomy”).

195. Cf. Jody Freeman & Adrian Vermeule, *Massachusetts v EPA: From Polities to Expertise*, SUP. CT. REV. 51 (2007) (discussing the Supreme Court’s “expertise-forcing” jurisprudence ensuring that “agencies actually do exercise expert judgment”); Heckler v. Chaney, 470 U.S. 821, 833 n.4 (1985) (applying arbitrary and capricious review, even in enforcement contexts characterized by high executive discretion, when an agency’s failure to exercise its discretion “amount[s] to an abdication of its statutory responsibilities”).

to the logics that machines induce by observing human actions.<sup>196</sup> The policies that these systems implement will change over time and be driven by machine rather than human reasoning, in a way that displaces the discretion of agency staff going forward.<sup>197</sup>

A key justification for delegating substantive policy choices to agencies, of course, is their ability to revise policy “in light of evolving societal, political, and technological circumstances.”<sup>198</sup> Yet when those revisions generate legal effects, administrative law requires engagement, reasoning, and transparency.<sup>199</sup> The challenge of public machine learning adoption, then, is to ensure such process as policy is made on a continuum—at design time, configuration time, and run time.<sup>200</sup>

While a range of agency processes might address the opacity of complex machine learning systems and account for the technocratic and democratic demands of reasoned governance,<sup>201</sup> this Part recommends elements for a framework of public machine learning adoption that satisfies both.

At its core is the reliance on centers of expertise—on the model of the USDS and the 18F “skunk works” team first developed by the Obama Administration—that develop and provide shared technical knowledge in ways that address expertise gaps across agencies while providing a systemic approach to the use of technology in government activity. We further identify different tools that such a centralized effort should employ, including

---

196. See *supra* Section III.B.1 (discussing decision making by machine learning systems).

197. See Mulligan & Bamberger, *supra* note 12 (discussing further the ways that values in technology change over time as technology is appropriated by users in new and unexpected ways, and how technology interactions with business models, organizational structures, and other technologies in ways that can transform its effects, use, and impact on values); Harry Surden, *Structural Rights in Privacy*, 60 SMU L. REV. 1605 (2007) (discussing ways that technology affects the “latent structural constraints” that work to protect values in addition to and in conjunction with legal measures).

198. Kenneth A. Bamberger, *Provisional Precedent: Protecting Flexibility in Administrative Policymaking*, 77 N.Y.U. L. REV. 1272, 1280 (2002); see Matthew C. Stephenson, *Public Regulation of Private Enforcement: The Case for Expanding the Role of Administrative Agencies*, 91 VA. L. REV. 93, 139 (2005) (“Flexibility, like expertise, is often invoked to justify delegation of substantive policy choices to agencies.”).

199. See, e.g., FCC v. Fox Television Stations, Inc., 556 U.S. 502 (2009) (applying arbitrary and capricious review to a change of agency policy applied in an adjudication); Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 30 (1983) (applying arbitrary and capricious review to a change in agency policy reached through rulemaking).

200. Clark et al., *supra* note 50, at 463 (discussing how values tussles play out at design, redesign, configuration, and run time). This “developer” perspective on the ability and need to address values at every stage of the process is captured in the security adage, “Secure by Design, Secure by Default, Secure in Deployment.” Steve Lipner, *The Trustworthy Computing Security Development Lifecycle*, 20TH ANN. IEEE COMPUTER SECURITY APPLICATIONS CONF. (2004).

201. See TAXPAYER ADVOCATE SERV., *supra* note 60.

algorithmic impact assessments, which not only involve deliberation about technical choices themselves, but also surface their policy implications publicly.

This framework uses two critical means to build on that visibility and foster public participation, political oversight, and informed agency engagement, during both system design and deployment. The first is institutional. It suggests that the adoption of processes that engage the public as policy is made through design. The second is technical. It suggests that reasoned agency deliberation about policy requires that machine learning systems adopted by governments reflect “contestable” design from the start—design that supports meaningful contestability throughout the system lifecycle by permitting an ongoing role for agency staff in shaping the policies embedded in systems.

Together, these tools focus on the development of expertise and the surfacing of politics while emphasizing judgment, coherence, efficiency, and transparency in setting administrative policy.

#### A. INFORMING AGENCY DELIBERATION WITH TECHNICAL EXPERTISE

##### 1. *Reviewing Piecemeal Efforts*

In other works, we have argued that technology should be used to govern only when an agency has access to relevant technical expertise and the ability to consider a wide scope of public values that may be implicated in the use of technology to regulate.<sup>202</sup> We identified a range of options for acquiring relevant expertise, including agency hiring, drawing on the expertise of other agencies, and soliciting expertise from external stakeholders.<sup>203</sup>

Flavors of these alternative approaches to leveraging expertise are visible in some novel processes around algorithmic systems set up by select agencies and pending legislative proposals that seek to address both the potential for inherent bias and privacy risks.

Several jurisdictions at all levels have chosen to create public or quasi-public bodies to aid in the analysis of algorithmic-system adoption. At the local level, New York City passed an algorithmic accountability bill assigning a task force to examine the way city government agencies use algorithms.<sup>204</sup> New York’s Automated Decision Systems Task Force, comprised of a cross-

---

202. Mulligan & Bamberger, *supra* note 12, at 759, 768–70. We also note that stakeholders must have the technical expertise to meaningfully participate and suggest alternative models for providing it. *Id.* at 775–76.

203. *Id.* at 768–70.

204. See Int. No. 1696-A, Automated decision systems used by agencies (NYC 2018), <http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0> [<https://perma.cc/8AZR-DYB9>].

disciplinary group of city officials and outside experts,<sup>205</sup> is tasked with recommending a process for reviewing the city's use of automated decision systems to ensure equity and opportunity. The Task Force has held two public workshops to date.<sup>206</sup> While the law brings experts in to assist the city in developing review processes and conducting reviews, in recent testimony submitted to the New York City Council Committee on Technology, two Task Force members wrote:

Task Force members have not been given any information about ADSs [automated decision systems] used by the City. To date, the City has not identified even a single system. Task Force members need to know about relevant systems used by the City to provide meaningful recommendations.<sup>207</sup>

They further reported that the Task Force had nevertheless made “meaningful progress in developing a methodology for eliciting relevant information about ADSs, using so-called “ADS Cards” that ask developers and operators to provide specific details about the system in question,” but that the City forced them to abandon the project.<sup>208</sup>

At the state level, the Pennsylvania Commission on Sentencing<sup>209</sup> has brought expertise into the creation of their recidivism risk system in numerous ways. The Commission was tasked with creating a recidivism risk tool—initially paper-based but over time governed by software.<sup>210</sup> To develop these tools, the Commission has conducted its own research,<sup>211</sup> partnered with

---

205. *Members*, NYC AUTOMATED DECISION SYSS. TASK FORCE, <https://www1.nyc.gov/site/adstaskforce/members/members.page> [<https://perma.cc/R5VZ-LMQL>] (last visited Oct. 10, 2019).

206. *Past: Forum #2: Transparency*, NYC AUTOMATED DECISION SYSS. TASK FORCE <https://www1.nyc.gov/site/adstaskforce/events/events.page> [<https://perma.cc/LPR2-73JV>] (last visited Oct. 10, 2019).

207. *Testimony regarding Update on Local Law 49 of 2018 in Relation to Automated Decision Systems (ADS) Used by Agencies before NYC Council Comm. on Technology* (Apr. 4, 2019) (testimony of Julia Stoyanovich & Solon Barocas), [https://dataresponsibly.github.io/documents/StoyanovichBarocas\\_April4,2019testimony.pdf](https://dataresponsibly.github.io/documents/StoyanovichBarocas_April4,2019testimony.pdf) [<https://perma.cc/KY2W-M6PY>].

208. *Id.*

209. The Commission was created in 1978 to develop and oversee the development of statewide guidelines to promote fairer and more uniform sentencing.

210. Judicial Code and Prisons and Parole Code, 95 PA. STAT. ANN. §§ 42, 61 (2010) (providing for the adoption of a risk assessment tool).

211. For overview of research, see *Research Overview of the Sentence Risk Assessment Instrument*, PA. COMMISSION ON SENT’G 4–5 (Oct. 2018), <http://pcs.la.psu.edu/guidelines/proposed-risk-assessment-instrument/additional-information-about-the-proposed-sentence-risk-assessment-instrument/research-overview-of-the-sentence-risk-assessment-instrument-1/view>; for reports to date, see PA. COMMISSION ON SENT’G, <http://pcs.la.psu.edu/>

academic institutions to hold workshops,<sup>212</sup> commissioned academic research,<sup>213</sup> and most recently commissioned an independent evaluation by the Urban Institute.<sup>214</sup>

At the federal level, the Organ Procurement and Transplantation Network (OPTN) and Task Force on Organ Procurement and Transplantation, which were established to regulate the donation and allocation of organs through the network,<sup>215</sup> have a lengthy history of engaging experts in the design of their organ allocation systems.<sup>216</sup> The OPTN's Board of Directors, which develops policies to govern the operation of the OPTN, relies on numerous expert

---

publications-and-research/risk-assessment [https://perma.cc/RH9P-CNRU] (last visited Aug. 1, 2019).

212. See *Pennsylvania Criminal Justice Roundtable*, PENNSTATE CRIM. JUST. RES. CTR. (May 19–20, 2011), <https://justicecenter.la.psu.edu/research/projects/pennsylvania-criminal-justice-roundtable> [https://perma.cc/WFQ7-JF27] (convening state criminal justice policymakers and experts in offender risk assessment and sentencing, including faculty at the Penn State Criminal Justice Research Center).

213. Matthew DeMichele & Julia Laskorunsky, *Sentencing Risk Assessment: A Follow-up Study of the Occurrence and Timing of Re-Arrest among Serious Offenders in Pennsylvania*, PA. COMMISSION ON SENT'G (May 2014), [https://justicecenter.la.psu.edu/research/projects/files/PCS%20\\_Risk%20Assessment\\_Tool.pdf](https://justicecenter.la.psu.edu/research/projects/files/PCS%20_Risk%20Assessment_Tool.pdf)/view [https://perma.cc/PZ9T-TB5N] (analyzing the relationship between offender and case characteristics and likelihood of recidivism).

214. See GEN. ASSEMB. COMM'N ON SENTENCING, PROPOSED SENTENCE RISK ASSESSMENT INSTRUMENT FOR 204 PA. CODE CHAPTER 305; RESPONSES TO PUBLIC COMMENTS; REQUEST FOR PROPOSALS; 48 Pa.B. 5445, at 2 (Sept. 1, 2018) (reporting that “in April 2018, following publication of a revised proposal, staff provided the Urban Institute with a complete set of files related to construction of the instrument (e.g., data, syntax, etc.) to begin the external review”).

215. National Organ Transplant Act, Pub. L. No. 98-507, 98 Stat. 2339 (1984) (establishing the Organ Procurement and Transplantation Network and banning organ sales). For a detailed legislative history covering the process and substantive considerations leading up to enactment, see Jed Adam Gross, *E. Pluribus UNOS: The National Organ Transplant Act and Its Postoperative Complications*, 8 YALE J. HEALTH POL'Y L. & ETHICS 145, 207–22 (2008).

216. See Gross, *supra* note 215, at 228–30 (describing the NOTA Task Force on Organ Transplantation, which included “medical professionals, social and behavioral scientists, a legal scholar, an ethicist with a background in religious studies, representatives of the public and private insurance sectors, and representatives of the general public” who were tasked with developing recommendations for organ transplantation, including allocation policies); *id.* at 220 (discussing the requirement that procurement organizations include transplant professionals on their board of directors or advisory board); *see also* Organ Procurement and Transplantation Network, 42 C.F.R. § 121.12 (2019) (establishing the Advisory Committee on Organ Transplantation, governed by the Federal Advisory Committee Act, to provide input on proposed OPTN policies and other matters); 42 C.F.R. § 121.3 (directing The OPTN Board of Directors to include “approximately 50% transplant surgeons or transplant physicians”); Mark D. Stegall et al., *Why Do We Have the Kidney Allocation System We Have Today? A History of the 2014 Kidney Allocation System*, 78 HUM. IMMUNOLOGY 4 (2017) (describing the deliberations and adoption of the kidney allocation system (KAS), including expert input, adopted in December 2014).

advisory committees.<sup>217</sup> Those committees engage in extensive fact-finding and deliberation.<sup>218</sup>

Since its inception, organ transplantation and allocation have been viewed as both highly technical and deeply political.<sup>219</sup> The algorithmic systems used to allocate organs are provided by the United Network for Organ Sharing (UNOS) under contract with the OPTN, which makes extensive information about the factors and weighting within allocation algorithms available.<sup>220</sup> OPTN/ONUS provides a central source of expertise around algorithmic allocation systems that governs the operation of all member transplant hospitals, organ procurement organizations, and histocompatibility labs in the United States. Thus, while numerous institutions are involved in procuring and transplanting organs, they benefit from a centralized set of expert resources. While bounded by legislation and overseen by the Department of Health & Human Services (HHS),<sup>221</sup> the OPTN/UNOS system for establishing policies to embed in the allocation algorithms is largely driven by experts enlisted through committees. Those committees are highly specialized, focusing on the specific factors that influence transplant success and considerations of just allocation with respect to particular organs.

The Taxpayer Advocate's office within the IRS offers an alternative model focused on agency development of internal expertise. While not set up specifically to address algorithmic systems, the Taxpayer Advocate has played a key role in identifying the problem with embedded policies in algorithmic

---

217. For a list of current committees, see *Committees*, U.S. DEP'T HEALTH & HUM. SERVS., <https://optn.transplant.hrsa.gov/members/committees> [https://perma.cc/LWR6-6WWM] (last visited Aug. 1, 2019). Proposed policies OPTN wants to enforce, including allocation policies, must be provided to the Secretary of the Department of Health and Human Services sixty days prior to implementation, the Secretary must publish significant proposed policies in the Federal Register, and they are not enforceable until approved by the Secretary. Department of Health and Human Services. 42 C.F.R. § 121.4 (b)(2) (2019).

218. For a description of a recent process for revising kidney allocations, see Mark D. Stegall et al., *Why Do We Have the Kidney Allocation System We Have Today? A History of the 2014 Kidney Allocation System*, HUM. IMMUNOLOGY 78.1 4, 4–8 (2017). For details on the allocation formula and discussion of use of a software system—Kidney-Pancreas Simulated Allocation Model (KPSAM)—to assess policy proposals, see Bhavna Chopra & Kalathai K. Sureshkumar, *Changing Organ Allocation Policy for Kidney Transplantation in the United States*, 5 WORLD J. TRANSPLANT 38 (2015).

219. See Jed Adam Gross, *supra* note 215, at 182–83 (describing issues from blood pressure to distributive justice in congressional hearings).

220. See *Organ Procurement and Transplantation Network Policies*, U.S. DEP'T HEALTH & HUM. SERVS., <https://optn.transplant.hrsa.gov/governance/policies/> [https://perma.cc/JC5R-XL4R] (last visited Oct. 11, 2019) (detailing allocation rules and weighting for various organs).

221. The Final Rule requires organ allocation to be based on sound medical judgment; make the best use of donated organs; avoid wasting organs and futile transplants; and promote patient access to transplants. 42 C.F.R. § 121.8 (2019).

systems.<sup>222</sup> Although the Advocate has not been fully successful in ensuring that embedded policies are vetted internally, consistent with other agency review and approval processes for all written policy, procedures, and guidance before issuance and publication, they have drawn agency attention to problematic aspects of systems, some of which have been reformed.<sup>223</sup>

Finally, a bill pending in California takes a different outward-looking approach. Rather than bringing expertise directly into government work, it would leverage the expertise of firms that provide AI-based products and services to public agencies. The bill would require firms to provide information about data curation and processing, as well as bias mitigation strategies, in their contracts with public agencies.<sup>224</sup> Requirements such as this could harness the knowledge and expertise of those close to the problem toward public ends—allowing the experts to identify and address problems, but opening them up for regulator and public scrutiny.<sup>225</sup>

Over the past year, engineers and other employees at firms that build machine learning systems have been increasingly active in objecting to their use in specific contexts or toward specific ends.<sup>226</sup> Providing engineers and other technical professionals with external justifications for considering the implications of their work on privacy, discrimination, and other substantive values can legitimize these nascent expressions of concern among technical professionals, encourage internal policing,<sup>227</sup> and provide a platform to support

---

222. TAXPAYER ADVOCATE SERV., *supra* note 60 (arguing that “automated systems and software applications require transparency, and employee guidance embedded in systems must be reviewed and continually analyzed for proper application” but noting that “policy guidance embedded in systems is neither reviewed internally nor published externally”).

223. *Id.*

224. Artificial Intelligence: Reporting, S.B. 444, 2019 Legis., Reg. Sess. (Cal. 2019), was a state senate bill in California introduced by Senator Umberg (D) on February 21, 2019 to establish that a contractor, vendor, or qualifying business shall maintain a written record of the data used relating to any use of artificial intelligence for the delivery of a product or service to a public entity. The records shall include all the following information: (1) the purpose of the data; (2) a description of the categories of the data; (3) the source of the data; (4) the demographics or information related to a characteristic listed in subdivision (a) of Section 11135 of the Government Code that is used as a source of input data for the creation of the artificial intelligence system. The business shall disclose to public entities that it relies on AI, information about the data, and “its internal policies for how bias in the artificial intelligence system is identified and mitigated.” *Id.* at § 3505(c)(3).

225. See generally Kaminsky, *supra* note 22, at 26–39 (calling for collaborative governance of algorithms used in the private sector).

226. Meredith Whittaker et al., *AI Now Report 2018*, AI Now INST. 40–42 (Dec. 2018), [https://ainowinstitute.org/AI\\_Now\\_2018\\_Report.pdf](https://ainowinstitute.org/AI_Now_2018_Report.pdf) [https://perma.cc/7WER-78L4] (summarizing host of employee actions at various companies to stop particular AI projects and system uses).

227. Such approaches could bolster algorithmic whistleblowing by employees envisioned by Sonia Katyal. Katyal, *supra* note 20.

collaborative work across institutions on issues such as bias mitigation strategies.

Each of these approaches has their advantages. Given the current level of uncertainty about how best to mitigate bias, leveraging the deep knowledge of the people and institutions close to the problem—as the OPTN and California proposal do in different ways—while exposing them to external review taps into the relevant expertise while fostering attention and accountability to a broader set of values.<sup>228</sup> The Pennsylvania Sentencing Commission, OPTN, and the California proposals lay different internal processes out for the public. OPTN and the Pennsylvania Sentencing Commission provide more detailed information about the properties of the algorithmic systems themselves. Requiring decision makers to explain themselves to outsiders with different perspectives on what policies ought to be embedded, or what bias mitigation strategies should be taken, can push those developing algorithmic systems to engage in more critical technical practice.<sup>229</sup> This approach may yield processes that make agencies, managers, or even individual engineers accountable for considering the impact of algorithmic systems from perspectives beyond speed and task performance, fostering a greater sense of responsibility for the outcomes they produce in the world.

In contrast to the Sentencing Commission and OPTN, the New York City Task Force’s mandate cuts across numerous agencies. While the factors contributing to its slow progress are not fully clear, one factor may be that focusing on algorithmic systems writ large, rather than algorithmic systems within a particular domain and with participation by substantive experts in that domain, may be perceived as more threatening or less beneficial.

## 2. *A Paradigm for Expert Decision Making*

However, these piecemeal attempts to inform case-by-case deliberation with internally- or externally-derived technical expertise about methodologies under consideration offer only a proverbial finger in the dike<sup>230</sup> against the oncoming flood of government usage of machine learning. While helpful, they are fragmentary, costly, and time-consuming. Furthermore, they often (with the possible exception of the New York City Task Force) fail to address the

228. See KENNETH A. BAMBERGER & DEIRDRE K. MULLIGAN, PRIVACY ON THE GROUND: DRIVING CORPORATE BEHAVIOR IN THE UNITED STATES AND EUROPE 190 (2015) (describing importance of exposing business practices to scrutiny) [hereinafter PRIVACY ON THE GROUND]; see generally Kenneth A. Bamberger, *Regulation as Delegation: Private Firms, Decisionmaking, and Accountability in the Administrative State*, 56 DUKE L.J. 377, 445–46 (2006).

229. Philip E. Tetlock, *Accountability: The Neglected Social Context of Judgment and Choice*, 7 RES. ORGANIZATIONAL BEHAV. 297, 314–21 (1985) (reviewing research evidence).

230. See MARY MAPES DODGE, HANS BRINKER OR THE SILVER SKATES: A STORY OF LIFE IN HOLLAND 105–09 (recounting a tale about a Dutch boy who saves his country by putting his finger in a leaking dike).

expertise gap across agencies by providing easily deployable, coordinated, and regularized policies and methods governing, and processes for assessing, algorithmic governance tools.

Indeed, centralization of expertise may be particularly important for attending to embedded policies in algorithmic systems. Organ donation is an area where the need for allocation strategies and their deeply political and technical nature were recognized and accounted for up front. But in many areas, algorithmic systems are replacing or aiding decision making that was conceived of and practiced in a more clinical way—relying on human judgment informed by expertise gained through education and training, refined through tacit knowledge and intuition developed experientially through practice, and perhaps discussion and feedback with others—and therefore lacks the level of formalization found in organ donation or even recidivism risk. To the extent algorithmic systems are being introduced in areas where clinical decision-making methods reign, the need for relevant technical expertise in addition to domain-specific expertise will be high; however, internal agency capacity to provide it may be low.

a) The Institutional Paradigm: USDS and the 18F “Skunk Works”

In this vein, efforts begun during the Obama Administration to create centers-of-expertise or shared technical knowledge offer a paradigm for a more systemic approach to the use of technology in government activity. This work has continued with the Office of Federal Procurement Policy (OFPP),<sup>231</sup> which, in conjunction with the USDS, created a Digital IT Acquisition Program (DITAP) to help contracting professionals gain the “expertise needed to support the delivery of digital information (i.e., data or content) and transactional services (e.g., online forms and benefits applications) across a variety of platforms, devices, and delivery mechanisms (e.g., websites, mobile applications, and social media).”<sup>232</sup>

President Obama created a centralized pool of technical expertise, USDS, housed within the Executive Office of the President of the United States. USDS is a transitory pool of experts in design, engineering, or product

---

231. OFPP was created by Congress to provide overall direction on “Government-wide procurement policies, regulations, procedures, and forms”; and to “promote economy, efficiency, and effectiveness” in procurement. 41 U.S.C. § 1101(b) (2012).

232. Memorandum from Lesley A. Field, Deputy Adm’r, Office of Mgmt. & Budget, on Establishment of Federal Acquisition Certification in Contracting Core-Plus Specialization in Digital Services (FAC-C-DS) to Chief Acquisition Officers & Senior Procurement Execs. A-1 (May 18, 2018), [https://techfarhub.cio.gov/assets/files/FAC\\_C\\_Digital\\_Services\\_5-18-18.pdf](https://techfarhub.cio.gov/assets/files/FAC_C_Digital_Services_5-18-18.pdf) [<https://perma.cc/75MU-WWS9>].

management brought in from outside government for “tours” of service.<sup>233</sup> USDS provides consultation services to federal agencies on information technology and guidance. For example, to develop greater technical expertise among the contracting professionals distributed throughout federal agencies, USDS created the Digital Services Playbook to propagate best practices from both the private and public sector across federal agencies. The Playbook is accompanied by the TechFAR Handbook, which provides guidance to agencies on how to comply with Federal Acquisition Regulations while using the Digital Services Playbook. Similarly, 18F—sometimes described as a skunk works<sup>234</sup> project for government—is an office within the General Services Administration (GSA) that collaborates with other agencies to fix technical problems, build products, and improve how government serves the public through technology.<sup>235</sup> They offer agencies access to expert teams of designers, software engineers, strategists, and product managers skilled in user-centered development, and other design and acquisition expertise.

But in addition to providing experts to government agencies on a task-specific model, 18F also develops guides on topics such as accessibility, agile development, and design methods to assist federal agencies.<sup>236</sup> An example of such guidance documents is the U.S. Web Design Standards (USWDS),<sup>237</sup> a joint product of 18F and USDS.<sup>238</sup> The USWDS provide guidance, templates, and models for developers and designers; it covers design including accessibility, front end and back end coding, and provides code and performance guidelines.<sup>239</sup> The mix of centralized expert staff who can be deployed for limited periods of time to assist agencies with complex

---

233. *How We Work*, U.S. DIGITAL SERV., <https://www.usds.gov/how-we-work> [<https://perma.cc/QZY3-YV6C>] (last visited Oct. 3, 2019).

234. See Dave Zvenyach, *Joining 18F*, V. DAVID ZVENYACH'S BLOGS (Jan. 20, 2015), <https://esq.io/blog/posts/joining-18f/> [<https://perma.cc/VM9T-F4C8>] (describing 18F as “a modern-day digital skunk works . . . [housing] some of the best and brightest developers in America, building applications that agencies need in a modern, experimental, and explicitly iterative manner”); accord Jason Bloomberg, *Digital Influencer Jez Humble: DevOps For Big Hairy Enterprises*, FORBES (Mar. 31, 2016, 9:54 AM), <https://www.forbes.com/sites/jasonbloomberg/2016/03/31/digital-influencer-jez-humble-devops-for-big-hairy-enterprises/#2f603f054a21> [<https://perma.cc/8B3G-C8R4>] (describing 18F as “a skunkworks-like team of designers, developers, and product specialists”).

235. *About*, 18F, <https://18f.gsa.gov/about/#our-team> [<https://perma.cc/2DT4-K9FQ>] (last visited Oct. 3, 2019).

236. *Guides*, 18F, <https://18f.gsa.gov/guides/> [<https://perma.cc/N2SD-QY7B>] (last visited Oct. 3, 2019).

237. U.S. WEB DESIGN SYS., <https://designsystem.digital.gov/> [<https://perma.cc/2V2Y-K2HN>] (last visited Oct. 3, 2019).

238. Mollie Ruskin et al., *Introducing the U.S. Web Design Standards*, 18F (Sept. 28, 2015), <https://18f.gsa.gov/2015/09/28/web-design-standards/> [<https://perma.cc/X33R-P43C>].

239. U.S. WEB DESIGN SYS., *supra* note 237.

technology projects and detailed guidance documents is appealing in the context of algorithmic systems.

Coordinated expert input through shared resources in the vein of the USWDS should frame agency decision making around algorithmic systems, as well as consulting services to provide hands-on assistance to agencies on an as-needed basis. The New York City Task Force was an effort to develop a methodology for eliciting relevant information from developers and operators of systems through targeted questions. It builds on research in the FAT\* community identifying specific information necessary to understand appropriate uses of and potential biases in systems.<sup>240</sup> The recently introduced Algorithmic Accountability Act of 2019,<sup>241</sup> while aimed at the private sector rather than the public sector, takes a similar approach by requiring automated decision system, data protection impact assessments, and regulations promulgated by the Federal Trade Commission to guide such assessments. A similar mix of standardized questions, guidance, and localized assessment is found in the privacy area, where Congress required administrative agencies to conduct privacy impact assessments<sup>242</sup> but authorized the Office of Management and Budget (OMB) to issue detailed guidance for agency implementation.<sup>243</sup>

The mix of expertise offered by 18F and USDS, and their track record of providing effective guidance and leadership, provides the most compelling

---

240. See Margaret Mitchell et al., *Model Cards for Model Reporting*, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 220 (2019) (proposing information about what the use context model was designed for, model performance benchmarked across different groups, and processes of validation and accompany models); see also Galen Harrison et al., *Towards Supporting and Documenting Algorithmic Fairness in the Data Science Workflow*, WORKSHOP ON TECH. & CONSUMER PROTECTION (May 23, 2019), <https://www.ieee-security.org/TC/SPW2019/ConPro/papers/harrison-conpro19.pdf> [https://perma.cc/VPZ9-B4UN] (proposing documentation and visualization of algorithms in data science processes); see generally Timnit Gebru et al., *Datasheets for Datasets*, ARXIV (Apr. 16, 2019), <https://arxiv.org/pdf/1803.09010.pdf> [https://perma.cc/82DT-G93P] (proposing information about attributes of datasets that should be documented and shared). These efforts resemble the efforts in the reproducible research community to provide information about data, code, computational steps, software environment, etc. See Victoria Stodden et al., *Enhancing Reproducibility for Computational Methods*, 354 SCIENCE 1240 (2016).

241. S. 1108, 116th Cong. (2019); H.R. 2231, 116th Cong. (2019) (providing the companion House bill).

242. E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2899, 2921 (2002) (requiring agencies to conduct a privacy impact assessment before “developing or procuring information technology that collects, maintains, or disseminates information that is in an identifiable form”).

243. See Memorandum from Joshua B. Bolten, Dir., Office of Mgmt. & Budget, on OMB Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002 to Heads of Exec. Dep’ts & Agencies (Sept. 26, 2003), [https://obamawhitehouse.archives.gov/omb/memoranda\\_m03-22/](https://obamawhitehouse.archives.gov/omb/memoranda_m03-22/) [https://perma.cc/39ZE-U2TC].

model to begin developing methods and tools for cross-agency efforts to identify and reason about embedded policies in algorithmic decision-making systems.

b) Models to Inform the Centralized Process

Existing legal frameworks addressing the use of predictive models in the area of credit and employment offer guidance regarding the possible content of centralized processes. For example, under the Equal Credit Opportunity Act (ECOA), in order to use age<sup>244</sup> as a predictive factor in granting credit, creditors must use an “empirically derived, demonstrably and statistically sound, credit scoring system.”<sup>245</sup> To meet these criteria, the system must be:

- (i) [b]ased on data that are derived from an empirical comparison of sample groups or the population of creditworthy and non-creditworthy applicants who applied for credit within a reasonable preceding period of time;
- (ii) [d]eveloped for the purpose of evaluating the creditworthiness of applicants with respect to the legitimate business interests of the creditor utilizing the system (including, but not limited to, minimizing bad debt losses and operating expenses in accordance with the creditor’s business judgment);
- (iii) [d]eveloped and validated using accepted statistical principles and methodology; and
- (iv) [p]eriodically revalidated by the use of appropriate statistical principles and methodology and adjusted as necessary to maintain predictive ability.<sup>246</sup>

While designed to regulate credit granting, these criteria, along with the research on aspects of models and data sets necessary to determine appropriate

---

244. While these requirements only apply to the use of credit scoring systems that use age as a predictive factor in granting credit, regulator comments suggest that in practice, statistical validation of credit scoring systems is a key tool that the Federal Trade Commission and Consumer Financial Protection Board use to assess compliance with ECOA and that regulated entities meet these guidelines to manage risk. *See Credit Scoring: Testimony Before the U.S. House Subcomm. on Financial Institutions & Consumer Credit*, 111th Cong. (Mar. 24, 2010), <https://www.federalreserve.gov/newsevents/testimony/braunstein20100323a.htm> [<https://perma.cc/ZU98-N33W>] (testimony of Sandra F. Braunstein, Dir., Div. of Consumer & Cmty. Affairs, Fed. Reserve).

245. *See* 12 C.F.R. § 1002.2(p) (defining empirically-derived and other credit scoring systems); 12 C.F.R. app. I § 1002.2(p) (2019) (“1.... The definition under §§ 1002.2(p)(1)(i) through (iv) sets the criteria that a credit system must meet in order to use age as a predictive factor.”).

246. 12 C.F.R. § 1002.2(p), n.217.

and fair uses, provide a useful starting point for thinking about data and models.

The Uniform Guidelines on Employee Selection Procedures (Uniform Guidelines)<sup>247</sup> issued in 1978 by the Equal Employment Opportunity Commission, provides another set of criteria that might be useful in the context of developing guidance on algorithmic systems. These guidelines are designed “to assist employers, labor organizations, employment agencies, and licensing and certification boards to comply with requirements of Federal law prohibiting employment practices which discriminate on grounds of race, color, religion, sex, and national origin.”<sup>248</sup> The Uniform Guidelines specify that selection procedures having an adverse impact on these protected characteristics will be prohibited:

The use of any selection procedure which has an adverse impact on the hiring, promotion, or other employment or membership opportunities of members of any race, sex, or ethnic group will be considered to be discriminatory and inconsistent with these guidelines, unless the procedure has been validated in accordance with these guidelines, or the provisions of section 6 below are satisfied.<sup>249</sup>

The guidelines thus provide a framework for determining the proper use of tests and other selection procedures. Moreover, while they do not require employers to conduct validity studies where no adverse impact results, they draw attention to the potential disparate impacts of selection procedures by providing guidance on how to construct and validate them—encouraging more thoughtful technical choices. For example, the Uniform Guidelines state that “where two or more selection procedures are available which serve the user’s legitimate interest . . . and which are substantially equally valid for a given purpose, the user should use the procedure which has been demonstrated to have the lesser adverse impact.”<sup>250</sup> While these guidelines already apply to the employment practices of agencies, they are equally applicable to the development of algorithmic systems more generally and could easily be adapted for such purposes. Other elements of the Uniform Guidelines are similarly relevant, such as the discussion of criterion-related, content, and construct validity studies to assess selection procedures.<sup>251</sup> In particular, the Uniform Guidelines clarify that the validity of a selection procedure can be validated by empirical data demonstrating that the selection procedure is

---

247. 29 C.F.R. § 1607 (2019).

248. 29 C.F.R. § 1607.1.

249. 29 C.F.R. § 1607.3.

250. 29 C.F.R. § 1607.3.B.

251. See, e.g., 29 C.F.R. § 1607.5.B.

predictive of or significantly correlated with important elements of job performance; by data showing that the content of the selection procedure is representative of important aspects of performance on the job; or through a construct validity study that consists of data showing that the procedure measures the degree to which candidates have identifiable characteristics which have been determined to be important in successful performance in the job.<sup>252</sup> In particular, the distinction between criterion-related validity and construct validity could be used to clarify different ways in which algorithmic systems could be reviewed. Although the Uniform Guidelines were issued thirty years ago, their focus on the *impact* of selection procedures—procedures used to predict how an applicant will do at a job—continues to make them relevant in the context of automated decisions today.

In conjunction with the growing set of guidance documents, toolkits, and other methods for analyzing algorithmic systems, these existing government regulations could provide a starting point for developing a suite of guidance documents and methods to standardize the questions and processes federal agencies use to assess and design algorithmic systems. With the availability of appropriate experts at 18F and USDS, this could provide a flexible, on-demand set of tools and personnel to fill the expertise gap plaguing federal agencies.

## B. INFUSING AGENCY DELIBERATION WITH POLITICAL VISIBILITY

### 1. *Impact Assessments: Bridging Technocracy and Democracy in Agency Deliberation*

Algorithmic impact assessments provide a critical tool to bridge the technocratic and democratic elements of deliberation.<sup>253</sup> Such tools not only enable and trigger agency deliberation about the technical aspects of system design, but surface the political implications of those choices, offering an important prerequisite for reasoned consideration of the policies they embed by agency staff, the public, and the political branches. They thus bridge the dual deliberation requirements of substantive expertise and political visibility.

More specifically, algorithmic impact assessments are boundary negotiation objects—objects used to “record, organize, explore and share ideas; introduce concepts and techniques; create alliances; create a venue for the exchange of information; augment brokering activities; and create shared understanding.”<sup>254</sup> If publicly disclosed, algorithmic impact assessments mediate between experts and the public, providing a common reference point,

---

252. *Id.*

253. See Reisman et al., *supra* note 22; Selbst, *supra* note 22.

254. Matthew J. Bietz & Charlotte P. Lee, *Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work*, PROC. 11TH EUR. CONF. ON COMPUTER SUPPORTED COOPERATIVE WORK 243, 247.

but importantly do not reflect a common understanding or consensus across these groups.<sup>255</sup> A good impact assessment provides a tool for exploring the problem space, helping the public collectively consider the points of policy within a machine learning system. They also allow the public and expert communities to in effect argue about the boundary between science and policy; in this way, they facilitate negotiation not just across the boundary, but about the location of it.

It is no accident that many of the “arbitrary and capricious” cases finding that agency action fails to reflect appropriate deliberation about relevant analytic techniques (models, assumptions, the use of data, and choices between false-negatives and false-positives) arise in the review of either cost-benefit or environmental-impact statements—two types of impact assessments well-enshrined in administrative law. Such tools are instrumental in “making bureaucracies think,”<sup>256</sup> and “take a hard look at the potential . . . consequences of their actions.”<sup>257</sup> While they do not mandate substantive outcomes, they force administrative agencies to turn their analytic capacity towards particular issues, and require explicit and publicly reviewable identification, recognition, and explanation of their choices about them.

Such requirements are especially important in surfacing issues in contexts with which government actors are unfamiliar, or where the issues addressed are “orthogonal to” or even “in tension with, an agency’s primary mission.”<sup>258</sup> Our research has demonstrated the particular importance of such impact assessments when agencies engage with the use of technology, given problems of technical illiteracy, design opacity, and the phenomena by which the

---

255. (2009) Privacy regulators and scholars generally advocate for publishing privacy impact assessments to support “contestation and public debate.” Jennifer Stoddart, *Auditing Privacy Impact Assessments: The Canadian Experience*, in PRIVACY IMPACT ASSESSMENT 453–54 (David Wright & Paul De Hert eds., 2012) (describing various regulators’ positions on publication). Unfortunately, proposed bills in the United States do not require publication. See The Algorithmic Accountability Act of 2019, S. 1108, 116th Cong. (2019); H.R. 2231, 116th Cong. (2019); see also Margot E. Kaminski & Andrew D. Selbst, *The Legislation That Targets the Racist Impacts of Tech*, N.Y. TIMES (May 7, 2019), <https://www.nytimes.com/2019/05/07/opinion/tech-racism-algorithms.html> [https://perma.cc/P6YV-43AV] (critiquing the lack of a publication requirement).

256. SERGE TAYLOR, MAKING BUREAUCRACIES THINK: THE ENVIRONMENTAL IMPACT STATEMENT STRATEGY OF ADMINISTRATIVE REFORM 251 (1984).

257. *The National Environmental Policy Act: A Study of Its Effectiveness After Twenty-Five Years*, COUNCIL ON ENVT'L QUALITY iii (Jan. 1997), <https://ceq.doe.gov/docs/ceq-publications/nepa25fn.pdf> [https://perma.cc/4KJH-4VK5] (discussing the National Environmental Policy Act’s “success” in making federal agencies take a “hard look” at the potential environmental consequences of their actions).

258. Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy Decisionmaking in Administrative Agencies*, 75 U. CHI. L. REV. 75, 83 (2008).

modality “hide[s] the political nature of its design,”<sup>259</sup> rendering policy implications invisible and making choices seem fixed, natural, and incontestable. In particular, the requirement imposed by the E-Government Act of 2002,<sup>260</sup> that agencies complete privacy impact assessments (PIAs) when developing or procuring information technology systems that include personally identifiable information, has forced agencies to address important privacy implications of systems intended to promote a range of public aims—implications that remained unnoticed and unaddressed (with important and expensive security consequences) where such requirements were not satisfied robustly.<sup>261</sup>

When agencies adopt technology, we have argued accordingly, the choices that impact legal rights must be addressed through the use of impact assessment tools.<sup>262</sup> On the one hand, those tools “create different frameworks and bring new considerations to bear in agency actions,” as well as bridge the gulf between the substantive domain expertise of agency staff and the frameworks and knowledge of outside experts.<sup>263</sup> On the other, they “facilitate participation by issue experts and by stakeholders who might otherwise be unaware of relevant risks and technological alternatives.”<sup>264</sup>

To have the desired ameliorative effect, experts must be involved in conducting impact assessments<sup>265</sup> and the outputs must be available to the public.<sup>266</sup> Current and proposed efforts to require government agencies to

---

259. Mulligan & Bamberger, *supra* note 12, at 778.

260. E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2899, 2921 (2002).

261. See Kenneth A. Bamberger & Deirdre K. Mulligan, *PIA Requirements and Privacy Decision-Making in US Government Agencies*, in *PRIVACY IMPACT ASSESSMENT* 225–50 (David Wright & Paul De Hert eds., 2012) (comparing DHS’s responsible adoption of RFID technology with that of the Department of State, which failed to discuss technical aspects of the program, alternative technologies, and risks); Bamberger & Mulligan, *supra* note 258 (discussing the same).

262. Mulligan & Bamberger, *supra* note 12, at 764–66, 780 (arguing that when agencies govern “by design,” they should use values-impact assessment tools, such as a “human rights impact assessment”).

263. *Id.* at 765.

264. *Id.*

265. Bamberger & Mulligan, *supra* note 258, at 104 (concluding that optimal use of privacy impact assessment turned on expert inter-disciplinary staff).

266. While different countries have taken different positions on whether privacy impact assessments or summaries ought to be shared with the public, there is a general preference for doing so due to the recognition that doing so can maintain public trust and confidence in systems and organizations. See David Wright & Paul De Hert, *Introduction to Privacy Impact Assessment*, in *PRIVACY IMPACT ASSESSMENT* 3–32 (David Wright & Paul De Hert eds., 2012); see also Elin Palm & Sven Ove Hansson, *The Case for Ethical Technology Assessment (eTA)*, 73 TECHNOLOGICAL FORECASTING & SOC. CHANGE 543, 547 (2006) (arguing for publication because “[i]t would be delusive to believe that technology developers are conscious of all the

conduct impacts of surveillance and algorithmic systems too often emphasize process while shorting or overlooking expertise.<sup>267</sup> There is an ongoing rise in chief privacy officers and other privacy staff, and more recently a move to redefine their roles to provide them greater latitude to address harms that may flow from data analysis and applications—reflected in new titles such as chief data governance officer or information steward.<sup>268</sup> More recently, growing concerns with AI—ranging from job displacement to bias to military applications—have ushered in a new set of professional staff with titles such as chief ethics officer.<sup>269</sup> More significantly, a major shift in privacy regulation in the European Union was to require data protection officers. The emphasis on professionals reflects the growing understanding that particular expertise is necessary to fully use tools such as impact assessments.

## 2. Other Political Visibility-Enhancing Processes

The impact assessments described above provide foundational tools for spanning boundaries between expert communities, policy makers, and the general public. By distilling the policy-relevant choices in technical design, impact assessments surface issues for political consideration.

However, the question of what aspects of a technical system are political is itself a value judgment, and who decides is itself a political matter. Despite the best efforts of experts,<sup>270</sup> the political and the technical defy clean

---

effects of their products[; i]n many cases, negative side effects come as a surprise to technology developers themselves”).

267. See OAKLAND, CAL., MUN. CODE §§ 9.64.010–9.64.070 (2018) (establishing requirement for impact assessments but not requiring new staff); accord SANTA CLARA CITY, CAL., ORDINANCE CODE § A40 (2016).

268. See, e.g., Molly Hulefeld, *What is a chief data ethics officer anyway*, IAPP PRIVACY ADVISORY (Nov. 27, 2018), <https://iapp.org/news/a/making-way-for-the-rise-of-the-chief-data-ethics-officer/> [https://perma.cc/Q39K-WDUV] (“Acxiom renamed its privacy program to become the[ ]data ethics, governance, protection and privacy program.”’’); *Mastercard Names JoAnn Stonier Chief Data Officer*, MASTERCARD (Feb. 8, 2018), <https://newsroom.mastercard.com/press-releases/mastercard-names-joann-stonier-chief-data-officer/> [https://perma.cc/BT43-QVX6] (announcing that JoAnn Stonier was moving from chief information governance and chief privacy officer to chief data officer, a new position designed to affirm the company’s commitment to data protection).

269. See, e.g., *Rise of The Chief Ethics Officer*, FORBES (Mar. 27, 2019), <https://www.forbes.com/sites/insights-intelai/2019/03/27/rise-of-the-chief-ethics-officer/#5e67f2ba5aba> [https://perma.cc/8P84-HFYR].

270. Jasianoff has documented how experts in science policy constantly attempted to demarcate the politics from the scientific in their practice in an effort to preserve claims of objectivity. See Sheila S. Jasianoff, *Contested Boundaries in Policy-Relevant Science*, 17 SOC. STUDS. SCI. 195, 199 (1987) (“To shore up their claims to cognitive authority, scientists have to impose their own boundaries between science and policy.”).

separation.<sup>271</sup> The complexity and density of technical and scientific matters can create barriers to broader participation in policy debates. This is surely true with respect to machine learning systems, where the complex interaction of design choices, data, and interfaces can produce clearly political outcomes yet be shielded from public scrutiny. Thus, agencies must further employ a broader set of processes to publicize system politics and elicit the public participation essential for legitimate administrative decision making.

The COMPAS debates provide a glimpse of the politics of expertise in action in machine learning systems. After journalists at ProPublica exposed the different false positive and false negative rates for black and white defendants, Northpointe, the developer of COMPAS, defended the system because it was “equally accurate for blacks and whites,” asserted its status as expert on the matter, and dismissed ProPublica’s analysis as wrong because they failed to “account the different base rates of recidivism for blacks and whites.”<sup>272</sup> Their rejoinder attempts to remove a legitimate political question—should models of fairness account for different base rates—from the public discussion by framing the question as one of whether to account for objective facts. Academics later pointed out that the differing perspectives on how to conceive of fairness espoused by Northpointe and ProPublica were mathematically incompatible, yet both defensible and possibly mutually desirable.<sup>273</sup> Here, as in many other instances, the technical and the political are entangled. Yet the important question about what metric should be used to support fair risk assessments appears to have escaped careful scrutiny by the public agency.

Fostering public engagement requires processes at a variety of levels, and times, to surface the politically salient questions latent in the design and use of machine learning systems. As an initial matter, public engagement would *both* be primed at the abstract level—i.e., the public would have a general

---

271. See generally STATES OF KNOWLEDGE: THE CO-PRODUCTION OF SCIENCE AND SOCIAL ORDER (Sheila Jasanoff ed., 2004) (presenting a collection of essays exploring the ways in which our methods of understanding and reasoning about the world and choices about how to live in the world are interdependent).

272. William Dieterich et al., *COMPAS Risk Scores: Demonstrating Accuracy Equity and Predictive Parity*, NORTHPOINTE 9 (July 8, 2016), [http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica\\_Commentary\\_Final\\_070616.pdf](http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf) [https://perma.cc/WQQ5-SEMC].

273. See Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear*, WASH. POST (Oct. 17, 2016), [https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm\\_term=.ce67b2c3fa95](https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.ce67b2c3fa95) [https://perma.cc/5TDR-E4MQ]; see also Jon Kleinberg et al., *Inherent Trade-Offs In The Fair Determination of Risk Scores*, ARXIV (Nov. 17, 2016), <https://arxiv.org/pdf/1609.05807.pdf> [https://perma.cc/MR4S-9W8W].

understanding of the important policy choices entangled with system design—and invited to participate in discourse with respect to specific systems.

a) Fostering Ongoing Public Engagement Through Agenda-Setting

Today, the politics of machine learning systems are an object of ongoing public scrutiny. The “black boxes” are being aired out, and centers of science policy have participated in opening up algorithmic systems to scrutiny. During the Obama Administration, the Office of Science and Technology Policy wrote several reports exploring the politics of the design and use of big data and artificial intelligence.<sup>274</sup> In particular, the reports detailed the biases that can result from training data and model design.<sup>275</sup> The sustained attention to the political issues embedded in technical systems was particularly important given President Obama’s efforts to strengthen “America’s role as the world’s engine of scientific discovery and technological innovation” and use technology to improve service delivery and governance across the public sector.<sup>276</sup> The Federal Trade Commission conducted workshops and reports that similarly focused attention on the policy implications of algorithmic systems, but with attention to its implications in the marketplace rather than the civic sector.<sup>277</sup> These reports focused agencies and the public on the politics

---

274. See, e.g., *Big Data: Seizing Opportunities, Preserving Values*, EXECUTIVE OFF. PRESIDENT (May 2014), <https://hsdl.org/?view&did=752636> [<https://perma.cc/AV9Y-L4FR>]; *Big Data and Differential Processing*, EXECUTIVE OFF. PRESIDENT (Feb. 2015), [https://obamawhitehouse.archives.gov/sites/default/files/whitehouse\\_files/docs/Big\\_Data\\_Report\\_Nonembargo\\_v2.pdf](https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf) [<https://perma.cc/Z9AL-48C7>]; *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, EXECUTIVE OFF. PRESIDENT (May 2016), [https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016\\_0504\\_data\\_discrimination.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf) [<https://perma.cc/VJ3S-HW3G>]; Ed Felten & Terah Lyons, *The Administration’s Report on the Future of Artificial Intelligence*, WHITE HOUSE BLOG (Oct. 12, 2016, 6:02 AM), <https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence> [<https://perma.cc/X4EQ-CFWT>]; see also John P. Holdren & Megan Smith, *Cabinet Exit Memo*, OFF. SCI. & TECH. POL’Y, EXECUTIVE OFF. PRESIDENT (Jan. 5, 2017), [https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp\\_exit\\_memo\\_final.pdf](https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_exit_memo_final.pdf) [<https://perma.cc/FN5W-7S6B>].

275. See *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, *supra* note 274.

276. Holdren & Smith, *supra* note 274, at 2 & n.2 (citing President Obama’s remarks on November 23, 2009 at the launch of his “Educate to Innovate” campaign for excellence in science, technology, engineering, and math education).

277. See *Big Data: A Tool for Inclusion Or Exclusion?*, FED. TRADE COMMISSION (Sept. 15, 2014), <https://ftc.gov/news-events/events-calendar/2014/09/big-data-tool-inclusion-or-exclusion> [<https://perma.cc/69PE-4ZGH>]; *Data Brokers: A Call for Transparency and Accountability*, FED. TRADE COMMISSION (May 2014), <https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf> [<https://perma.cc/6QSU-SD8A>].

of algorithmic choices, and sent a strong political signal to agencies that technological design ought to be scrutinized as policy.

Agenda-setting activities such as these White House reports, other Administration strategy documents that address the ethical, legal, and social aspects of artificial intelligence,<sup>278</sup> and studies and reports by organizations such as the National Academy of Science,<sup>279</sup> throw open the doors to ongoing public engagement through plain language, case studies, context, and explicit identification of the policy judgments that warrant public engagement along with technical expertise. They also encourage agencies to surface questions about specific technical system designs for public scrutiny.

#### b) Fostering Public Engagement on Specific Systems

Broader participation in the adoption and design of specific systems poses some practical challenges. As discussed in Part II, unlike regulations which agencies write themselves, the software code of machine learning systems is typically authored and owned by companies. Empirical work suggests that agencies routinely have little impact on the design of such systems, largely procuring them “off the shelf.” Where an agency desires to engage the public with the design of a technical system, contracts<sup>280</sup> and intellectual property<sup>281</sup> may present formidable obstacles. Testing—a tried and true method of exploring how a system performs for various populations or under different conditions (usability) as well as security properties—may be contractually limited.<sup>282</sup> Companies may even demand secrecy of training materials. While revealing source code may not be a necessary or desirable means of exposing the values embedded in systems, allowing regulators and experts of their choosing to examine, test, and tinker with systems is an initial prerequisite for surfacing values for public consideration.

---

278. See *The National Artificial Intelligence Research and Development Strategic Plan*, NAT'L SCI. & TECH. COUNCIL, at 8–40 (Oct. 2016), [https://nitrd.gov/PUBS/national\\_ai\\_rd\\_strategic\\_plan.pdf](https://nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf) [<https://perma.cc/JBV5-Z5HK>].

279. See, e.g., NAT'L ACADS. SCIS., ENG'G & MED., PROACTIVE POLICING: EFFECTS ON CRIME AND COMMUNITIES (David Weisburd & Malay K Majmundar eds., 2018), <https://doi.org/10.17226/24928> [<https://perma.cc/7A3H-USQV>] (discussing biases in statistical predictions).

280. See Joseph Lorenzo Hall, *Contractual Barriers to Transparency in Electronic Voting*, PROC. 2007 USENIX/ACCURATE ELECTRONIC VOTING TECH. WORKSHOP (2007), [https://www.usenix.org/legacy/event/evt07/tech/full\\_papers/hall/hall\\_html/jhall\\_evt07.html.html](https://www.usenix.org/legacy/event/evt07/tech/full_papers/hall/hall_html/jhall_evt07.html.html) [<https://perma.cc/GUY2-HY4X>] (discussing the use of contracts to limit public oversight).

281. See, e.g., Aaron Burstein et al., LEGAL ISSUES FACING ELECTION OFFICIALS IN AN ELECTRONIC-VOTING WORLD (Mar. 15, 2007), [https://www.law.berkeley.edu/files/Legal\\_Issues\\_Facing\\_ElectionOfficials.pdf](https://www.law.berkeley.edu/files/Legal_Issues_Facing_ElectionOfficials.pdf) [<https://perma.cc/7DKX-JPG8>]; Levine, *supra* note 114.

282. Hall, *supra* note 280.

As noted elsewhere, public participation is important during both design and deployment.<sup>283</sup> The U.S. Public Participation Playbook provides best practices for agencies to engage the public. While the strategies it has identified through collaboration with agency staff and citizen engagement experts are technology agnostic, they provide guidelines for and examples of successful citizen engagement. Of particular importance for machine learning is the Playbook's mindset of ongoing public engagement and feedback, its emphasis on identifying engagement strategies for specific stakeholder groups, and its emphasis on using prototypes and systems as well as more traditional means of eliciting feedback.

Combining the high-level airing of issues in the White House reports discussed above with participatory processes at the USDS and 18F could provide a powerful first step in a framework for public engagement with the politics of machine learning systems. The development of the Web Standards by USDS and 18F was done through an open process and produced publicly available resources for both use and interrogation. In the words of USDS and 18F, “we’re working in the open to create a resource that everyone can own and contribute to.”<sup>284</sup> Admittedly, this sort of openness is not conducive to broad participation by the lay public, but it does open up technical design for interrogation by experts who can support such participation and deliberation.

However, USDS and 18F go beyond such technocratic openness. For example, 18F played a key role in implementing the Digital Accountability and Transparency Act,<sup>285</sup> building a prototype of the proposed technical implementation to facilitate public feedback. In effect, through prototyping, 18F brought public participation into the agile software design process.<sup>286</sup> The prototyping used in this instance was one of many strategies for fostering and improving public participation in shaping government programs, including regulations, developed by 18F in conjunction with numerous government agencies.

Yet the broader public must be brought in, as well. The Federal Trade Commission’s practices used to develop policy around the privacy impact of technology use and design model suggest two important elements of agency process: publicity and engagement.<sup>287</sup> The Commission conducted and

---

283. See Mulligan & Bamberger, *supra* note 12, at 772 (arguing that “[t]he traditional sequential perspective of ‘policymaking’ (during which there is an opportunity for input) followed by ‘implementation’ is inconsistent with [regulation through] design”).

284. Ruskin, *supra* note 238.

285. Digital Accountability and Transparency Act of 2014, Pub. L. No. 113-101, 128 Stat. 1146.

286. See *Implementing A Government-wide Law*, 18F, <https://18f.gsa.gov/what-we-deliver/data-act/> [<https://perma.cc/WAC6-7FKF>] (last visited Oct. 9, 2019).

287. PRIVACY ON THE GROUND, *supra* note 228, at 189–90 (discussing FTC efforts).

publicized research on both child-directed and general audience websites, increasing the transparency of technology privacy behavior and its policy implications, spurring public attention and providing the basis for engagement. The participatory fora that followed empowered privacy advocates (indeed, many advocacy organizations were founded in response to this opportunity) who both provided input and served as a means for publicizing policy threats more broadly.

Meaningful public participation focused more directly on the adoption of specific systems requires similar scaffolding by agencies. Prototypes and simulations can be powerful means of publicizing technology choices by translating between mathematical formulations and policy choices. For example, the debate about how to model fairness in COMPAS (described above) may seem esoteric to members of the public, but a simulation that allowed individuals to play with various fairness metrics could powerfully expose them to the political implications. An example of such a tool is Google's What-if Tool. The What-if Tool is an interactive visual interface that allows individuals to probe machine learning models. The tool provides options to explore various algorithmic fairness constraints, compare counterfactuals, selectively add and remove data points, compare models on a single data set, among others.<sup>288</sup> Interactive exploratory tools provide opportunities for experiential learning by the public, building greater understanding of aspects of model design, and potentially framing questions about specific model design or data selection choices facing an agency.

The models of the Oakland Privacy Advisory Committee, the New York City Algorithmic Taskforce, and the Pennsylvania Sentencing Commission, moreover, use public events in which experts engage with impact assessments and systems before the public to further elaborate the policies that warrant attention. In effect, these models bolster the technological expertise of the public. Given that civil society organizations often have limited technical capacity, government provisioning of such experts to the public and stakeholders through expert committees may be an important component of the public participation infrastructure. Without agency enlistment of such expertise for the public, the public may miss risks and opportunities posed by machine learning systems and be unable to formulate appropriate and viable solutions.

### *3. Contestable Design*

To ensure that the informed agency engages with and deliberates about the policymaking that occurs through system deployment and use, government

---

288. WHAT-IF TOOL, <https://pair-code.github.io/what-if-tool/> [<https://perma.cc/HVJ6-HHWN>] (last visited Oct. 9, 2019).

machine learning systems must be designed to promote contestability.<sup>289</sup> That is, they must be designed to reveal their “thinking” and receive feedback from and collaborate with human users at runtime.<sup>290</sup> By fostering user engagement within the system, contestable systems use that engagement to iteratively identify and embed domain knowledge and contextual values as decision making becomes a collaborative effort in sociotechnical systems. Contestable systems thus provide a means for system logic to be overseen by, to learn from, and to be shaped by, the domain expertise and experience of human agency staff actually vested with administrative discretion. Contestability is thus one way “to enable responsibility in knowing,” to use Judith Simon’s phrase, as the production of knowledge is spread across humans and machines.<sup>291</sup>

As legal scholars, we are reluctant to argue for generalized design choices for agency machine learning systems. However, existing literature suggests that particular design choices support contestability by fostering user understanding of models and outputs, collaborative construction of systems, dynamic feedback and control, and within-model challenges to system outputs.<sup>292</sup> These approaches suggest the beginnings of a list of design

---

289. See Tad Hirsch et al., *Designing Contestability: Interaction Design, Machine Learning, and Mental Health*, 2017 PROC. ACM DESIGNING INTERACTIVE SYSS. CONF. 95, 98 (2017) (identifying “contestability” as a new design concern—focused on anticipating and designing for the ways technology can reshape knowledge production and power and describing three lower-level design principles to support contestability: 1) improving accuracy through phased and iterative deployment with expert users in environments that encourage feedback; 2) heightening legibility through mechanisms that “unpack aggregate measures” and “trac[e] system predictions all the way down” so that “users can follow, and if necessary, contest the reasoning behind each prediction”; and relatedly, in an effort to identify and vigilantly prevent system misuse and implicit bias, 3) identifying “aggregate effects” that may imperil vulnerable users through mechanisms that allow “users to ask questions and record disagreements with system behavior” and engage the system in self-monitoring).

290. Contestable design aligns with Mireille Hildebrandt’s call for “‘agonistic machine learning,’ i.e., demanding that companies or governments that base decisions on machine learning must explore and enable alternative ways of datafying and modelling the same event, person or action.” Mireille Hildebrandt, *Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*, 20 THEORETICAL INQUIRIES L. 83, 106 (2019). This “highlights the contestability at the level of the inner workings of machine learning systems.” *Id.* at 118. Also, this “responds to the need to call out the ethical and political implications of *who decides task T, performance metric P and experience E, and to investigate how this is done, taking into account which (and whose) concerns are at stake*.” *Id.* at 110.

291. Judith Simon, *Distributed Epistemic Responsibility in a Hyperconnected Era*, in THE ONLIFE MANIFESTO: BEING HUMAN IN A HYPERCONNECTED ERA 145–59, 146 (Luciano Floridi ed., 2015), [https://doi.org/10.1007/978-3-319-04093-6\\_17](https://doi.org/10.1007/978-3-319-04093-6_17) [<https://perma.cc/2BQV-DBEG>].

292. For insights on how contestable systems advance individual understanding, see, for example, Motahhare Eslami, *Understanding and Designing Around Users’ Interaction with Hidden Algorithms in Sociotechnical Systems*, 2017 CSCW ’17 COMPANION 57 (describing several studies finding that seamful designs, which expose algorithmic reasoning to users, facilitated understanding, improved user engagement, and in some instances altered user behavior);

requirements that permit contestability, which frames an agenda for research and experimentation about governmental systems looking forward.

a) Design Should Expose Built-in Values

At its core, contestability requires systems to be designed in a way that exposes value-laden features and parameters. Most simply, this visibility can prompt awareness, reflection, and feedback by agency decision makers relying on these systems. Reminding agency staff about the original choice of model and training data can prompt future deliberation about their appropriateness, particularly where a system consistently produces outcomes that agency staff perceive as incorrect in certain classes of cases. Decision makers relying on recidivism risk systems, such as Northpointe's COMPAS, must be made fully aware of how gender and other protected attributes are used, to both educate them about the policy choices underlying their decision supports and to encourage feedback from ground-level staff about their strengths and limitations. Brauneis and Goodman underscore the ways that contestability can foster ongoing deliberation about policy, describing the ways that the Hunchlab predictive policing software

allows each community to set weights for the relative seriousness of each type of crime—how much more important is it to stop a murder than a burglary? It also allows tailored weights for patrol

---

Motahhare Eslami et al., "*I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds*, 2015 PROC. 33RD ANN. ACM CONF. ON HUM. FACTORS COMPUTING SYSS. 153 (describing the lasting effects on how users engage with Facebook to influence the News Feed algorithm after an experimental design intervention that visualized its curatorial voice); Susan Joslyn & Jared LeClerc, *Decisions with Uncertainty: The Glass Half Full*, 22 CURRENT DIRECTIONS PSYCHOL. SCI. 308 (2013) (describing how displaying uncertainty in weather predictions can lead to more optimal decision making and trust in a forecast: transparency about probabilistic nature of prediction engenders trust even when predictions are wrong); Simone Stumpf et al., *Toward Harnessing User Feedback For Machine Learning*, 2007 PROC. 12TH INT'L CONF. ON INTELLIGENT USER INTERFACES 82; Simone Stumpf et al., *Interacting Meaningfully with Machine-Learning Systems: Three Experiments*, 67 INT'L J. HUMAN-COMPUTER STUD. 639 (2009) (noting that explainable systems can improve user understanding and use of system and enable users to provide deep and useful feedback to improve algorithms); Travis Moor et al., *End-User Debugging of Machine-Learned Programs: Toward Principles for Baring the Logic*, SEMANTIC SCHOLAR (2009), [https://pdfs.semanticscholar.org/9a4f/a9f116668927575113d6d2e8572e39925650.pdf?\\_ga=2.190583149.793734117.1566762743-966572889.1566762743&\\_gac=1.216591076.1566762743.CjwKCAjw44jrBR AHEiwAZ9igKEyD8gmFQSr6pK9M8nWVxjHzez0odTFCvRLu4dS7rXCjRvZGxRwsmBoCVXwQAvD\\_BwE](https://pdfs.semanticscholar.org/9a4f/a9f116668927575113d6d2e8572e39925650.pdf?_ga=2.190583149.793734117.1566762743-966572889.1566762743&_gac=1.216591076.1566762743.CjwKCAjw44jrBR AHEiwAZ9igKEyD8gmFQSr6pK9M8nWVxjHzez0odTFCvRLu4dS7rXCjRvZGxRwsmBoCVXwQAvD_BwE) [<https://perma.cc/E7BA-MXS3>] (noting that salient explanations helped users adjust their mental models); Saleema Amershi et al., *Power to The People: The Role of Humans in Interactive Machine Learning*, 35 AI MAG. 105 (2014) (providing an overview of interactive machine learning research with case studies, and discussing value of interactive machine learning approaches for the machine learning community as well as users).

efficacy—[for example,] indoor crimes are less likely to be deterred by increased police presence.<sup>293</sup>

Moreover, design decision visibility can enable active participation by system users in consequential decisions about their configuration or use. For example, the confidence thresholds that determine an agency's preference for false positives or false negatives when using Amazon's Rekognition Web Service should be prominently exposed to staff and easily configurable. In these ways, contestable design expands front-line staff's knowledge of the policies and values embedded in the machine learning systems they use, while offering them opportunities to configure and interrogate them at run time. Such designs help agency staff learn about machine learning systems as the systems learn about agency staff. Machine learning systems, then, should be designed to allow users to both make key decisions about values-significant parameters and understand their significance. This requires moving away from defaults for these parameters and toward contestable systems that require engagement during setup and use.

Contestable design, moreover, is a prerequisite for continuous feedback from domain experts. Rather than traditional forms of contesting automated decisions—“out-of-band” processes (those external to the regular operation of the system itself, like exception handling and appeals)—contestable design brings argumentation, and therefore opportunities for learning and recalibration within the system itself. Such continuous within-system learning is appropriate as “our models are, and will continue to be, fallible” and is particularly important in areas where the risks of “‘getting it wrong’ can be quite high.”<sup>294</sup> Active, critical, real-time engagement with the reasoning of machine learning systems’ inputs, outputs, and models reduces the risk that machine learning systems will replace the logical and ethical frameworks that comprise expert agency judgment, and respond to risks posed by fallible models—regardless of whether fallibility is a product of design choices, shifts in policy, or inattention to gaps between phenomena and the representations we choose to capture them.

Contestability is a more active and dynamic principle than explanation<sup>295</sup> to guide design; for these reasons it breeds user engagement. Where the passivity of “explainable” algorithmic systems imagines engagement, reflection, and questioning as out-of-band activities—via exception handling, appeals processes, etc.—contestable systems foster active, critical engagement

---

293. Brauneis & Goodman, *supra* note 6, at 150.

294. Hirsch et al., *supra* note 289, at 97.

295. See *id.* at 98; see also Daniel Klutzz et al., *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*, in AFTER THE DIGITAL TORNADO: NETWORKS, ALGORITHMS, HUMANITY (Kevin Werbach ed., forthcoming 2020).

within the system. Explanations are also typically static, insofar as they are focused on conveying a single message; contestability, in contrast, aims to support interactive exploration of, and in some instances tinkering with, machine logic.

b) Design Should Trigger Human Engagement

Designing for contestability further requires the design of human-system interaction in a way that promotes an active and ongoing role for agency decision makers by overcoming the over-reliance on, and deference to, decision-support systems arising from automation bias—“the use of automation as a heuristic replacement for vigilant information seeking and processing”<sup>296</sup> and automation complacency—insufficient attention and monitoring of automation outputs.

It is certain that without thoughtful interventions, agency staff will be less attentive and engaged with the decisions supported by machine learning systems. Where the goal of introducing machine learning systems is to achieve a hybrid production of knowledge that builds on the strengths of human and machine ways of knowing—rather than to fully displace human decision making—chosen designs and policies must keep humans “in the game.”

Research has established that particular design choices determine the extent to which expert judgment continues to be exercised in system-supported decision making in ways that skew policy regarding, for example, a preference for false-negatives over false-positives. The extent to which users substantively review a machine learning system output depends largely on whether the system signals that the result is anomalous or normal. For example, research has shown that radiologists scrutinize mammography films identified by decision-support systems as positive for cancer, catching many of the false positives that the system produced.<sup>297</sup> But films identified as normal rarely receive such inspection, allowing nearly all false negatives to evade detection. The perception that the system was over-inclusive—supported by the experience of identifying false positives—contributed to a belief in the infrequency of false negatives, when in fact the system systematically failed to identify certain images of cancer.

---

296. Kathleen L. Mosier & Linda J. Skitka, *Automation Use and Automation Bias*, PROC. HUM. FACTORS & ERGONOMICS SOC’Y ANN. MEETING 344 (1999).

297. Anrey A. Povyakalo et al., *How to Discriminate Between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography*, 33 MED. DECISION MAKING 98 (2013); see Goddard, *supra* note 182; see also Adrian Bussone et al., *The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems*, INT’L CONF. ON HEALTHCARE INFORMATICS 160 (2015) (discussing research findings on automation bias and self-reliance).

Similarly, in ongoing research with legal professionals using machine learning systems to aid in document review for discovery,<sup>298</sup> we have found that human review focuses on documents identified as relevant and thus appropriate for disclosure, according less attention to those the machine learning system designates as non-responsive. In both instances, perceptions of the machine's performance interact with experts' risk models to yield different levels of engagement with different outputs of the same underlying system. Yet, ideally we would want agency experts to be attentive to misidentifications or classifications (depending upon the task) produced by false positives and negatives—although, depending upon the use of the system, attention to one or the other might be directed by policy.

Iterative re-delegation of tasks and communicating about uncertainty are both design strategies that have been found successful as to maintain human skill and sense of responsibility and maintain attention to how machines are executing delegated tasks. For example, adaptive allocation,<sup>299</sup> in which an automated task is reallocated back to a human for periods of time, has been found to reduce automation complacency and improve subsequent attention to tasks by humans,<sup>300</sup> reduce human distraction, and promote a sense of responsibility.<sup>301</sup> Communicating the confidence that a system has in the conclusions it offered has been found to foster feelings of user responsibility<sup>302</sup> and, where coupled with a feedback loop, improve decisions in the moment and over time.<sup>303</sup>

---

298. Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 861 (2019).

299. See RAJA PARASURAMAN ET AL., THEORY AND DESIGN OF ADAPTIVE AUTOMATION IN AVIATION SYSTEMS, REP. NO. NAWCADWAR-92033-60 (July 17, 1992), <https://apps.dtic.mil/dtic/tr/fulltext/u2/a254595.pdf> [<https://perma.cc/YBJ6-UAXQ>].

300. See Raja Parasuraman, *Effects of Adaptive Task Allocation on Monitoring of Automated Systems*, 38 HUM. FACTORS 665 (1996).

301. *Id.*

302. Matthew Kay et al., *When (ish) Is My Bus?: User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems*, PROC. 2016 CONF. ON HUM. FACTORS COMPUTING SYSS. 5092 (2016) (finding that some bus riders felt that providing information about the uncertainty of an arrival time would make them more responsible for actions that led them to miss the bus: “you’re more likely to be unhappy than if you missed the bus and can just blame the app”).

303. Communicating the uncertainty related with the probabilistic nature of machine learning systems to users improves decision making, and if it is coupled with feedback mechanism, it can leverage human knowledge to increase accuracy of model over time. See Rocio Garcia-Retamero & Edward T. Cokely, *Communicating Health Risks with Visual Aids*, 22 CURRENT DIRECTIONS IN PSYCHOL. SCI. 392–39 (2013) (displaying a grid of pictograms, each representing a patient success or fatality improved the accuracy of people’s risk assessment); see also Ulrich Hoffrage & Gerd Gigerenzer, *Using Natural Frequencies to Improve Diagnostic Inferences*, 73 ACAD. MED. 538 (1998) (noting that more medical experts could accurately estimate the positive predictive value (precision) of a test when presented with discrete counts

c) Design Should Promote Contestation About Social and Political Values

Finally, for a number of reasons, contestability is of heightened importance where functions delegated to machine learning systems are deeply connected to social and political values such as fairness—values that are often ambiguous and contested.

First, given the numerous competing definitions of fairness,<sup>304</sup> there may well be multiple conflicting views on which definition should apply to a given function or in a given context; and when a definition of fairness is chosen, it may be susceptible to different formalizations.<sup>305</sup> We have more generally argued that, because of the strength and durability of design decisions, policymakers should be cautious about “baking” choices about human and public rights into technology systems and should steer such determinations to the least fixed point of technical intervention—permitting ongoing debate about such value choices and “designing technological hooks that permit different value choices in different contexts.”<sup>306</sup> The discussion of the debates around what fairness required in COMPAS offers a clear example of the need to maintain visibility about these contested design elements.

Second, protecting and respecting values such as fairness and privacy may often hinge on *process*. Fairness and privacy, for example, are often *defined*—at least in part—by access to procedures that afford individuals meaningful participation, and to information (rules and data used to make decisions about them). These procedural aspects of values can be supported through contestable system design that minimizes the automation and opacity of those decisions and ensures that human judgment will be brought to bear in the ongoing shaping of, and in the assessment of the products of, decision-support systems.<sup>307</sup> Contestability keeps agency experts in control of these values

---

or outcomes); Stumpf et al., *Toward Harnessing*, *supra* note 292, at 82 (noting that “user feedback has the potential to significantly improve machine learning systems”).

304. See Deirdre K. Mulligan et al., *This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology*, PROC. 2019 ACM ON HUMAN-COMPUTER INTERACTION 3, 119 (2019) (defining the following terms: formal equality (blind to all other variables)—to each person an equal share; need-based—to each person according to individual need; effort-based—to each person according to individual effort; social contribution—to each person according to societal contribution; and merit-based—to each person according to merit).

305. For example, one could operationalize a given fairness definition around groups—seeing demographic parity or equal positive predictive values, or equal negative predictive values, or equal false positive or false negative rates, or accuracy equity—or one could operationalize it through an individual fairness metric, such as equal thresholds or devising a similarity metric. For a discussion, see *id.*

306. Mulligan & Bamberger, *supra* note 12, at 750.

307. See Min Kyung Lee & Su Baykal, *Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division*, PROC. 2017 ACM CONF.

questions, even while specific tasks and functions are handed off to machine learning systems. They can allow agency experts to revisit and tune machine-learning decisions to context-specific information that influences the perceived or actual fairness of a system. Contestable design responds to the demand of philosophers and ethicists that systems be designed to respond to diverse contexts ruled by different moral frameworks<sup>308</sup> and to support collaborative development of ethical requirements.<sup>309</sup>

## V. CONCLUSION

In 1967 John Culkin, interpreting one of Marshal McLuhan's five postulates, offered the now-famous line: "We shape our tools and thereafter they shape us."<sup>310</sup> The fear of abdicating important policy decisions to the control of tools, rather than autonomously wielding them in service of the public interest, lies at the heart of current concerns with the governmental adoption of machine learning systems. Such abdication strikes at the heart of administrative legitimacy and good governance, and suggests that machine learning systems are tools to be procured at our peril.

In many instances they are more texts than tools, and we suggest engaging them accordingly. They have their limits, their biases, and their blind spots. We should question and bicker with them, but we should also learn from and teach them. We should not blindly defer to them or bring them into our deliberations without knowing their backstories, working assumptions, and theories.

Through policy choices and design, we can build purposeful tools that are aligned with values chosen based on reason, expertise, transparency, and robust and ongoing deliberation and oversight. We can bind these new systems through carefully constructed "policy knots" that align them with the requirements of administrative law through policy, practice, and design.

---

ON COMPUTER SUPPORTED COOPERATIVE WORK & SOC. COMPUTING 1035 (2017) (presenting discussion with others, and ability to interrogate systems logic can improve perceptions of fairness).

308. See Batya Friedman & Helen Nissenbaum, *Software Agents and User Autonomy*, PROC. 1ST INT'L CONF. ON AUTONOMOUS AGENTS 466 (1997).

309. See Matteo Turilli, *Ethical Protocols Design*, 9 ETHICS & INFO. TECH. 49 (2007).

310. John M. Culkin, *A Schoolman's Guide to Marshall McLuhan*, SATURDAY REV. 51, 70–72 (Mar. 18, 1967) (offering a "barously brief" distillation of Marshal McLuhan's writings, John M. Culkin expanded on one of McLuhan's five postulates, *Art Imitates Life*).

858

BERKELEY TECHNOLOGY LAW JOURNAL

[Vol. 34:781

# ACQUIRING ETHICAL AI

*David S. Rubenstein\**

## Abstract

Artificial intelligence (AI) is transforming how the federal government operates. Under the right conditions, AI systems can solve complex problems, reduce administrative burdens, improve human decisions, and optimize resources. Under the wrong conditions, AI systems can lead to widespread discrimination, invasions of privacy, and the erosion of democratic norms. A burgeoning literature has emerged to square algorithmic governance with the precepts of constitutional and administrative law. Federal procurement law, however, remains a dangerous blind spot in the reformist agenda. This Article pivots into that neglected space and emerges with comprehensive framework for acquiring ethical AI. Toward that end, the Article makes three main contributions. First, it provides an original account that yokes the ambitions of algorithmic governance, the imperative of ethical AI, and the levers of procurement law. Second, this Article argues that the procurement system is uniquely situated to check and enable algorithmic governance in ways that other legal frameworks miss. Third, the Article prescribes a set of concrete regulatory reforms to instantiate ethical AI throughout the procurement process: from acquisition planning to market solicitation, bid evaluation, source selection, and contract performance. Procurement law will not solve all the challenges of algorithmic governance. Just as surely, those challenges cannot be solved without procurement law.

---

\* James R. Ahrens Chair in Constitutional Law and Director, Robert Dole Center for Law and Government, Washburn University School of Law. The author thanks Daniel Ho, Aziz Huq, Martin Murillo, and Nicole Petroff for their very helpful comments and suggestions on earlier drafts. The author also thanks Kaitlyn Bull, Ande Davis, Penny Fell, Barbara Ginsberg, Creighton Miller, Chris Smith, and Zach Smith for excellent research assistance, as well as the *Florida Law Review* for careful and constructive editing.

INTRODUCTION .....	749
I. AI TODAY.....	758
A. <i>Machine Learning Systems</i> .....	759
B. <i>Humans in AI Systems</i> .....	761
1. Problem Formulation and System Objectives.....	762
2. Data Selection and Preparation .....	763
3. Model Training.....	763
4. Model Testing and Evaluation .....	764
5. Model Selection and System Configuration .....	765
II. TOWARD ETHICAL ALGORITHMIC GOVERNANCE .....	768
A. <i>Good (Algorithmic) Governance</i> .....	768
B. <i>Ethical Challenges</i> .....	771
1. Safety.....	772
2. Fairness .....	773
3. Transparency .....	778
4. Accountability .....	781
C. <i>The Rise of Ethical AI</i> .....	782
1. Ethical AI in Industry.....	783
2. Ethical AI in Government .....	785
III. FROM PRINCIPLES TO PRACTICE .....	787
A. <i>The Gap Between Ethical AI Principles and Practice</i> ..	787
1. Industry Challenges.....	788
2. Government Challenges .....	793
B. <i>The Gap Between Ethical AI and Algorithmic Governance</i> .....	796
IV. OPERATIONALIZING ETHICAL AI THROUGH PROCUREMENT LAW .....	797
A. <i>Acquisition Planning: AI Risk Assessments</i> .....	799
B. <i>Market Solicitations: Calling for Ethical AI</i> .....	804
C. <i>Evaluation and Source Selection: Requiring Ethical AI</i> .....	806
1. Evaluation Criteria .....	807
2. Responsibility Determination.....	810
D. <i>Contract Performance: Pathways and Pitfalls</i> .....	813
1. COTS AI Solutions .....	813
2. Customized AI Solutions .....	814
CONCLUSION.....	819

## INTRODUCTION

Artificial intelligence (AI) is transforming how the federal government operates.<sup>1</sup> For example, the Department of Justice uses AI in law enforcement;<sup>2</sup> the Social Security Administration uses AI for adjudicatory functions;<sup>3</sup> the Department of Homeland Security uses AI to regulate immigration;<sup>4</sup> the Internal Revenue Service uses AI to detect tax fraud;<sup>5</sup> the Department of Veterans Affairs uses AI to deliver health services;<sup>6</sup> the Pentagon uses AI to augment its military and intelligence capabilities;<sup>7</sup> the General Services Administration uses AI to streamline

---

1. See generally DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY, MARIONO-FLORENTINO CUÉLLAR, ADMIN. CONF. OF THE U.S., GOV'T BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (2020) [hereinafter ACUS REPORT], <https://www.cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf> [<https://perma.cc/AU59-KGAG>] (providing a rich account of this transformation along with its implications for regulatory practice and administrative law).

2. See *Letter to Attorney General Barr RE: The Use of the PATTERN Risk Assessment in Prioritizing Release in Response to the COVID-19 Pandemic*, LEADERSHIP CONF. ON CIVIL & HUM. RTS. (Apr. 3, 2020), <https://civilrights.org/resource/letter-to-attorney-general-barr-re-the-use-of-the-pattern-risk-assessment-in-prioritizing-release-in-response-to-the-covid-19-pandemic/> [<https://perma.cc/M349-RZYM>] (discussing and critiquing the Department's use of an AI criminal risk-assessment tool); James Vincent, *FBI Used Facial Recognition to Identify a Capitol Rioter from His Girlfriend's Instagram Posts*, VERGE (Apr. 21, 2021), <https://www.theverge.com/2021/4/21/22395323/fbi-facial-recognition-us-capital-riots-tracked-down-suspect> [<https://perma.cc/RZW9-MJ86>].

3. See ACUS REPORT, *supra* note 1, at 38–40.

4. See Aaron Boyd, *CBP Is Upgrading to a New Facial Recognition Algorithm in March*, NEXTGOV (Feb. 7, 2020), <https://www.nextgov.com/emerging-tech/2020/02/cbp-upgrading-new-facial-recognition-algorithm-march/162959/> [<https://perma.cc/RH9L-6MUT>]; Kate Evans & Robert Koulish, *Manipulating Risk: Immigration Detention Through Automation*, 24 LEWIS & CLARK L. REV. 789, 793 (2020).

5. TREASURY INSPECTOR GEN. FOR TAX ADMIN., U.S. DEP'T OF TREASURY, REFERENCE NO. 2017-20-080, THE RETURN REVIEW PROGRAM INCREASES FRAUD DETECTION; HOWEVER, FULL RETIREMENT OF THE ELECTRONIC FRAUD DETECTION SYSTEM WILL BE DELAYED 4 (2017), <https://www.treasury.gov/tigta/auditreports/2017reports/201720080fr.pdf> [<https://perma.cc/LXW4-PP78>] (noting “machine learning algorithms” for generating fraud risk scores).

6. See Anagha Srikanth, *How the VA Is Using Artificial Intelligence to Improve Veterans' Mental Health*, THE HILL (Sept. 8, 2020), <https://thehill.com/changing-america/well-being/mental-health/515536-how-the-va-is-using-artificial-intelligence-to> [<https://perma.cc/DH6U-WGUH>]; Am. Homefront Project, *VA Embraces Artificial Intelligence To Improve Veterans' Health Care*, CPR NEWS (Feb. 19, 2020), <https://www.cpr.org/2020/02/19/va-embraces-artificial-intelligence-to-improve-veterans-health-care/> [<https://perma.cc/LYT2-8BW8>].

7. See DEP'T OF DEF., SUMMARY OF THE 2018 DEPARTMENT OF DEFENSE ARTIFICIAL INTELLIGENCE STRATEGY 15 (2018), <https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF> [<https://perma.cc/WK76-SLZD>]; Memorandum from Kathleen H. Hicks, Deputy Sec'y of Def., to Senior Pentagon Leadership et al. 1 (May 26, 2021) (discussing the Department of Defense's “embrace[]” of AI and the need to “adopt responsible behavior, processes, and outcomes in a manner that reflects the Department's commitment to its ethical principles”).

business operations;<sup>8</sup> the Department of Health and Human Services uses AI for regulatory analysis and reform.<sup>9</sup> The list goes on<sup>10</sup> and is projected to grow exponentially with the government's digital transformation.<sup>11</sup>

This presages a new era of "algorithmic governance," in which federal responsibilities and functions will increasingly migrate from humans to machines.<sup>12</sup> As emergent technology, AI bears the burden of proof—and it's far from an easy case. Under the right conditions, AI systems can solve complex problems, reduce administrative burdens, improve human decisions, optimize government resources, and drive agency missions.<sup>13</sup> Under the wrong conditions, however, AI systems pose serious threats to civil rights and democratic norms.<sup>14</sup> Already, AI systems have wrongly deprived individuals of unemployment and medical benefits,<sup>15</sup> wrongly identified individuals for criminal arrest,<sup>16</sup>

---

8. See, e.g., Dave Nyczepir, *GSA Leads Rise in Automation Projects Governmentwide*, FEDSCOOP (May 11, 2021), <https://www.fedscoop.com/automation-projects-rise-gsa/> [<https://perma.cc/PVJ5-UQXG>] (reporting that The General Services Administration has four fully operational AI projects, with many more in the works).

9. See Press Release, U.S. Dep't of Health & Hum. Servs., HHS Launches First-of-Its-Kind Regulatory Clean-Up Initiative Utilizing AI (Nov. 17, 2020), <https://www.pressreleasepoint.com/hhs-launches-first-its-kind-regulatory-clean-initiative-utilizing-ai> [<https://perma.cc/SEEN-LZGW>].

10. See, e.g., ACUS REPORT, *supra* note 1, at 25–29 (discussing a "suite of algorithmic tools" used by the SEC "to identify violators of federal security laws"); *id.* at 46–52 (discussing AI use cases by the U.S. Patent and Trademark Office); Jori Heckman, *USPS Gets Ahead of Missing Packages with AI Edge*, FED. NEWS NETWORK (May 6, 2021), <https://federalnewsnetwork.com/artificial-intelligence/2021/05/usps-rolls-out-edge-ai-tools-at-195-sites-to-track-down-missing-packages-faster/> [<https://perma.cc/DQS8-RPZ4>] (reporting that the U.S. Postal Service is using AI to examine and categorize packages it receives).

11. See KEVIN DROEGEMEIER ET AL., OFF. OF MGMT. & BUDGET, FEDERAL DATA STRATEGY 2020 ACTION PLAN 11 (2020), <https://strategy.data.gov/assets/docs/2020-federal-data-strategy-action-plan.pdf> [<https://perma.cc/8KCC-VXV6>] (discussing the government's information technology modernization efforts and data initiatives to support the adoption of AI technologies); NAT'L SEC. COMM'N ON A.I., FINAL REPORT 4 (2021) [hereinafter NSCAI FINAL REPORT] ("We envision hundreds of billions in federal spending [for AI technologies] in the coming years.").

12. See Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 636 (2017) ("[I]mportant decisions that were historically made by people are now made by computer systems."); WILL HURD & ROBIN KELLY, *RISE OF THE MACHINES: ARTIFICIAL INTELLIGENCE AND ITS GROWING IMPACT ON U.S. POLICY* 7–8 (2018), <https://www.hSDL.org/?view&did=816362> [<https://perma.cc/9KYV-59BX>] (reporting concerns about job loss due to AI-driven automation).

13. See *infra* Section II.A (discussing the putative benefits of algorithmic governance).

14. See *infra* Section II.B (discussing the causes and manifestations of AI risk).

15. See, e.g., Stephanie Wykstra & Undark, *It Was Supposed to Detect Fraud. It Wrongfully Accused Thousands Instead. How Michigan's Attempt to Automate Its Unemployment System Went Terribly Wrong*, ATLANTIC (June 7, 2020), <https://www.theatlantic.com/technology/archive/2020/06/michigan-unemployment-fraud-automation/612721/> [<https://perma.cc/8KT7-DWLS>].

16. See Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. TIMES (Aug. 3, 2020), <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html> [<https://perma.cc/8KT7-DWLS>].

and wrongly denied access to government food programs.<sup>17</sup> More generally, the government's adoption of AI can amplify racial and gender biases, encroach on civil liberties, and create barriers to government transparency and accountability.<sup>18</sup>

From this starting position, the question is not whether algorithmic governance will be for better or worse, but rather *whose lives* will be benefitted and burdened, in *which ways*, under *what rules* of engagement, and *who should decide*.<sup>19</sup> Amidst the swirling uncertainty one thing is clear: the outcomes will by heavily influenced by the technology industry. Despite pockets of excellence, the government's demand for AI systems far exceeds its in-house capacity to design, develop, field, and monitor this technology at scale.<sup>20</sup> Accordingly, many if not most of the tools and operational support for algorithmic governance will be procured by contract from technology firms.<sup>21</sup>

The “outsourcing of algorithmic governance” brings many affordances; chief among them is the government’s ability to capitalize

---

cc/PP5D-7E9X]; *see also* Complaint ¶¶ 39–49, *Williams v. City of Detroit*, No. 2:21-cv-10827 (E.D. Mich. filed Apr. 13, 2021), [https://www.aclumich.org/sites/default/files/field\\_documents/001\\_complaint\\_1.pdf](https://www.aclumich.org/sites/default/files/field_documents/001_complaint_1.pdf) [https://perma.cc/7D4L-F77Z] (collecting data regarding racial bias and misidentification in facial-recognition systems used by police).

17. *See, e.g.*, Florangela Davila, *USDA Disqualifies Three Somalian Markets from Accepting Federal Food Stamps*, SEATTLE TIMES (Apr. 10, 2002), <https://archive.seattletimes.com/archive/?date=20020410&slug=somalis10m> [https://perma.cc/MR7K-9JZX] (describing how the U.S. Department of Agriculture’s monitoring system for suspicious transactions denied three Somalian-owned markets from accepting food stamps based on “unusual, irregular, and/or inexplicable” activity at each store).

18. *See infra* Section II.B; *see also* Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103, 129 (2018) (explaining that AI’s ability to scale government processes and decision-making magnifies any error or bias); CATHY O’NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 29–31 (2016) (providing a trenchant account of how AI systems disproportionately harm marginalized and vulnerable populations); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 180–88 (2017) (arguing that government decision-making systems create a “digital poorhouse”); Solon Barocas & Andrew D. Selbst, *Big Data’s Disparate Impact*, 104 CALIF. L. REV. 671, 677 (2016) (“[D]ata mining holds the potential to unduly discount members of legally protected classes and to place them at systematic relative disadvantage.”).

19. *See* KATE CRAWFORD, ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE 8 (2021) (providing a similar framing of these sociopolitical issues).

20. NAT’L SEC. COMM’N ON A.I., SECOND QUARTER RECOMMENDATIONS 34 (2020) (“[T]here is a severe shortage of AI knowledge in [the Department of Defense] and other parts of government.’ . . . Current initiatives are helpful, but only work around the edges, and are insufficient to meet the government’s needs.” (quoting NAT’L SEC. COMM’N ON A.I., INTERIM REPORT 35 (2019)); cf. ACUS REPORT, *supra* note 1, at 18, 89 (finding that approximately half of AI applications covered in the study were developed in-house by federal agency personnel).

21. *See infra* notes 256–61 and accompanying text (discussing the asymmetry between the government’s demand for AI tools and its capacity to develop, implement, and monitor, these tools in-house).

on the industry's innovation, institutional know-how, and high-skilled workforce.<sup>22</sup> But these alliances subsume worrisome reliances. Currently, AI technologies are virtually unregulated in the private market.<sup>23</sup> Unless and until that changes, federal agencies will be acquiring unregulated technology for use in high-stakes government contexts.<sup>24</sup> Moreover, AI systems are embedded with value-laden tradeoffs between what is technically feasible, socially acceptable, economically viable, and legally permissible.<sup>25</sup> Without proper planning and precaution, the government may acquire AI with embedded policies from private actors whose financial motivations and legal sensitivities may not align with the government or the people it serves.<sup>26</sup>

Simply put, acquiring AI is not business as usual. The technology is inherently risky, regardless of who develops and deploys it.<sup>27</sup> But the government's risk profile requires special attention.<sup>28</sup> Most notably,

---

22. See David S. Rubenstein, *The Outsourcing of Algorithmic Governance*, YALE J. ON REGUL.: NOTICE & COMMENT (Jan. 19, 2021), <https://www.yalejreg.com/nc/the-outsourcing-of-algorithmic-governance-by-david-s-rubenstein/> [https://perma.cc/9G VS-7PXT]; NSCAI FINAL REPORT, *supra* note 11, at 24 ("The government lags behind the commercial state of the art in most AI categories, including basic business automation. It suffers from technical deficits that range from digital workforce shortages to inadequate acquisition policies, insufficient network architecture, and weak data practices.").

23. To some extent, the regulation of AI development and deployment may be regulated under existing federal law. See Memorandum from Russell T. Vought, Dir., Off. of Mgmt. & Budget, to the Heads of Exec. Dep'ts & Agencies 2 (Nov. 17, 2020), <https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf> [https://perma.cc/HW3J-FHQ5]; Elisa Jillson, *Aiming for Truth, Fairness, and Equity in Your Company's Use of AI*, FED. TRADE COMM'N: BUS. BLOG (Apr. 19, 2021, 9:43 AM), <https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai> [https://perma.cc/2Q6Z-DQV5] (describing potential FTC enforcement in cases of AI-derived discrimination). Still, it is widely acknowledged that existing federal laws and regulations are inadequate to address the novel ways that AI systems are developed and operationalized. See *infra* notes 192–97 and accompanying text (discussing dramatic uptick in legislative proposals to regulate AI). This regulatory void is becoming increasingly difficult to justify, given the ubiquity and social significance of AI in the areas of finance, education, manufacturing, labor and employment, transportation, recreation, journalism, medicine, insurance, agriculture, energy, and countless more.

24. See DAVID S. RUBENSTEIN, GREAT DEMOCRACY INITIATIVE, FEDERAL PROCUREMENT OF ARTIFICIAL INTELLIGENCE: PERILS AND POSSIBILITIES 4 (2020).

25. *Id.*

26. *Id.*

27. Sean McGregor, *When AI Systems Fail: Introducing the AI Incident Database*, P'SHIP ON A.I.: BLOG (Nov. 18, 2020), <https://www.partnershiponai.org/aiincident> database [https://perma.cc/9ZMH-RH7S] ("Failures of [AI] systems pose serious risks to life and wellbeing, but even well-intentioned intelligent system developers fail to imagine what can go wrong when their systems are deployed in the real world.").

28. Cf. Daniel Guttman, *Public Purpose and Private Service: The Twentieth Century Culture of Contracting Out and the Evolving Law of Diffused Sovereignty*, 52 ADMIN. L. REV. 859, 862 (2000) (explaining that "in practice, two different sets of regulations have come to govern those doing the basic work of government"—those that apply to federal officials, on the one hand, and to federal contractors, on the other).

government action is subject to constitutional and administrative law requirements, whereas private action is not.<sup>29</sup> Further, the polity generally expects the government to serve the public interest in safe, fair, transparent, and accountable ways. But these norms of good governance pose major challenges for machine learning AI systems, which are agnostic to democratic values, and often opaque, fickle, and brittle.<sup>30</sup>

A rich literature has emerged to address the challenges of algorithmic governance. Generally speaking, the reformist agenda is keyed to how law, technology, or both, can be configured in ways that are normatively desirable and operationally feasible.<sup>31</sup> To date, most of this legal scholarship has concentrated on constitutional due process,<sup>32</sup> equal protection,<sup>33</sup> free speech and assembly,<sup>34</sup> and criminal rights.<sup>35</sup> Scholars

---

29. See *id.*; Lillian Bevier & John Harrison, *The State Action Principle and Its Critics*, 96 VA. L. REV. 1767, 1786 (2010) (“Constitutional rules are almost all addressed to the government.”). For an incisive treatment of the constitutional state action doctrine as applied to government AI vendors, see Kate Crawford & Jason Schultz, *AI Systems as State Actors*, 119 COLUM. L. REV. 1941, 1971–72 (2019) (arguing that courts should adopt a version of the state action doctrine to apply to vendors who supply AI systems for government decision-making).

30. See *infra* Section I.A (discussing machine learning AI systems), Section II.B (discussing a suite of sociotechnical challenges associated with machine learning AI systems).

31. See Ryan Calo & Danielle K. Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. 797, 835 (2021) (canvassing the “ongoing project that responds to automation’s disruption of rights and values through a combination of legal and technical reforms”).

32. See, e.g., Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1281–88 (2008) (discussing how government use of automated systems in governmental administrative proceedings raises due process concerns and prescribing reforms); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 121–28 (2014) (same).

33. See, e.g., Aziz Z. Huq, *Constitutional Rights in the Machine Learning State*, 105 CORNELL L. REV. 1875, 1917–27 (2020); Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 193 (2017) (“[A] simple prohibition on the use of protected characteristics such as race and sex in an automated decision process is easy to implement, but would do little to prevent biased outcomes.”); Barocas & Selbst, *supra* note 18, at 677 (discussing the specific technical issues that give rise to models whose use in decision-making may have a disproportionately adverse impact on protected classes).

34. See, e.g., Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM L. REV. 1265, 1273, 1295–306 (2020) (exploring the “procedural and substantive conflicts between proprietary [algorithmic] decision-making on the one hand and government transparency obligations under the First Amendment and [Freedom of Information Act] on the other”); see also Woodrow Hartzog & Evan Selinger, *Facial Recognition Is the Perfect Tool for Oppression*, MEDIUM (Aug. 2, 2018), <https://medium.com/s/story/facial-recognition-is-the-perfect-tool-for-oppression-bc2a08f0fe66> [<https://perma.cc/7FEU-HQEJ>]; Sigal Samuel, *Activists Want Congress to Ban Facial Recognition. So They Scanned Lawmakers’ Faces.*, VOX (Nov. 15, 2019, 10:10 AM), <https://www.vox.com/future-perfect/2019/11/15/20965325/facial-recognition-ban-congress-activism> [<https://perma.cc/R8JP-BY38>].

35. See, e.g., Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 331–32 (2015); Michael L. Rich, *Machine Learning, Automated*

have also begun the necessary work of squaring algorithmic governance with separation of powers doctrine<sup>36</sup> and precepts of administrative law.<sup>37</sup>

Federal procurement law, however, remains a dangerous blind spot in the reformist agenda. It is no novelty to observe, as others have, that the government's market dependencies and information asymmetries exacerbate the challenges of algorithmic governance.<sup>38</sup> And a handful of scholars have urged contracting officials, at all levels of government, to protect and promote public interests when acquiring AI from private vendors.<sup>39</sup> Yet the regulatory hooks and incentive structures required to meet these challenges remain woefully undertheorized, unspecified, and unutilized. This Article pivots into that neglected space and emerges with a comprehensive framework for "acquiring ethical AI." No less than other areas of law, federal procurement law will need retrofitting to

---

*Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871, 878–79 (2016); Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1350–53, 1397 (2018).

36. See, e.g., Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1154, 1176–84 (2017) (discussing how federal agency use of machine learning AI systems could potentially implicate the constitutional nondelegation doctrine); Mariano-Florentino Cuéllar, *Cyberdelegation and the Administrative State*, in ADMINISTRATIVE LAW FROM THE INSIDE OUT 134, 134, 156–57 (Nicholas R. Parrillo ed., 2017) (discussing how agency use of AI relates to delegation principles).

37. See ACUS REPORT, *supra* note 1; Coglianese & Lehr, *supra* note 36; Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 6 (2019); David Engstrom & Daniel Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. ON REG. 800 (2020); Deirdre K. Mulligan & Kenneth A. Bamberger, *Procurement as Policy: Administrative Process for Machine Learning*, 34 BERKELEY TECH. L.J. 773, 782 (2019); Wendy Wagner & Martin Murillo, *Is the Administrative State Ready for Big Data?*, in DATA & DEMOCRACY (Apr. 30, 2021), <https://s3.amazonaws.com/kfai-documents/documents/684b5fd17e/4.29.2021-Wagner-and-Murillo.pdf> [<https://perma.cc/U6VN-8B9T>]. Danielle Citron's seminal account of the "automated administrative state" argued that "[a]utomation jeopardizes the due process safeguards owed individuals and destroys the twentieth-century assumption that policymaking will be channeled through participatory procedures that significantly reduce the risk that an arbitrary rule will be adopted." Citron, *supra* note 32, at 1281. These concerns, aired more than a decade ago, have only intensified because machine learning AI systems are generally more complex and less transparent than the technologies that Citron interrogated. See Calo & Citron, *supra* note 31, at 818 ("In the decade since the publication of *Technological Due Process*, governments have doubled down on automation despite its widening problems.").

38. See Brauneis & Goodman, *supra* note 18, at 152–63 (spotlighting the transparency deficits that accrue when state and local government adopt AI systems developed by third parties); Mulligan & Bamberger, *supra* note 37, at 782 (explaining how a "procurement mindset" can forfeit the government's responsibility to make important design choices with public input); see also ACUS REPORT, *supra* note 1, at 88–90 (outlining some pros and cons of the government's insourcing and outsourcing for AI solutions); MONA SLOANE ET AL., AI AND PROCUREMENT PRIMER 3 (Summer 2021) (observing that "existing public procurement processes and standards are in urgent need of revision and innovation").

39. See, e.g., Brauneis & Goodman, *supra* note 18, at 164; Cary Coglianese & Erik Lampmann, *Contracting for Algorithmic Accountability*, 6 ADMIN. L. REV. ACCORD 175, 180 (2021).

regularize and legitimize algorithmic governance. Toward those ends, this Article makes three major contributions.

First, it provides an original account that yokes the ambitions of algorithmic governance, the principles of ethical AI, and the levers of procurement law. Broadly conceived, ethical AI envisages a cluster of principles relating to safety, fairness, transparency, accountability, privacy, and human well-being.<sup>40</sup> Hardly an abstract concern, ethical AI is a global imperative backed by the United States,<sup>41</sup> G20,<sup>42</sup> and all the leading technology firms (Amazon, Google, Facebook, Microsoft, and IBM, just to name a few).<sup>43</sup> Make no mistake, the institutional motivations fueling the ethical AI movement are pluralistic and opportunistic.<sup>44</sup> Yet the convergence of public and private interests around core ethical AI principles is what matters most for this discussion.<sup>45</sup> For the government and industry alike, AI innovation is a complex ambition that mediates technical capability and human values. Awful AI does not sell—politically or commercially.<sup>46</sup> Once these sociopolitical dynamics are accounted for, the normative case for acquiring ethical AI is also pragmatic.

---

40. See *infra* Sections II.B–C; Anna Jobin et al., *The Global Landscape of AI Ethics Guidelines*, 1 NATURE MACH. INTEL. 389, 390 fig.1 (2019), <https://www.nature.com/articles/s42256-019-0088-2.pdf> [<https://perma.cc/68YX-NM8Z>] (surveying 84 distinct ethical AI frameworks and finding that they have largely converged around a core set of concepts and principles, including safety, fairness, accountability, transparency, and privacy).

41. See, e.g., Exec. Order No. 13,960, 85 Fed. Reg. 78,939, 78,940–41 (Dec. 8, 2020); OFF. OF SCI. & TECH. POL’Y, EXEC. OFF. OF THE PRESIDENT, AMERICAN ARTIFICIAL INTELLIGENCE INITIATIVE: YEAR ONE ANNUAL REPORT, at i (2020) (“In a time of global power competition, our leadership in AI has never been more of an imperative.”); ORG. FOR ECON. COOP. & DEV. ARTIFICIAL INTELLIGENCE IN SOCIETY 16–17 (2019) [hereinafter OECD] [www.oecd-ilibrary.org/docserver/eedfee77-en.pdf](http://www.oecd-ilibrary.org/docserver/eedfee77-en.pdf) [<https://perma.cc/CJF5-JNAM>] (providing a comprehensive survey of the many ways that AI is projected to transform social structures and power dynamics across markets and borders).

42. See G20 MINISTERIAL STATEMENT ON TRADE AND DIGITAL ECONOMY app. at 11–14 (2019), <https://www.mofa.go.jp/files/000486596.pdf> [<https://perma.cc/MJ9A-5U82>].

43. See *infra* Section II.C (discussing proliferation of ethical AI throughout the public and private sectors); see also Alex Hern, “Partnership on AI” Formed by Google, Facebook, Amazon, IBM, and Microsoft, GUARDIAN (Sept. 28, 2016, 5:00 PM), <https://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms> [<https://perma.cc/74S6-EH9E>].

44. See *infra* notes 199–05 and accompanying text.

45. See *infra* Section II.C (discussing political and market demand for ethical AI).

46. See *infra* Section II.C; David Dao et al., *Awful AI*, GITHUB, <https://github.com/david-dao/awful-ai> [<https://perma.cc/PTU9-P92Q>] (providing a “curated list to track current scary usages of AI—hoping to raise awareness to its misuses in society” (emphasis omitted)); *Artificial Intelligence Incident Database*, <https://incidentdatabase.ai> [<https://perma.cc/2AXQ-BBA7>] (providing a systematized collection of incidents where intelligent systems have caused safety, fairness, or other real-world problems, for the express purpose of “learn[ing] from [AI’s] failings”).

*Second*, this Article argues that the procurement system is uniquely suited to both check and enable algorithmic governance. Currently, the government is procuring AI systems that may be inoperable, either because the technology is untrustworthy or unlawful in application. The inscrutability of acquired AI systems, for example, might violate constitutional or statutory requirements for government transparency and accountability.<sup>47</sup> Even if those thresholds are met, the inputs and outputs of AI systems may violate anti-discrimination norms, privacy laws, and domain-specific legal constraints.<sup>48</sup> Litigation will no doubt surface these legal tensions.<sup>49</sup> Indeed, the AI docket is already littered with cautionary cases.<sup>50</sup> Yet many of these governance challenges can be addressed

---

47. See, e.g., Citron, *supra* note 32, at 1281–88 (discussing how agency use of automated systems raises due process concerns); Mulligan & Bamberger, *supra* note 37, at 782 (“[T]he policy choices embedded in system design fail the prohibition against arbitrary and capricious agency actions absent a reasoned decision-making process that enlists the expertise necessary for reasoned deliberation, provides justifications for such choices, makes visible the political choices being made, and permits iterative human oversight and input.”). But cf. Cary Coglianese, *Using Machine Learning to Improve the U.S. Government*, REGUL. REV. (Aug. 12, 2019), <https://www.theregreview.org/2019/08/12/coglianese-using-machine-learning-to-improve-us-government/> [<https://perma.cc/Z9PL-YR4W>] (arguing that “with proper planning and implementation, the federal government’s use of algorithms, even for highly consequential purposes, should not face insuperable or even significant legal barriers under any prevailing administrative law doctrines”).

48. See Huq, *supra* note 33, at 1881 (explaining that machine learning AI technology “places pressure on the formulation of due process, equality, and privacy interests in subtly different ways”).

49. For comprehensive surveys of AI-related litigation and trends, see *Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems*, AI NOW INST. 5 (2018), <https://ainowinstitute.org/litigatingalgorithms.pdf> [[https://perma.cc/L584-83ZQhttps://ainowinstitute.org/litigatingalgorithms-2019-us.pdf](https://perma.cc/L584-83ZQ) [<https://perma.cc/X9H9-B3W5>]; see also Coglianese & Lampmann, *supra* note 39, at 177 (“Without question, agencies that choose to use AI tools need to be mindful of the possibility that their choices could later come under not just the spotlight of media attention but also the scrutiny of judicial review.”).

50. See, e.g., *Barry v. Lyon*, 834 F.3d 706, 710–11 (6th Cir. 2016) (holding that Michigan’s public benefits system erroneously terminated food assistance benefits of more than 20,000 individuals based on crude data matching algorithm in violation of due process guarantees); *Cahoo v. SAS Analytics Inc.*, 912 F.3d 887, 892 (6th Cir. 2019) (noting that defendants “designed, created, and implemented” allegedly flawed software that erroneously terminated unemployment benefits of thousands of Michigan residents without adequate notice); *Ark. Dep’t of Hum. Servs. v. Ledgerwood*, 530 S.W.3d 336, 338–40 (Ark. 2017) (affirming a temporary restraining order against an unlawful switch to computer algorithm that reduced the attendant-care services for multiple patients with severe illnesses by an average of 43%); *Hous. Fed’n of Tchrs., Local 2415 v. Hous. Indep. Sch. Dist.*, 251 F. Supp. 3d 1168, 1171–74 (S.D. Tex. 2017) (finding due process violation because teachers had no way to replicate and challenge their algorithmic scores); *Latif v. Holder*, 28 F. Supp. 3d 1134, 1161–62 (D. Or. 2014) (ordering the Federal Bureau of Investigation to “fashion new procedures” for its no-fly list policy and to “provide plaintiffs with

through the procurement system in ways that are more efficient, effective, and prior to harm.<sup>51</sup>

Third, this Article offers a set of pragmatic recommendations to operationalize ethical AI throughout the procurement process: from acquisition planning to market solicitation, bid evaluation, source selection, and contract performance.<sup>52</sup> By centering ethical AI across the procurement lifecycle, agency officials and vendors will be incented to think more holistically—and competitively—about the AI tools passing through the acquisition gateway for government use. Less directly, though equally important, the government’s purchasing power and virtue signaling can spur technological innovation and galvanize public trust in AI technologies inside and outside of government.<sup>53</sup> Thus conceived and constructed, the acquisition gateway is more than a marketplace: it is a policymaking space for mediating the possibilities and perils of modern AI systems.

The Article proceeds as follows. Part I offers a primer on machine learning AI systems and spotlights a range of human value judgments embedded in the technology. Part II expounds upon AI’s sociotechnical challenges and the political economies of ethical AI. Part III homes in on the residual gaps between ethical AI principles and practice, the implications of those gaps for algorithmic governance, and the limitations of existing laws and technologies to bridge the gulf. Part IV explicates how the federal procurement system can help—indeed, why it must. Procurement law will not solve all the challenges of algorithmic governance. Just as surely, those challenges cannot be solved without procurement law.

---

the requisite due process . . . without jeopardizing national security”). *But cf.* *State v. Loomis*, 881 N.W.2d 749, 753 (Wis. 2016) (holding that using an AI risk-assessment tool to aid judges with sentencing decisions “does not violate a defendant’s right to due process”).

51. *See generally* Part IV (offering a set of recommendations to promote AI safety, fairness, transparency, and accountability).

52. *See infra* Sections IV.A–B.

53. For a similar claim about the potential for positive externalities, see Coglianese & Lampmann, *supra* note 39, at 181 (“[T]he expectations that governments insist upon in their procurement contracts can help set the bar for algorithmic accountability throughout the economy, promoting the diffusion of norms about responsible AI across both the public and private sectors.”).

## I. AI TODAY

Despite the hype—or perhaps because of it—“artificial intelligence” has no “universally accepted definition.”<sup>54</sup> The lexical rifts are largely attributable to the field’s evolution and multi-disciplinarity, which spans computer science, mathematics, psychology, sociology, neuroscience, philosophy, linguistics, and intersects with countless more.<sup>55</sup> AI’s lexical dissensus also reflects clashing ideologies. As Kate Crawford explains, “[e]ach way of defining artificial intelligence is doing work, setting a frame for how it will be measured, valued, and governed.”<sup>56</sup> Sensitive to these concerns, and without normative pretense, this Article employs the term AI to capture a range of computer-based technologies that make predictions, classifications, recommendations, and automated decisions.<sup>57</sup>

Only a decade ago, AI was a fringe subject of academic study with sparse real-world applications. Rather abruptly, however, AI has emerged from research labs to disrupt every major market and facet of society.<sup>58</sup>

---

54. U.S. GOV’T ACCOUNTABILITY OFF., GAO-18-142SP 15 (2018) (observing that “[t]here is no single universally accepted definition of AI, but rather differing definitions and taxonomies”); *see also* Forrest E. Morgan et al., *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*, RAND CORP. 8–9 & 9 n.4 (2020), [www.rand.org/pubs/research\\_reports/RR3139-1.html](http://www.rand.org/pubs/research_reports/RR3139-1.html) [<https://perma.cc/4XUX-7PQZ>] (explaining the definitional challenges, and noting that “[i]t was striking how averse the experts we interviewed were to providing definitions of artificial intelligence”).

55. *See* U.S. GOV’T ACCOUNTABILITY OFF., GAO-18-142SP, *supra* note 54, at 15.

56. CRAWFORD, *supra* note 19, at 7.

57. One provision of U.S. law broadly defines AI to include the following:

- (1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.
- (2) An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
- (3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
- (4) A set of techniques, including machine learning, that is designed to approximate a cognitive task.
- (5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision making, and acting.

10 U.S.C. § 2358.

58. *See* U.S. GOV’T ACCOUNTABILITY OFF., GAO-21-519SP, ARTIFICIAL INTELLIGENCE: AN ACCOUNTABILITY FRAMEWORK FOR FEDERAL AGENCIES AND OTHER ENTITIES 5 (2021) (noting that AI applications “rang[e] from medical diagnostics and precision agriculture, to advanced manufacturing and autonomous transportation, to national security and defense”); OECD, *supra* note 41, at 16–17 (providing a comprehensive survey of the many ways that AI is projected to transform social structures and power dynamics across markets and borders).

AI's ubiquity is masked by its obscurity. Yet, increasingly and impalpably, the technology is embedded or connected to smartphones and drones, cars and cattle, workstations and police stations, classrooms and war rooms, energy grids and traffic grids, social platforms and financial platforms, medical systems and surveillance systems.<sup>59</sup> As such, the technology is radically changing how nations, institutions, and individuals interact, experience, and perceive the world.<sup>60</sup>

### A. Machine Learning Systems

AI's ascendancy and dissemination over the past decade owes to the conflation of several developments: the availability of exponentially more data and computing power; the democratization of the internet of things;<sup>61</sup> and breakthroughs in "machine learning" technologies.<sup>62</sup> Unlike traditional computer algorithms that require manual coding, machine learning algorithms learn and improve from exposure to large amounts of data.<sup>63</sup> A full exposition of machine learning is beyond this Article's remit. But a basic understanding of the technology will be important for the discussion ahead. The challenges of algorithmic governance, and this Article's procurement prescriptions, are anchored to how machine learning systems are designed, developed, and deployed.

Stripped to its essentials, machine learning is a statistical technique that learns from data to make classifications or predictions for new data inputs.<sup>64</sup> For example, if the objective of an AI system is to detect and

---

59. See NSCAI FINAL REPORT, *supra* note 11, at 33.

60. *Id.* ("Americans have not yet grappled with just how profoundly the [AI] revolution will impact our economy, national security, and welfare."); Eleonore Pauwels, *The New Geopolitics of Artificial Intelligence*, WORLD ECON. F. (Oct. 15, 2018), <https://www.weforum.org/agenda/2018/10/artificial-intelligence-ai-new-geopolitics-un> [<https://perma.cc/6JBY-X4NB>] ("The multilateral system urgently needs to help build a new social contract to ensure that . . . [AI] is deployed safely and aligned with the ethical needs of a globalizing world."); *see also supra* notes 58–59 and accompanying text.

61. "The Internet of Things, or IoT, refers to the billions of physical devices around the world that are now connected to the internet, all collecting and sharing data." Steve Ranger, *What Is the IoT? Everything You Need to Know About the Internet of Things Right Now*, ZDNET (Feb. 3, 2020, 6:45 AM), <https://www.zdnet.com/article/what-is-the-internet-of-things-everything-you-need-to-know-about-the-iot-right-now> [<https://perma.cc/J2FD-B29T>].

62. See NSCAI FINAL REPORT, *supra* note 11, at 20–21; *see also* Mireille Hildebrandt, *Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*, 20 THEORETICAL INQUIRIES L. 83, 84–85 (2019).

63. See U.S. GOV'T ACCOUNTABILITY OFF., GAO-21-519SP, *supra* note 58, at 14 (distinguishing "first-wave" and "second-wave" AI technologies on this basis); *see also* David Lehr & Paul Ohm, *Playing with Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC DAVIS L. REV. 653, 678 (2017).

64. See IBM Cloud Educ., *Machine Learning*, IBM (July 15, 2020), <https://www.ibm.com/cloud/learn/machine-learning> [<https://perma.cc/X7J9-GWW7>]; Sendhil

diagnose cancer, a computer can be fed many thousands of labeled images of malign and benign tumors, learn to distinguish the images based on patterns and correlations in the pixel data, and then generate an algorithmic model that can make diagnostic predictions.<sup>65</sup> Or, if the objective of an AI system is to predict violence within prison populations, a machine learning algorithm can be trained with data of prior incidents, inmate characteristics (e.g., age, education, weight, criminal history, gang affiliations), and other proxy variables deemed to correlate with prison violence. In turn, a warden or prison guard can deploy the trained AI model to predict incidents of inmate violence and take prophylactic action.<sup>66</sup>

The machine learning AI systems in circulation today are powerful but “narrow,” insofar as they can handle discrete tasks in bounded domains (like in the examples above).<sup>67</sup> Deep neural networks are a subset of sophisticated machine learning algorithms that have been trained to classify images, recognize faces, translate languages, predict human emotions, personalize online experiences, and much (much) more.<sup>68</sup> Some of the most technologically advanced AI systems coordinate or aggregate multiple algorithmic models. For example, autonomous vehicles utilize a range of algorithms to perform driving and navigation functions.<sup>69</sup> Still, at present, the most advanced AI systems do

---

Mullainathan & Jann Spiess, *Machine Learning: An Applied Econometric Approach*, 31 J. ECON. PERSP. 87, 88 (2017) (defining machine learning in terms of its capacity for “out of sample” prediction). There are several different approaches to machine learning. For a short and accessible overview of the main approaches, see Nicholas Diakopoulos, *Algorithmic Accountability: Journalistic Investigation of Computational Power Structures*, 3 DIGIT. JOURNALISM 398, 399 (2015). For extended treatments, see generally IAN H. WITTEN ET AL., *DATA MINING: PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES* (4th ed. 2017); KEVIN P. MURPHY, *MACHINE LEARNING* (2012); STUART J. RUSSELL & PETER NORVIG, *ARTIFICIAL INTELLIGENCE: A MODERN APPROACH* (4th ed. 2021).

65. Cf. Daoud Meerzaman, *Machine Learning and Computer Vision Offer a New Way of Looking at Cancer*, NAT’L CANCER INST. (Jan. 27, 2019), <https://datascience.cancer.gov/news-events/blog/machine-learning-and-computer-vision-offer-new-way-looking-cancer> [https://perma.cc/7NS2-7JYY].

66. See Stefanie Kanowitz, *How Predictive Analytics Keeps Corrections Staff, Inmates Safe*, GCN (Aug. 18, 2021), [https://gcn.com/articles/2021/08/18/predictive-analytics-corrections.aspx?s=gcnda\\_190821&oly\\_enc\\_id=&m=1](https://gcn.com/articles/2021/08/18/predictive-analytics-corrections.aspx?s=gcnda_190821&oly_enc_id=&m=1) [https://perma.cc/RW46-WHRP].

67. See U.S. GOV’T ACCOUNTABILITY OFF., GAO-18-142SP, *supra* note 54, at 15–16 (distinguishing between narrow and general AI).

68. See Bertrand Leong, *Rise of the Machines: Deep Learning, Machine Learning, AI, and Big Data*, SING. INST. MGMT. (2018), <https://m360.sim.edu.sg/article/Pages/Rise-of-the-Machines.aspx> [https://perma.cc/AFR4-5FFG]; Bernard Marr, *What Is Deep Learning AI? A Simple Guide With 8 Practical Examples*, BERNARD MARR & CO., <https://bernardmarr.com/what-is-deep-learning-ai-a-simple-guide-with-8-practical-examples/> [https://perma.cc/PXV4-EGZZ].

69. See Rilind Elezaj, *How AI Is Paving the Way for Autonomous Cars*, MACH. DESIGN (Oct. 17, 2019), <https://www.machinedesign.com/mechanical-motion-systems/article/21838234/>

not have common sense, causal reasoning,<sup>70</sup> or situational awareness “to determine the relevance of new ‘unknowns.’”<sup>71</sup> Moreover, unlike humans, AI systems cannot generalize or reliably transfer knowledge across experiential domains.<sup>72</sup> That type of “artificial general intelligence” does not (yet) exist and is beyond this Article’s scope.<sup>73</sup>

### B. Humans in AI Systems

The math and science behind AI systems can make them seem objective and neutral. However, a critical literature debunks that myth.<sup>74</sup> Beyond bits and bytes, AI systems are social artifacts that embed and project human choices, biases, and values.<sup>75</sup> These human inputs and outputs are hardly obvious. Precisely for that reason, it is crucial for policymakers and stakeholders to appreciate that value-laden choices of great social and legal consequence may be encased in AI systems prior to their adoption and deployment.

The discussion below provides a stylized account of just some of the human choices and tradeoffs that occur behind the AI curtain. This prelude supplies contextual mooring for the social, technical, and institutional challenges of algorithmic governance, which Parts II and III expound upon.

---

how-ai-is-paving-the-way-for-autonomous-cars [https://perma.cc/YH5L-JHH2]; U.S. GOV’T ACCOUNTABILITY OFF., GAO-18-142SP, *supra* note 54, at 65–68 (discussing the technology of automated vehicles).

70. See Brian Bergstein, *What AI Still Can’t Do*, TECH. REV. (Feb. 19, 2020), [www.technologyreview.com/2020/02/19/868178/what-ai-still-can-t-do/](http://www.technologyreview.com/2020/02/19/868178/what-ai-still-can-t-do/) [https://perma.cc/3QL9-JYL6] (describing the inability of AI to engage in causal reasoning).

71. David Leslie, *Understanding Artificial Intelligence Ethics and Safety*, ALAN TURING INST. 32 (2019), [https://www.turing.ac.uk/sites/default/files/2019-06/understanding\\_artificial\\_intelligence\\_ethics\\_and\\_safety.pdf](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf) [https://perma.cc/96KG-VB77].

72. See U.S. GOV’T ACCOUNTABILITY OFF., GAO 21-519SP, *supra* note 58, at 16.

73. See *id.* (noting the speculative nature of general AI).

74. See, e.g., CRAWFORD, *supra* note 19, at 8 (arguing that AI is neither artificial nor intelligent); MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN 5–7 (2019) (discussing the emerging science of ethical algorithm design to address the social implications of AI technologies); O’NEIL, *supra* note 18 (arguing that AI systems are “weapons of math destruction” that shape individual action and social dynamics); SAFIYA UMOJA NOBEL, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM 1–2 (2018) (“While we often think of . . . ‘big data’ and ‘algorithms’ as being benign, neutral, or objective, they are anything but.”).

75. For critical treatments of AI’s power dynamics, see *supra* note 74 (collecting sources), *see also* Sarah M. West et al., *Discriminating Systems: Gender, Race and Power in AI*. AI NOW INST. (Apr. 2019), <https://ainowinstitute.org/discriminatingsystems.pdf> [https://perma.cc/KCR9-CT6B] (describing technology industry’s diversity challenges and how those challenges manifest in AI applications and ideations).

## 1. Problem Formulation and System Objectives

The design of a machine learning system generally begins with the identification of a specific problem that AI may help to solve. To illustrate, suppose a federal agency has a huge backlog of applications for government benefits.<sup>76</sup> There are many ways to solve that problem. But if the plan involves AI, then agency officials must select a relatively specific objective for the AI system.<sup>77</sup> Here, assume that the agency wants an AI system to predict which applications in the backlog are likely to be granted, for the well-intended purpose of expediting the delivery of government benefits to qualifying applicants.

Despite good intentions, the agency's problem formulation and AI objective may have negative collateral effects on various stakeholders. For instance, the unflagged cases might be further delayed because resources are channeled to the flagged cases. Worse still, agency personnel may unfairly (and unwittingly) stigmatize the unflagged cases as unmeritorious because the AI system did not predict a win.<sup>78</sup> Meanwhile, agency personnel may be shuffled or shifted to accommodate the dual-track system. Even if institutionally justified, the disruptions may seed confusion or discontent in the ranks.

Already, this sketch illustration exposes the human underbelly of AI systems. The agency's formulation of the problem (backlog) and chosen objective for the AI system (to flag meritorious cases) were not preordained, much less conscribed to math or science. Furthermore, as discussed below, each of the foregoing choices will lead to countless more choices and path dependencies throughout the AI development lifecycle.

---

76. See Aaron Boyd, *VA Wants to Automate Digitization of Its 5-Mile-High Electronic Health Record Backlog*, NEXTGOV (July 9, 2020), <https://www.nextgov.com/it-modernization/2020/07/va-wants-automate-digitization-its-5-mile-high-electronic-health-record-backlog/166769/> [https://perma.cc/A33G-36U2]; Abigail Hauslohner, *The Employment Green Card Backlog Tops 800,000, Most of Them Indian. A Solution Is Elusive.*, WASH. POST (Dec. 17, 2019, 5:26 PM), [https://www.washingtonpost.com/immigration/the-employment-green-card-backlog-tops-800000-most-of-them-indian-a-solution-is-elusive/2019/12/17/55def1da-072f-11ea-8292-c46ee8cb3dce\\_story.html](https://www.washingtonpost.com/immigration/the-employment-green-card-backlog-tops-800000-most-of-them-indian-a-solution-is-elusive/2019/12/17/55def1da-072f-11ea-8292-c46ee8cb3dce_story.html) [https://perma.cc/THS8-DLR2]; see also ACUS REPORT, *supra* note 1, at 37–45 (discussing use of AI system for social security case processing).

77. See MURPHY, *supra* note 64, at 2.

78. Cf. Citron, *supra* note 32, at 1271–72 (discussing “automation bias”); *infra* notes 153–57 and accompanying text (same).

## 2. Data Selection and Preparation

Once the system objectives have been established, developers must assemble a corpus of data to train a machine learning algorithm. Because AI models are “only as good as the data” that trains them,<sup>79</sup> data selection and preparation are arguably the most important parts of the development process.

In our hypothetical, the agency’s previously decided cases will be a primary source of training data. All else equal, the accuracy of machine learning algorithms improve with exposure to more data.<sup>80</sup> Thus, five years of training data may be better than three years of comparable data. But if the data quality depreciates over time, then a tradeoff between data quality and quantity is unavoidable.<sup>81</sup> These sorts of discretionary judgments will generally be made by data scientists with input from subject matter experts.<sup>82</sup>

In some contexts, data selection may also entail legal judgment, political choice, or some combination thereof. For instance: should the training data exclude cases decided prior to a relevant statutory amendment, and should cases from certain regions or subpopulations be included?<sup>83</sup> Whatever the answers, they will be informed by normative and analytic judgments that will directly impact the AI system’s performance in potentially consequential ways. As Michael Kearns and Aaron Roth explain, “[w]hen maximizing accuracy across multiple different populations, an algorithm will naturally optimize better for the majority population, at the expense of the minority population[.]”<sup>84</sup>

## 3. Model Training

Once the data is ready, the development team can use it to train a machine learning algorithm. In general, the goal of this phase is to optimize an algorithm’s objective function, which is a “mathematical expression of the algorithm’s goal.”<sup>85</sup> In lay terms, as applied to our

---

79. Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 106 (2014).

80. See Lehr & Ohm, *supra* note 63, at 678 (“To reap the predictive benefits of machine learning, a sufficiently large number of observations is required.”).

81. See WITTEN ET AL., *supra* note 64, at 60 (“Domain experts need to be consulted to explain anomalies, missing values, the significance of integers that represent categories rather than numeric quantities, and so on.”).

82. *Id.*

83. Cf. John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 WASH. L. REV. 1725, 1794 (2018) (discussing the mismatch that can occur, in the criminal bail context, if training data is collected before the implementation of bail reforms).

84. KEARNS & ROTH, *supra* note 74, at 78. This occurs because, by definition, “there are more people from the majority group, and hence they contribute more to the overall accuracy of the model.” *Id.*

85. Lehr & Ohm, *supra* note 63, at 671.

hypothetical, the goal is to train an AI model that can accurately predict which cases in the backlog are meritorious (and, as much as possible, to flag *only* those cases).<sup>86</sup> No attempt will be made here to capture all the ingenuity, craft, analysis, and discretionary judgment that model training entails.<sup>87</sup> But, as one pivotal example, AI developers must make tradeoffs between two types of prediction errors: false positives and false negatives.<sup>88</sup>

In our hypothetical, the *false positives* are the cases that the AI system erroneously flags as likely winners. The *false negatives* are the meritorious cases that the AI system misses.<sup>89</sup> For certain types of algorithms, the ratio of false positives to false negatives can be predetermined and forced by code.<sup>90</sup> For other algorithms, the error rates can be manipulated by adjusting sensitivity thresholds.<sup>91</sup> Whether to err on the side of false negatives or false positives, and how much so, are policy choices of great significance. As such, different stakeholder could choose differently, based on different considerations and calculations. Indeed, as David Lehr and Paul Ohm explain, “it is very rare for a stakeholder to view being wrong in one way as equally harmful as being wrong in the opposite way.”<sup>92</sup> Moreover, context matters. The optimal error ratio for our hypothetical case-flagging system, whatever it may be, will probably not be the optimal error ratios for predicting cancer or prison violence. And, certainly, the considerations that inform those judgments are not the same. The key point here is that *humans* decide not only what to shoot for during model training, but also the metrics for “success.”

#### 4. Model Testing and Evaluation

After an AI model is trained, it must be tested and evaluated to explore its “fit” between the data and its target objective. The goal of this phase is to determine if, and how well, the trained model can generalize to make accurate predictions for *new* data inputs (i.e., outside of the original

---

86. Cf. *id.* at 671–72 (discussing objective functions).

87. Entire courses and books are filled with the math, science, and ingenuity of model training. See, e.g., WITTEN ET AL., *supra* note 64; Andrew Ng et al., *Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization*, COURSERA, [https://www.coursera.org/learn/deep-neural-network?trk\\_location=query-summary-list-link](https://www.coursera.org/learn/deep-neural-network?trk_location=query-summary-list-link) [https://perma.cc/RC69-EZ38].

88. See Lehr & Ohm, *supra* note 63, at 691–92 (discussing false positives and false negatives).

89. See *id.*

90. See *id.* (discussing opportunities to code an “asymmetric cost ratio”).

91. *Id.*

92. *Id.*

training data).<sup>93</sup> There are many different testing and validation methods, each with their own tradeoffs and limitations.<sup>94</sup> Moreover, AI developers must make normative choices about what to test for and why. Thus, for our hypothetical, the developers may (or may not) test the model to determine if it performs equally well on benefit applications filed by men and women, equally well on applications filed by Black men and White women, and so on.<sup>95</sup> In addition, the clustering of testing variables entails its own set of choices and empirical voids. Testing for demographic parity on the dimensions of race and gender, for example, may fail to properly account for intersectional differences within those groups along the dimensions of age and nationality.<sup>96</sup>

## 5. Model Selection and System Configuration

Throughout the development process, several different AI models may be trained, tested, and evaluated. Selecting which model(s) to deploy in an AI system is generally informed by a range of objectives, performance metrics, risks, and constraints.<sup>97</sup> Moreover, each of those considerations may be influenced by a mix of social, legal, financial, technical, political, and logistical considerations.

One important set of design choices pertain to model “interpretability.” In the AI field, interpretability refers to the ability of humans to understand an AI model’s logic or decisional pathway from inputs to outputs.<sup>98</sup> Some machine learning models are more scrutible

---

93. See Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, COMM’NS OF THE ACM, Oct. 2012, at 78, 81–82 (discussing challenges relating to overfitting and underfitting).

94. See *id.*; Lehr & Ohm, *supra* note 63, at 699–701.

95. Cf. Barocas & Selbst, *supra* note 18, at 677 (explaining why an AI system might not perform equally across groups and subgroups); see also *infra* Section II.B.2 (discussing technical and non-technical causes of algorithmic bias and other forms of unfairness).

96. See Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, in FACCT ’20: PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 39 (2020), <https://arxiv.org/pdf/2001.00973.pdf> [<https://perma.cc/GH6A-HXKC>] (“Algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values.”).

97. See Lehr & Ohm, *supra* note 63, at 690 (discussing six considerations for model selection: “the kind of output variable, the ability to implement an ‘asymmetric cost ratio,’ the ability to explain or offer reasons for the predictions, the potential for overfitting, the opportunities for tuning, and practical resource limitations”).

98. See Brent Mittelstadt et al., *Explaining Explanations in AI*, in FACCT ’19: PROCEEDINGS OF THE 2019 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 279, 280 (2019), <https://arxiv.org/pdf/1811.01439.pdf> [<https://perma.cc/NEM3-KHLU>] (highlighting how “poorly interpretable models” are unable to reveal how classifications result from the inputs).

than others.<sup>99</sup> Linear regression algorithms, for example, are relatively easy for humans to comprehend but have limited functionality.<sup>100</sup> By contrast, deep learning neural networks drive some of the most powerful, sophisticated, and functional AI systems, but their complexity renders them inscrutable to humans.<sup>101</sup> The inputs and outputs of these black-box systems may be known, but the computational formulas that churn inputs into outputs may entail thousands, millions, or billions of parameters.<sup>102</sup> At that level of complexity, an AI's inner logic defies human comprehension.<sup>103</sup> Thus, in contexts where different machine learning algorithms might be suited for a particular task, AI developers may need to make a tradeoff between model accuracy and interpretability.<sup>104</sup> Put otherwise, models with better overall accuracy may be pitted against simpler models that, if selected, would allow humans to know when and why the AI's predictions are wrong.

Other important design considerations anticipate post-deployment human interactions with the AI model. For example, to run the model, humans may need to collect and input feature variables (e.g., criminal

---

99. See FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 3–4, 9 (2015) (shining critical light on the “black box” nature of algorithmic systems); Jatinder Singh et al., *Responsibility & Machine Learning: Part of a Process* 4–5 (Oct. 27, 2016) (unpublished manuscript), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2860048](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2860048) [https://perma.cc/DW2M-5PG7].

100. See Singh et al., *supra* note 99, at 4.

101. See *id.*; see also Scott Wisdom et al., *Interpretable Recurrent Neural Networks Using Sequential Sparse Recovery* 1 (Nov. 22, 2016) (unpublished manuscript), <https://arxiv.org/pdf/1611.07252.pdf> [https://perma.cc/TL24-M94R] (“Interpreting the learned features and outputs of machine learning models is problematic. This difficulty is especially significant for deep learning approaches [like neural networks], which are able to learn effective and useful function maps due to their high complexity.”).

102. A model parameter is a configuration variable that is internal to the model. OpenAI’s GPT-3, a language model capable of natural language processing tasks, “has a whopping 175 billion parameters.” Khari Johnson, *OpenAI Debuts Gigantic GPT-3 Language Model with 175 Billion Parameters*, VENTUREBEAT (May 29, 2020, 8:34 AM), <https://venturebeat.com/2020/05/29/openai-debuts-gigantic-gpt-3-language-model-with-175-billion-parameters/> [https://perma.cc/4GZQ-SJCL]; see also William Fedus et al., *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*, ARXIV (Jan. 11, 2021), <https://arxiv.org/abs/2101.03961> [https://perma.cc/6S54-BV3U]; Divye Singh, *New Contender in Trillion Parameter Model Race*, MEDIUM (June 8, 2021), <https://medium.com/geekculture/new-contender-in-trillion-parameter-model-race-6ef0675ddd46> [https://perma.cc/9UTN-425H].

103. See Singh et al., *supra* note 99, at 5–6; see also Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a “Right to an Explanation” Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 56–57 (2017) (discussing the “curse of dimensionality” that results in complex machine learning models when large amounts of variables are combined in complex ways so as to defy comprehension).

104. Cf. Cynthia Rudin, *Please Stop Explaining Black Box Models for High Stakes Decisions*, 1 NATURE MACH. INTEL. 206, 207–08 (2019) (urging the use of interpretable models in high-stakes contexts and challenging the myth that there is necessarily a tradeoff between accuracy and interpretability).

history, nationality, age, etc.) to generate an AI output. Humans may also need to decide what to do, if anything, with that output. In such systems, there is a so-called human *in-the-loop*<sup>105</sup>—which in our example might be agency adjudicators assigned to the flagged cases. But AI systems can also be designed to work autonomously post-deployment,<sup>106</sup> with or without a human monitor *on-the-loop*.<sup>107</sup> For all configurations, however, AI developers must exercise human judgment and foresight about who will use or control the system, in what contexts, for what purposes, and under what constraints.<sup>108</sup>

The resulting configurations can have major downstream implications.<sup>109</sup> For instance, the design of the human–computer interface may influence a developer’s *own* sense of responsibility, insofar as accountability is offloaded—rightly or wrongly—to operators and end-users during deployment.<sup>110</sup> Irrespective of a developer’s intentions, a human in-the-loop may be held to account for reasons beyond their control. And, in these or other scenarios, a human may become “a rubber stamp for the machine, providing nothing more than a cosmetic reason to lull [stakeholders] into feeling better about the results.”<sup>111</sup>

\* \* \*

As the foregoing discussion hopes to impress, machine learning AI systems are infused with countless value-laden choices and tradeoffs. Far too often, these human decisions are latent, unexpressed, ad hoc, post hoc, or myopically informed. Drawing them to the surface is a critical first step toward understanding why AI systems are less like calculators and

---

105. See Singh et al., *supra* note 99, at 13.

106. See *id.* at 14. An email spam filter is an example of autonomous AI.

107. See Joel E. Fischer et al., *In-the-Loop or On-the-Loop? Interactional Arrangements to Support Team Coordination with a Planning Agent*, CONCURRENCY & COMPUTATION PRAC. & EXPERIENCE, Mar. 6, 2017, at 2, <https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.4082> [<https://perma.cc/5USE-7BW6>] (distinguishing between humans in-the-loop and on-the-loop, and studying contexts in which one structuring might be preferable to others).

108. Some or all of this information may not be known during the development stage, or might be known but change over time in unforeseeable ways.

109. See Will Orr & Jenny L. Davis, *Attributions of Ethical Responsibility by Artificial Intelligence Practitioners*, INFO. COMM’N & SOC’Y 719, 725 (2020) (describing a “pattern of ethical dispersion” in machine-learning AI development, whereby “powerful bodies set the parameters, practitioners translate these parameters into tangible hardware and software, and then relinquish control to users and machines, which together foster myriad and unknowable outcomes”); see also Meg Leta Jones, *The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles*, 18 VAND. J. ENT. & TECH. L. 77, 90–91 (2015) (revealing how legal approaches that ignore the complex relations between humans and machines fail to protect the values legal approaches sought to protect).

110. See Orr & Davis, *supra* note 109, at 7 (discussing how perceptions of ethical responsibility are dispersed in AI development and deployment).

111. Lehr & Ohm, *supra* note 63, at 716.

more like calculated policy. This orientation, in turn, foists pressure on the organizing question of *who decides* what an AI's embedded policies should be.<sup>112</sup> In our case-flagging hypothetical, the government presumably made those important choices. Still, who decides *within* government can impact the functionality and features of deployed AI systems. By the same token, the government's decision to *outsource* AI development or deployment can be highly consequential. Especially if private vendors become the de facto deciders.<sup>113</sup>

## II. TOWARD ETHICAL ALGORITHMIC GOVERNANCE

This Part maps the promises and pitfalls of algorithmic governance. The discussion begins with AI's positive potential because that is what drives the demand for algorithmic governance in the first place. From that starting position, the discussion pivots to a range of recursive sociotechnical challenges inhering in machine learning systems: specifically as pertains to safety, fairness, transparency, and accountability. The Part concludes with a contextual rendering of the rise of ethical AI frameworks to address these challenges.<sup>114</sup>

### A. Good (Algorithmic) Governance

Ideally, AI can make government more efficient and effective across a wide range of functions: law enforcement, adjudication, rulemaking, national security, resource allocation, in-house management, delivery of public services, and beyond.<sup>115</sup> Moreover, with proper design, AI systems can provide greater accuracy than human deciders alone.<sup>116</sup> Further, AI

---

112. See *infra* notes 250–54 and accompanying text (generally discussing the government's build-or-buy choice).

113. Cf. GRY HASSELBACH ET AL., WHITE PAPER ON DATA ETHICS IN PUBLIC PROCUREMENT OF AI-BASED SERVICES AND SOLUTIONS 11 (2020) (“The government’s choice among competing market solutions will generally entail “a prioritization of interests and values embedded in [product] design.”).

114. See *infra* Section II.C; see also *infra* Part III (canvassing the many challenges of translating ethical AI principles in practice, for both industry and government actors).

115. See ACUS REPORT, *supra* note 1, at 6 (“Rapid developments in AI have the potential to reduce the cost of core governance functions, improve the quality of decisions, and unleash the power of administrative data, thereby making government performance more efficient and effective.”); see also *supra* notes 1–11 and accompanying text (providing examples of federal agency uses of AI).

116. See Coglianese & Lehr, *supra* note 37, at 6 (describing how machine learning algorithms produce “unparalleled accuracy” compared to other statistical methods and human judgment).

systems may be more transparent and accountable than government agents, who might conceal or be unaware of their own cognitive biases.<sup>117</sup>

The centralization of AI decision-making may also promote greater consistency across cases, both in public-facing operations (such as adjudication and law enforcement) and inward-facing operations (such as personnel retention).<sup>118</sup> What's more, centralized AI decision-making can facilitate audits of external and internal government programs.<sup>119</sup>

The government's adoption of AI technologies may also be fiscally responsible. By "automating repetitive tasks" and "augmenting" the capabilities of federal workers, taxpayer dollars can be saved or rerouted to better use.<sup>120</sup> According to one rosy estimate, the government's widespread adoption of AI could yield \$500 billion in cost reductions over the next decade.<sup>121</sup>

Suffice to say, AI has the potential to augment, enable, and vastly improve government operations. Beyond better, however, the government's rapid uptake of AI is arguably imperative "to protect [the

---

117. *See id.* ("We find reason to be optimistic that, notwithstanding machine learning's black-box qualities, responsible governments can provide sufficient transparency about their use of algorithms to supplement, and possibly even replace, human judgments."); David Freeman Engstrom & Daniel E. Ho, *Artificially Intelligent Government: A Review and Agenda*, in RESEARCH HANDBOOK ON BIG DATA LAW 64 (Roland Vogl ed., 2021) ("The perhaps counterintuitive result is that the displacement of enforcement discretion by algorithm might, on net, yield an enforcement apparatus that is less opaque and more legible to agency heads and reviewing courts alike than the existing system."); Kroll et al., *supra* note 12, at 656–77 (explaining how, through proper design, AI systems can be made more transparent and accountable).

118. The gains in consistency depend, in part, on whether humans-in-the-loop can adjust or override algorithmic scores and under what conditions or constraints. Cf. Evans & Koulish, *supra* note 4, at 794 (exposing how immigration enforcement agents manipulated an automated risk classification system used for immigration detention and release).

119. Alice Xiang, *Reconciling Legal and Technical Approaches to Algorithmic Bias*, 88 TENN. L. REV. (forthcoming 2021) (manuscript at 12), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3650635](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3650635) [<https://perma.cc/SUW2-2L5Q>] (noting the "potential for algorithms to centralize decision-making, which can make auditing decisions easier," relative to "hundreds or thousands of human decision-makers"). That said, the audibility and oversight of AI decision-making will depend on whether those decisions are documented, traceable, and transparent—none of which can be assumed under current conditions. *See* U.S. GOV'T ACCOUNTABILITY OFF., GAO-21-519SP, *supra* note 58 (providing an auditing framework for federal AI systems); cf. U.S. GOV'T ACCOUNTABILITY OFF., GAO-21-518, FACIAL RECOGNITION TECHNOLOGY: FEDERAL LAW ENFORCEMENT AGENCIES SHOULD BETTER ASSESS PRIVACY AND OTHER RISKS 20 (2021) (reporting that more than a dozen "federal agencies do not have awareness of what non-federal systems with facial recognition technology are used by [federal] employees," and "have therefore not fully assessed the potential risks of using these systems, such as risks related to privacy and accuracy").

120. CHRISTINA BONE ET AL., THE COMING AI PRODUCTIVITY BOOM AND HOW FEDERAL AGENCIES CAN MAKE THE MOST OF IT 2, 4 (2020).

121. *Id.* That estimate, however, was based on projections of government adoptions of AI systems at a much faster rate than current capabilities.

nation's] security, promote its prosperity, and safeguard the future of democracy.”<sup>122</sup> That was a top-line message delivered by the National Security Commission on Artificial Intelligence (NSCAI) to the President and Congress in 2021.<sup>123</sup>

While this Article is principally focused on civilian and domestic contexts, the global “AI arms race” is quite relevant here. Foremost, the global competition exerts gravitational pull on the government’s *entire* AI trajectory.<sup>124</sup> More concretely, the race anchors the government’s ambition to “retain [America’s] innovation leadership”<sup>125</sup>—which depends mightily on the industry’s capacities and cooperation. Further, the U.S./China juxtaposition—and narratives around it—crystallize the need for AI innovation and ideation that reflects American values.<sup>126</sup> As put by the NSCAI: If AI systems violate civil rights, or “have significant negative consequences, then leaders will not adopt them, operators will not use them, Congress will not fund them, and the American people will not support them.”<sup>127</sup>

---

122. NSCAI FINAL REPORT, *supra* note 11, at 8.

123. *Id.* at 8–9 (“[N]ational security professionals must have access to the world’s best [AI] technology to protect themselves, perform their missions, and defend us.”). The NSCAI was established by Section 1051 of the John S. McCain National Defense Authorization Act for Fiscal Year 2019, Pub. L. No. 115-232, 132 Stat. 1636 (2018). The NSCAI’s task is to make recommendations to the President and Congress to “advance the development of artificial intelligence [AI], machine learning, and associated technologies to comprehensively address the national security and defense needs of the United States.” *Id.*

124. See, e.g., Michael Kratsios, Opinion, *Why the US Needs a Strategy for AI*, WIRED (Feb. 11, 2019, 9:00 AM), <https://www.wired.com/story/a-national-strategy-for-ai/> [https://perma.cc/4NEB-S6SH] (“An AI future that enriches the lives of our citizens, promotes innovation, and ensures our national and economic security requires continued American leadership.”); Gary Grossman, *The AI Arms Race Has Us on the Road to Armageddon*, VENTURE BEAT (Apr. 19, 2021, 2:10 PM), <https://venturebeat.com/2021/04/19/the-ai-arms-race-has-us-on-the-road-to-armageddon/> [https://perma.cc/J3YJ-8YU6] (“For now, the AI arms race is a cold war, mostly between the U.S., China, and Russia, but worries are it will become more than that.”).

125. See NSCAI FINAL REPORT, *supra* note 11, at 11 (“The United States . . . must do what it takes to retain its innovation leadership and position in the world . . . and organize to win it by orchestrating and aligning U.S. strengths.”).

126. See Darren Byler, *China’s Hi-tech War on Its Muslim Minority*, GUARDIAN (Apr. 11, 2019, 1:00 PM), <https://www.theguardian.com/news/2019/apr/11/china-hi-tech-war-on-muslim-minority-xinjiang-uighurs-surveillance-face-recognition> [https://perma.cc/63LM-ZM93]; Paul Mozur, *Inside China’s Dystopian Dreams: A.I., Shame, and Lots of Cameras*, N.Y. TIMES (July 8, 2018), <https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html> [https://perma.cc/8J6E-R4J5]; Drew Donnelly, *An Introduction to the China Social Credit System*, NEW HORIZONS (Aug. 2, 2021), <https://nhglobalpartners.com/china-social-credit-system-explained> [https://perma.cc/VRZ8-KKSD].

127. NSCAI FINAL REPORT, *supra* note 11, at 133.

### B. Ethical Challenges

AI's social prospects are simultaneously alluring and alarming: we want efficient and effective AI systems, but not if they efficiently or effectively cause harm. Satisfying *all* of these conditions may not be possible; AI's promises and perils are hard to decouple.

Facial recognition systems, for example, can be deployed to find lost children and terrorists, but can also be used for Orwellian surveillance and social repression.<sup>128</sup> AI can help to mitigate climate change, but the computing resources and raw materials powering large-scale AI models are environmentally unsustainable.<sup>129</sup> AI can uncover and rectify social biases, but can also learn those biases from training data and project them into the future.<sup>130</sup>

If there is any sense in which AI is neutral, it is the technology's dual capacity for good and evil when deployed by humans in real-world settings. That's the rub and root of AI's sociotechnical challenges in general, and for algorithmic governance especially. Although necessarily partial, the discussion below is centered around four pillars of ethical AI: safety, fairness, transparency, and accountability.<sup>131</sup> Broadly conceived, these principles link to the government's legal obligations and norms of good governance.

---

128. See Clare Garvie & Laura M. Moy, *America Under Watch: Face Surveillance in the United States*, GEO. L. CTR. ON PRIV. & TECH. (May 16, 2019), <https://www.americaunderwatch.com/> [https://perma.cc/4BXQ-8X98] ("When used on public gatherings, face surveillance may have a chilling effect on our First Amendment rights to unabridged free speech and peaceful assembly."); see also GEORGE ORWELL, *NINETEEN EIGHTY FOUR* (1949) (depicting a canonical surveillance state).

129. See Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in FACCT '21: PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 610, 612–13 (2020), <https://dl.acm.org/doi/10.1145/3442188.3445922> [https://perma.cc/4FNP-CXBF] (providing a critical account of the environmental costs of training large-language AI models); Amy Stein, *Artificial Intelligence and Climate Change*, 37 YALE J. ON REG. 890, 892–93, 919 (2020) (arguing for the enhanced use of AI to address climate change, and discussing the "need to ensure that AI's negative environmental impacts are outweighed by its positive ones"); see also Thomas Griffin, *Why We Should Care About the Environmental Impact of AI*, FORBES (Aug. 17, 2020, 9:10 AM), <https://www.forbes.com/sites/forbestechcouncil/2020/08/17/why-we-should-care-about-the-environmental-impact-of-ai> [https://perma.cc/HV3A-B28Y].

130. See Steven Mills, *Foreword* to MONTREAL AI ETHICS INST., *The State of AI Ethics: January 2021*, at 11, 12 (2021), <https://montrealethics.ai/wp-content/uploads/2021/01/The-State-of-AI-Ethics-Report-January-2021.pdf> [https://perma.cc/AQ8W-H4WB]; see also Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2226 (2019) ("Counterintuitively, algorithmic assessment could play a valuable role in a system that targets the risky for support rather than for restraint.").

131. See Jobin et al., *supra* note 40, at 391 (mapping and analyzing a corpus of ethical AI principles and guidelines from around the globe, and finding a convergence around the principles of "transparency, justice and fairness, non-maleficence, responsibility and privacy").

## 1. Safety

In the AI field, “safety” connotes a range of system attributes, including model accuracy, reliability, robustness, and security.<sup>132</sup> When exposed to real-world elements, AI systems make mistakes that most humans never would. In part, that is because the variance and complexity of the real world may not be captured in the training data, which is all the algorithm knows.<sup>133</sup> Of course, humans make mistakes too. But the efficiency and scalability of AI systems make them uniquely concerning.<sup>134</sup> As Robert Brauneis and Ellen Goodman explain, “[t]he ability of these algorithmic processes to scale, and therefore to influence decisions uniformly and comprehensively, magnifies any error or bias that they embody[.]”<sup>135</sup> Here, math is relevant: an efficient AI decision-making system that makes 100,000 predictions at a 10% error rate may harm 10,000 individuals; by contrast, an inefficient human that makes 100 predictions at a 25% error rate may harm 25 individuals.

The safety of AI systems can also be compromised by adversarial attacks. For example, a malicious actor can manipulate data upon which an AI model will be trained and tested.<sup>136</sup> Such “data poisoning” can result in “curated misclassification, systemic malfunction, and poor performance.”<sup>137</sup> A malicious actor can also manipulate deployed AI models to induce gross miscalculations.<sup>138</sup> This brittleness, at scale, can have profound consequences in high-stakes and safety-critical contexts. Certainly, humans can be manipulated, bribed, or spied upon by malicious actors in ways that undermine or endanger public interests. Yet AI systems have similar human vulnerabilities throughout the developmental pipeline *plus* attack surfaces in the data, code, cloud, and hardware, across sprawling supply chains.<sup>139</sup>

---

132. See Leslie, *supra* note 71, at 30.

133. Fábio Kepler, *Why AI Fails in the Wild*, UNBABEL (Nov. 15, 2019), <https://unbabel.com/blog/artificial-intelligence-fails/> [https://perma.cc/VHN9-RNKT] (“When unrepresentative data is used for training, sometimes with no considerations about how the training data was collected or where it came from, it can be very problematic to apply a model to different situations from the ones it knows.”); see also Colin Smith et al., *Hazard Contribution Modes of Machine Learning Components*, in THE AAAI-20 WORKSHOP ON ARTIFICIAL INTELLIGENCE SAFETY 4 (2020), <http://ceur-ws.org/Vol-2560/paper41.pdf> [https://perma.cc/2D27-GECB] (discussing unexpected performance, for example, “through unanticipated feature interaction . . . that was also not previously observed during model validation”).

134. See Brauneis & Goodman, *supra* note 18, at 129; see also O’NEIL, *supra* note 18, at 29–31 (discussing the scalability of algorithms and consequent risk of widespread harm).

135. Brauneis & Goodman, *supra* note 18, at 129.

136. Leslie, *supra* note 71, at 32–33.

137. *Id.* at 33.

138. *Id.* at 32–33.

139. Cf. Exec. Order No. 14,028, 86 Fed. Reg. 26,633, 26,637 (May 12, 2021) (noting, in a

The foregoing safety risks compound when AI systems interact with other technologies, analog systems, or new environmental conditions.<sup>140</sup> For example, the output from one AI system may become another system's input, and so on. The resulting domino effects and feedback loops can cause "systems of systems" to drift from their anticipated performance, often imperceptibly and dangerously.<sup>141</sup> This arguably occurred in Michigan, for example, when the state used an AI system created by a commercial vendor to detect fraudulent claims for unemployment benefits.<sup>142</sup> For a variety of reasons, the system had a high error rate that caused tens of thousands of individuals to suffer life-changing financial harm, not only from the denial of benefits, but also in the form of collateral penalties, interest, and lost wages.<sup>143</sup> The resulting damage, and class-action litigation, spans years and is still ongoing.<sup>144</sup>

## 2. Fairness

"Fairness" has no agreed-upon meaning.<sup>145</sup> A concept like "justice" may work just as well or better, but it too has no agreed-upon meaning.<sup>146</sup> For present purposes, what matters is the breadth of concerns that fairness (or justice) captures, including nondiscrimination, due process, autonomy, inclusivity, equal opportunity, and fair dealing. To greater and lesser extents, AI may cohere with these human values—but that alignment will not obtain by default.

---

lengthy executive order to improve the nation's cybersecurity posture, that "the development of commercial software often lacks transparency, sufficient focus on the ability of the software to resist attack, and adequate controls to prevent tampering by malicious actors").

140. See Singh et al., *supra* note 99, at 15.

141. *Id.* (noting that the interactions can be direct or more indirect, through "butterfly effects," where subtle actions of a system can affect others in potentially dramatic ways).

142. See Cahoo v. SAS Inst. Inc., 377 F. Supp. 3d 769, 771 (E.D. Mich. 2019).

143. See Kate Crawford et al., *AI Now 2019 Report*, AI Now INST. 36 (2019), [https://ainowinstitute.org/AI\\_Now\\_2019\\_Report.html](https://ainowinstitute.org/AI_Now_2019_Report.html) [<https://perma.cc/E4ZL-DXEL>].

144. See Cahoo v. Fast Enters. LLC, 528 F. Supp. 3d 719, No. 17-10657, 2021 WL 1146119 (E.D. Mich. Mar. 25, 2021); Cahoo v. SAS Analytics Inc., 912 F.3d 887, 892 (6th Cir. 2019); see also Alejandro De La Garza, *States' Automated Systems Are Trapping Citizens in Bureaucratic Nightmares with Their Lives on the Line*, TIME (May 28, 2020, 2:24 PM), <https://time.com/5840609/algorithm-unemployment/> [<https://perma.cc/CA99-9CT3>].

145. See Abigail Z. Jacobs & Hanna Wallach, *Measurement and Fairness*, in FACCT '21: PROCEEDINGS OF THE ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 375, 375 (2021), <https://arxiv.org/pdf/1912.05511.pdf> [<https://perma.cc/3HEG-6YBN>].

146. See Reuben Binns, *Fairness in Machine Learning: Lessons from Political Philosophy*, in FACCT '18: PROCEEDINGS OF THE 2019 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1, 1 (2018), <http://proceedings.mlr.press/v81/binns18a/binns18a.pdf> [<https://perma.cc/CA4Z-VHRL>] (drawing parallels between justice and fairness, and noting that "attempts to formalise 'fairness' in machine learning contain echoes of these old philosophical debates").

Intentionally or not, AI models absorb social biases contained in training data.<sup>147</sup> For example, an AI system deployed to predict criminal activity will exhibit higher false positive rates for Black defendants if the training data is an artifact of discriminatory policing.<sup>148</sup> Likewise, AI language models that learn from a corpus of text scraped from the internet will exhibit the anti-Semitic, anti-Muslim, and gender biases captured in online content.<sup>149</sup> Especially in complex AI systems, it can be difficult to identify algorithmic biases until they manifest, and quite difficult to fix post hoc.<sup>150</sup>

Of course, social bias is not specific to AI; humans are biased too. This cynical observation, however, only sharpens the point: human biases throughout society get captured in data, which gets imbued in AI models that make biased classifications and predictions. Through these dynamics—and with objective veneer—historical data is effectively laundered into the future and dispersed through networks of AI and analog systems.<sup>151</sup>

For sensitive government decisions, humans in-the-loop may be expected to exercise human judgment as a check on algorithmic classifications and predictions.<sup>152</sup> Studies show, however, that humans

---

147. See U.S. GOV'T ACCOUNTABILITY OFF., GAO 21-519SP, *supra* note 58, at 9 (“Biases arise from the fact that AI systems are created using data that may reflect preexisting biases or social inequities.”).

148. See, e.g., Rashida Richardson et al., *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 15, 20 (2019) (finding that nine of thirteen police departments that were studied likely used “dirty data” to train and used predictive policing algorithms, many of which were acquired from the private sector with federal funding); Letter from Members of Congress to Merrick Garland, Att’y Gen. (Apr. 15, 2021) (on file with author) (requesting information about federal funding and oversight of predictive policing algorithms and asserting that these algorithms “likely . . . amplify biases against historically marginalized groups”).

149. Tom Brown et al., *Language Models Are Few-Shot Learners* 36–39 (July 22, 2020) (unpublished manuscript), <https://arxiv.org/pdf/2005.14165.pdf> [<https://perma.cc/X6K2-L6ES>] (“[I]nternet-trained models have internet-scale biases; models tend to reflect stereotypes present in their training data.”); see also Abubakar Abid et al., *Persistent Anti-Muslim Bias in Large Language Models* 9–10 (Jan. 18, 2021) (unpublished manuscript), <https://arxiv.org/abs/2101.05783> [<https://perma.cc/D7H5-5KGU>].

150. See Brown et al., *supra* note 149, at 32–39 (discussing preliminary findings of bias in GPT-3 along the dimensions of gender, race, and religion); *id.* at 39 (noting that efforts to “remove” bias have “been shown to have blind spots”); Karen Ho, *This Is How AI Bias Really Happens—and Why It’s So Hard to Fix*, MIT TECH. REV. (Feb. 4, 2019), <https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix> [<https://perma.cc/C39Y-GELM>].

151. See U.S. GOV’T ACCOUNTABILITY OFF., GAO 21-519SP, *supra* note 58, at 9 (observing that AI systems have the “potential to amplify existing biases and concerns related to civil liberties, ethics, and social disparities”).

152. See *supra* notes 106–07 and accompanying text (discussing humans in-the-loop and on-the-loop).

are prone to over rely on computer recommendations—a phenomenon known as “automation bias.”<sup>153</sup> Such complacency can result in human failure to identify or rectify AI errors.<sup>154</sup> Risks at the human–computer interface also run in the opposite direction—a phenomenon known as “algorithmic aversion.”<sup>155</sup> More specifically, humans that do not trust or understand the technology might seek to compensate for actual or perceived AI failures and shortfalls.

Even if well intended, these human compensations may be unwarranted, unfair, or illegal in some settings.<sup>156</sup> For instance, if an AI system used for government hiring exhibits bias toward men, to what extent can or should the hiring official upwardly adjust the algorithmic score for women?<sup>157</sup> Likewise, if an AI risk-assessment system used for bail determinations exhibits bias against Black defendants, to what extent can or should judges ignore or discount AI recommendations for Black (or White) defendants?<sup>158</sup> As these examples lay bare, correcting for

---

153. See Kate Goddard et al., *Automation Bias: Empirical Results Assessing Influencing Factors*, 83 INT'L J. MED. INFORMATICS 368, 368–69 (2014); Citron, *supra* note 32, at 1271–72.

154. Moreover, in the long run, systemic automation bias can have atrophying effects on domain expertise and human judgment. See ACUS REPORT, *supra* note 1, at 8 (“Managed poorly, government deployment of AI tools can hollow out the human expertise inside agencies with few compensating gains, widen the public-private technology gap, increase undesirable opacity in public decision-making, and heighten concerns about arbitrary government action and power.”).

155. Cf. Berkeley J. Dietvorst et al., *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*, 144 J. EXPERIMENTAL PSYCH. 114, 114 (2015) (finding that “people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster”).

156. Alice Xiang & Daniel E. Ho, *From Affirmative Action to Affirmative Algorithms: The Legal Challenges Threatening Progress on Algorithmic Fairness*, P’SHIP ON A.I.: BLOG (Nov. 9, 2020), <https://www.partnershiponai.org/affirmativealgorithms> [<https://perma.cc/B7GA-X83X>] (“The ways in which many of us in the AI community have moved to mitigate bias in the algorithms we develop may pose serious legal risks of violating equal protection.”).

157. Compare Kim, *supra* note 33, at 191 (arguing that, “despite its limitations, auditing for discrimination should remain an important part of the strategy for detecting and responding to biased algorithms,” and moreover, that “the law permits the use of auditing to detect and correct for discriminatory bias.”), with Kroll et al., *supra* note 12, at 694–95 (expressing skepticism about auditing as a strategy for detecting and correcting algorithmic bias, on both technological and legal grounds).

158. Cf. Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [<https://perma.cc/JRR9-5D29>] (finding that one system erroneously categorized Black defendants as future criminals at nearly twice the rate as it did for White defendants); Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It’s Actually Not That Clear.*, WASH. POST (Oct. 17, 2016, 5:00 AM), <https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/> [<https://perma.cc/7M7V-GPKL>] (countering that the problem ProPublica identified is “mathematically guaranteed” given historical data showing disparate recidivism rates for Black and White offenders combined with a particular definition of fairness).

perceived or actual unfairness in algorithmic governance requires a mix of legal, technical, social, and pragmatic considerations—none of which are settled.<sup>159</sup>

Another source of algorithmic bias stems from cultural and experiential blind spots in the technology industry. According to recent studies, only 26% of computing related jobs are held by women at the leading technology firms,<sup>160</sup> and the share of technical workers who are Black sits below 4%.<sup>161</sup> While the causes of these diversity challenges are complex and contestable, the consequences are widely acknowledged: demographic hegemony “affects how AI companies work, what products get built, who they are designed to serve, and who benefits from their development.”<sup>162</sup>

Facial recognition technology is perhaps the most notorious manifestation of these sociotechnical challenges.<sup>163</sup> In 2018, pioneering work by Joy Buolamwini and Timnit Gebru demonstrated that three prominent facial recognition systems performed significantly worse on people of color, especially women of color.<sup>164</sup> The disparities were not intentional; the AI models were optimized to fit the training data of predominantly White faces. This revelation prompted additional studies,

---

159. See, e.g., Engstrom & Ho, *supra* note 37, at 806 (finding it “far from certain” that current doctrine “will resolve the most pressing cases” in algorithmic governance); Huq, *supra* note 33, at 1917–27 (discussing the difficulties that arise in transposing the equal protection doctrine to the machine learning context). For additional sources and viewpoints on these issues, see *supra* note 33.

160. Sam Daley, *Women in Tech Statistics Show the Industry Has a Long Way to Go*, BUILT IN (May 5, 2021), [https://builtin.com/women-tech/workplace-statistics](https://builtin.com/women-tech/women-in-tech-workplace-statistics) [https://perma.cc/N4VF-SZ4Q].

161. Michael Ellison, *This Is How Big Tech Is Failing Its Black Employees*, FAST CO. (Oct. 21, 2020), <https://www.fastcompany.com/90565387/why-big-techs-lofty-diversity-reports-fell-so-far-from-expectations> [https://perma.cc/VP4S-8JP4] (“The share of technical workers who are Black at Facebook, Google, and Microsoft has inched up less than one percentage point since 2014 and still sits below 4% at each company.”).

162. West et al., *supra* note 75, at 5; see also RUHA BENJAMIN, RACE AFTER TECHNOLOGY: ABOLITIONIST TOOLS FOR THE NEW JIM CODE 4 (2019) (arguing “that human social bias is engineered into automated technology because (overwhelmingly White and male) programmers fail to recognize how their understanding of technology is informed by their identities”).

163. There are many more examples of this problem in hiring, lending, medicine, and beyond. See, e.g., Xiang, *supra* note 119, at 17–18 (discussing biased algorithms and applications in hiring and healthcare); Ted Knutson, *AI Lending Discrimination Needs To Be Tackled with Legislation Says House Financial Services Chair*, FORBES (May 7, 2021, 2:29 PM), <https://www.forbes.com/sites/tedknutson/2021/05/07/ai-lending-discrimination-needs-to-be-tackled-with-legislation-says-house-financial-services-chair> [https://perma.cc/XAR9-Z4VZ] (discussing discrimination in AI systems used for lending decisions).

164. Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. OF MACH. LEARNING RSCH. 1, 1 (2018); see also *Overview of Gender Shades Project*, MIT MEDIA LAB, <https://www.media.mit.edu/projects/gender-shades/overview/> [http://perma.cc/AFB3-GHVF].

including by the National Institute of Standards and Technology (NIST), which found that several facial recognition systems—including those used for law enforcement—were exponentially more likely to misidentify people of color.<sup>165</sup>

Algorithmic bias can be mitigated with better data practices, technical patches, and workarounds.<sup>166</sup> But even if the technology can be made perfectly accurate, that would not address structural concerns relating to power, autonomy, and liberty.<sup>167</sup> For example, the government's ability to engage in sprawling and accurate digital surveillance says nothing about whether that capability should be wielded, for what purposes, in what contexts, and over what communities or subpopulations.<sup>168</sup>

Moreover, even when working as intended, AI models are statistical simplifications that necessarily treat people as members of groups, not as individuals.<sup>169</sup> That might be of less consequence or concern when an algorithm recommends songs based on group characteristics. But when deployed in high-stakes or sensitive government contexts, algorithmic stereotyping can undermine a person's sense of autonomy, unfairly

---

165. Drew Harwell, *Federal Study Confirms Racial Bias of Many Facial-Recognition Systems, Casts Doubt on Their Expanding Use*, WASH. POST (Dec. 19, 2019), <https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/> [https://perma.cc/AK25-NYNT] (“Asian and African American people were up to 100 times more likely to be misidentified than white men, depending on the particular algorithm and type of search.”); see also Jacob Snow, *Amazon’s Face Recognition Falsey Matched 28 Members of Congress with Mugshots*, ACLU (July 26, 2018, 8:00 AM), <https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28> [https://perma.cc/8WCR-42J6] (reporting that Amazon’s facial recognition system wrongly matched 28 members of Congress to criminal mug shots).

166. In 2015, for example, Google’s image recognition system was found to classify African-Americans as “gorillas.” Tom Simonite, *When It Comes to Gorillas, Google Photos Remains Blind*, WIRED (Jan. 11, 2010, 7:00 AM), <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/> [https://perma.cc/S2HE-RRNZ]. Google “fixed” the problem by removing gorillas from the service’s lexicon. *Id.* This workaround is highly revealing of the *technical challenges* of retrospective solutions to algorithmic bias, as well as the *social challenges* of grappling with structural bias.

167. See West et al., *supra* note 75, at 18 (“[T]hough improving the performance of AI systems might be a necessary step toward making them more inclusive, there are some contexts in which ‘fixing’ such inaccuracies may not fix the overall problems presented by such systems—and some problems that cannot be fixed by a technical solution at all.”).

168. See *id.*; Julia Powles & Helen Nissenbaum, *The Seductive Diversion of “Solving” Bias in Artificial Intelligence*, ONEZERO (Dec. 7, 2018), <https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53> [https://perma.cc/TT4E-4JZC].

169. See Mulligan & Bamberger, *supra* note 37, at 787.

deprive them of benefits and opportunities,<sup>170</sup> or affirmatively cause harm.<sup>171</sup>

### 3. Transparency

AI transparency can be frustrated by a host of technical and non-technical reasons.<sup>172</sup> Regardless of the cause, the opacity of AI systems can be highly problematic. Most basically, if stakeholders do not understand how an AI system works, they may not understand when or why it fails. Without transparency, moreover, the government may lack moral or legal justification to act upon an AI model's outputs.<sup>173</sup> “[T]he algorithm made me do it” will simply not do in contexts where an individual's rights or well-being are compromised.<sup>174</sup>

---

170. This sense of unfairness was on full display after the British government cancelled the annual “A-Level” qualification exams for university placements due to the COVID-19 pandemic. Bryan Walsh, *How an AI Grading System Ignited a National Controversy in the U.K.*, AXIOS (Aug. 19, 2020), <https://wwwaxios.com/england-exams-algorithm-grading-4f728465-a3bf-476b-9127-9df036525c22.html> [<https://perma.cc/298X-F8KA>]; see Kelsey Piper, *The UK Used a Formula to Predict Students' Scores for Canceled Exams. Guess Who Did Well.*, VOX (Aug. 22, 2020, 7:30 AM), <https://www.vox.com/future-perfect/2020/8/22/21374872/uk-united-kingdom-formula-predict-student-test-scores-exams> [<https://perma.cc/6H7D-RYFU>]. As a “stand-in for actual scores,” an AI system predicted how students would have performed on the exam. *Id.* Because the algorithm placed significant weight on the past performance of the students’ schools, students “lost the chance to be treated as individuals.” Walsh, *supra*. Moreover, because many of the lower performing schools were also less affluent, students lost the opportunity to outscore their predicted performance and earn a place in more affluent learning institutions. *Id.*; see also *Exam Algorithms: Some Lessons*, FTI CONSULTING (Aug. 21, 2020), <https://www.fticonsulting.com/emea/insights/articles/exam-algorithms-some-lessons> [<https://perma.cc/H92M-8KRU>].

171. See Reva Schwartz et al., *Draft NIST Special Publication 1270: A Proposal for Identifying and Managing Bias in Artificial Intelligence*, NAT'L INST. STANDARDS & TECH. 9 (June 2021), <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf> [<https://perma.cc/H3ZF-L6GJ>] (“[AI] tools that are designed to use aggregated data about groups to make predictions about individual behavior—a practice initially meant to be a remedy for non-representative datasets—can lead to biased outcomes.”).

172. See Jenna Burrell, *How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC’Y, Jan.–June 2016, at 1, 3–5.

173. See Mulligan & Bamberger, *supra* note 37, at 782 (arguing that “the policy choices embedded in system design fail the prohibition against arbitrary and capricious agency actions absent a reasoned decision-making process”); see also *supra* note 47; *infra* notes 267–268 and accompanying text (discussing possible legal implications of model inscrutability for constitutional and administrative law).

174. See FAIRNESS, ACCOUNTABILITY & TRANSPARENCY MACH. LEARNING, [www.fatml.org](http://www.fatml.org) [<https://perma.cc/EE2L-SNSY>] (“[T]here is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to ‘the algorithm made me do it.’”); Victoria Burton-Harris & Philip Mayor, *Wrongfully Arrested Because Face Recognition Can’t Tell Black People Apart*, ACLU (June 24, 2020), <https://www.aclu.org/news/privacy-technology/wrongfully-arrested-because-face-recognition-cant-tell-black-people-apart/> [<https://perma.cc/6AL9-ENJ8>] (following the arrest of a Black man that was misidentified by a facial recognition system, “[o]ne officer responded, ‘The computer must have gotten it wrong.’”).

Some AI models *cannot* be explained because of their complexity.<sup>175</sup> Other AI models *will not* be explained because they don't have to be.<sup>176</sup> Most pertinent here, federal agencies that procure AI solutions may not be privy to the vendor's trade secrets—for example, when the technology is acquired as a “commercial item off the shelf.”<sup>177</sup> Even if the government has access to vendor trade secrets, federal law or nondisclosure agreements may prevent disclosure of those secrets to litigants, lawmakers, or other stakeholders.<sup>178</sup>

Moreover, AI transparency may be self-defeating or dangerous in many government contexts. For example, full transparency of the input variables and source code of a fraud-detection system might facilitate “gaming” or “hacking” by adversarial actors.<sup>179</sup> For similar reasons, cybersecurity concerns may trump AI transparency in safety-critical domains, such as energy, transportation, telecommunication, voting systems, waterways, and more.<sup>180</sup> Finally, AI transparency may run

---

175. See *supra* notes 101–04 and accompanying text.

176. See PASQUALE, *supra* note 99, at 180–81; Sonia K. Katyal, *The Paradox of Source Code Secrecy*, 104 CORNELL L. REV. 1183, 1186–87 (2019) (explaining how “source code that underlies and governs automated decision making is hidden from public view, comprising an unregulated ‘black box’ that is privately owned and operated”); Brauneis & Goodman, *supra* note 18, at 159 (complaining that “the information allegedly protected by trade secret law may lie at the heart of essential public functions and constitute political judgments long open to scrutiny”).

177. See *infra* Section IV.D.1 (discussing commercial off-the-shelf acquisitions); see also 48 C.F.R. § 12.212 (2021) (providing, for the acquisition of commercially available computer software, that vendors generally “shall not be required to . . . [r]elinquish to, or otherwise provide, the Government rights to use, modify, reproduce, release, perform, display, or disclose commercial computer software or commercial computer software documentation except as mutually agreed to by the parties”).

178. Katherine Fink, *Opening the Government’s Black Boxes: Freedom of Information and Algorithmic Accountability*, 21 INFO. COMM’N & SOC’Y 1453, 1456–59 (2017) (reviewing current state of law and practice with respect to whether algorithms would be considered “records” under the Freedom of Information Act (FOIA) and reviewing agency bases for withholding algorithms and source code under FOIA requests); Wexler, *supra* note 35, at 1396–417 (describing and critiquing how trade secrecy has been used to prevent criminal defendants to gain access to information about AI risk-assessment tools used in the criminal justice system).

179. See, e.g., Engstrom & Ho, *supra* note 117, at 68 (discussing the risk of gaming the suite of tools under development at the U.S. Patent and Trademark Office to help examiners adjudicate patent applications); Leslie, *supra* note 71, at 32–34 (discussing a range of adversarial attacks and failure modes); NSCAI FINAL REPORT, *supra* note 11, at 47 (same). But cf. Ignacio N. Cofone & Katherine J. Strandberg, *Strategic Games and Algorithmic Secrecy*, 64 MCGILL L.J. 623, 623 (2019) (detailing the relationship between gaming and transparency, and arguing that the threat of gaming is overblown in many contexts and often addressable in ways that do not require secrecy).

180. See, e.g., David S. Levine, *Secrecy and Unaccountability: Trade Secrets in Our Public Infrastructure*, 59 FLA. L. REV. 135, 135 (2007) (describing and critiquing how government outsourcing creates transparency and accountability gaps around critical public infrastructures); Brauneis & Goodman, *supra* note 18, at 175–76; see also Exec. Order No. 14,028, 86 Fed. Reg.

headlong into privacy laws in a wide variety of contexts where the government is provided personal data.<sup>181</sup> This tension arises, for example, when the inputs or outputs of AI systems contain protected or sensitive information that can be traced to individuals (even when names and other identifying information are scrubbed from the data).<sup>182</sup>

In short, AI transparency is undercut by a mix of technical, commercial, and legal issues that coalesce to keep much of algorithmic governance in the dark. Whether justifiably so is a matter of debate—normatively, doctrinally, and contextually.<sup>183</sup>

---

26,633, 26,633 (May 17, 2021) (“[T]he trust we place in our digital infrastructure should be proportional to how trustworthy and transparent that infrastructure is, and to the consequences we will incur if that trust is misplaced.”).

181. See, e.g., 5 U.S.C. § 552a(b) (prohibiting disclosure of records without the prior written consent of the person whom the records pertain to, excepting for reasons such as routine use for, *inter alia*, census purposes, matters of the House of Congress or any of its committees or subcommittees, etc.); Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 26, 29, and 42 U.S.C.) (setting forth privacy and security standards for protecting personal health information); *see also* Engstrom & Ho, *supra* note 117, at 65 (explaining that “privacy and data security constraints, while designed to safeguard privacy and minimize public burdens, can also impose significant costs on agencies, reduce the efficacy of algorithmic tools, and stymie agency innovation”).

182. See KEARNS & ROTH, *supra* note 74, at 30–33 (discussing how data can be extracted from AI models and de-anonymized); Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 111. This is a major concern, especially because of malicious threats to information security. See, e.g., Zolan Kanno-Youngs & David E. Sanger, *Border Agency’s Images of Travelers Stolen in Hack*, N.Y. TIMES (June 10, 2019), <https://www.nytimes.com/2019/06/10/us/politics/customs-data-breach.html> [<https://perma.cc/GW6H-9TNQ>]; Julie Hirschfield Davis, *Hacking of Government Computers Exposed 21.5 Million People*, N.Y. TIMES (July 9, 2015), <https://www.nytimes.com/2015/07/10/us/office-of-personnel-management-hackers-got-data-of-millions.html> [<https://perma.cc/FF2P-NBQB>].

183. See, e.g., Coglianese & Lehr, *supra* note 37, at 40–49 (arguing that government use of AI can generally comport with constitutional due process, as well as administrative law’s reason-giving and transparency norms); Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM L. REV. 1265, 1273, 1295–306 (2020) (exploring the “procedural and substantive conflicts between proprietary [algorithmic] decision-making on the one hand and government transparency obligations under the First Amendment and [Freedom of Information Act] on the other”); Citron, *supra* note 32, at 1281–88 (discussing how agency use of automated systems raises due process concerns); Mulligan & Bamberger, *supra* note 37, at 782 (arguing that “policy choices embedded in system design fail the prohibition against arbitrary and capricious agency actions absent a reasoned decision-making process that enlists the expertise necessary for reasoned deliberation, provides justifications for such choices, makes visible the political choices being made, and permits iterative human oversight and input”). See also Brauneis & Goodman, *supra* note 18, at 152–63 (spotlighting the transparency deficits that accrue when state and local government adopt AI systems developed by third parties).

#### 4. Accountability

The foregoing transparency challenges have major implications for government accountability. The less stakeholders know, the more difficult it becomes to ascertain whether an AI system is being used, and if so, whether that use is properly authorized, justified, and legal. As of this writing, there is no publicly available register of AI systems currently used by which federal actors, for what purposes, from what sources, and under what authority.<sup>184</sup> This is highly problematic for two related reasons: first, the opacity shutters out stakeholder input; second, the opacity breeds public distrust around the government's use of AI systems (including benign and potentially beneficial uses).<sup>185</sup>

Judicial review is another way that our legal system might hold government actors accountable for their use of AI systems. Conceivably, courts could also hold government actors accountable for the technical and non-technical value judgments embedded in or emanating from AI systems. The opacity of AI systems, however, can stymie a court's ability to perform these functions.

Beyond judicial settings, government watchdogs, journalists, and stakeholders are similarly constrained in their ability to "look under the hood" of AI tools affecting the polity's rights and interests.<sup>186</sup> As the Government Accountability Office (GAO) acknowledged in a 2021 report: "The U.S. government, industry leaders, professional associations, and others have begun to develop principles and frameworks to address [transparency and fairness] concerns, but there is limited information on how these will be implemented to allow for third-party assessments and audits of AI systems."<sup>187</sup>

Lines of accountability, moreover, are frequently tangled because AI systems are assemblages of datasets, technology stacks, and complex human networks.<sup>188</sup> When things go wrong, it can be far from clear which

---

184. See Rubenstein, *supra* note 24, at 20–21 (urging the creation of a federal registry of federal AI use cases). A 2020 executive order calls for such a catalog. See Exec. Order No. 13,960, 85 Fed. Reg. 78,939, 78,941 (Dec. 8, 2020). A few European cities have launched AI registries, with more jurisdictions likely to follow this type of proactive disclosure. See Khari Johnson, *Amsterdam and Helsinki Launch Algorithm Registries to Bring Transparency to Public Deployments of AI*, VENTURE BEAT (Sept. 28, 2020, 11:41 AM), <https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/> [<https://perma.cc/E5Y8-YJ2F>].

185. See Rubenstein, *supra* note 24, at 14–15; Schwartz et al., *supra* note 171, at 5 ("A consistent finding in the literature is the notion that trust can improve if the public is able to interrogate systems and engage with them in a more transparent manner.").

186. See Brauneis & Goodman, *supra* note 18, at 159 (expounding on this concern); Katyal, *supra* note 176, at 1259 (same).

187. U.S. GOV'T ACCOUNTABILITY OFF., GAO-21-519SP, *supra* note 58, at 9–10.

188. The use of "open" data and source code in AI system is common, and double-edged.

actors and institutions (if any) should be held accountable or to what extent.

### C. *The Rise of Ethical AI*

The foregoing challenges around AI safety, fairness, transparency, and accountability are beginning to register in social and political discourse. This reckoning may be credited to a series of high-profile AI episodes that do not require technical savvy to appreciate. In particular, the 2018 media coverage of the Cambridge Analytica–Facebook scandal was a watershed moment that exposed how AI ecosystems secretly exploit consumer data for commercial and political ends.<sup>189</sup>

The news was hardly surprising to a small cadre of academics, journalists, and industry insiders who—years prior—had foretold the dangers of digital surveillance and the power of AI to shape human behaviors.<sup>190</sup> When the Cambridge Analytica–Facebook scandal broke, however, “techlash” went mainstream.<sup>191</sup> As just one measure, only a

---

A.I. Now Inst., *A New AI Lexicon: OPEN*, MEDIUM (July 12, 2021), <https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-open-3ec7daa300a> [https://perma.cc/H33H-99EH?type=image] (“Despite the potential benefits of open data, there has been little research or discussion on the assumptions and applications of open data in the context of AI technologies, specifically how data is collected and made available.”).

189. See Alvin Chang, *The Facebook and Cambridge Analytica Scandal, Explained with a Simple Diagram*, VOX (May 2, 2018, 3:25 PM), <https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram> [https://perma.cc/2X9K-VGKQ] (discussing the shutdown of a political consulting firm that harvested user data from Facebook); Alex Hern, *Cambridge Analytica: How Did It Turn Clicks into Votes?*, GUARDIAN (May 6, 2018, 3:00 PM), <https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie> [https://perma.cc/CB57-RPRE].

190. See, e.g., WOLFIE CHRISTL & SARAH SPIEKERMANN, NETWORKS OF CONTROL: A REPORT ON CORPORATE SURVEILLANCE, DIGITAL TRACKING, BIG DATA & PRIVACY 7 (2016) (“While the media and special interest groups are aware of these developments for a while now, we believe that the full degree and scale of personal data collection, use and—in particular—abuse has not been scrutinized closely enough.”); PASQUALE, *supra* note 99, at 8–10; O’NEIL, *supra* note 18, at 13; WENDY HUI KYONG CHUN, CONTROL AND FREEDOM 1 (2008); Citron, *supra* note 32, at 1262; see also Shoshana Zuboff, *Surveillance Capitalism and the Challenge of Collective Action*, NEW LAB. F. (Jan. 24, 2019), <https://journals.sagepub.com/doi/full/10.1177/1095796018819461> [https://perma.cc/V5P6-PEQE].

191. Matthew Le Bui & Safiya Umoja Noble, *We’re Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness*, in THE OXFORD HANDBOOK OF ETHICS OF AI 163–67 (Markus D. Dubber et al. eds., 2020) (connecting the rise of techlash to the Cambridge Analytica–Facebook scandal); see also Rana Foroohar, *Year in a Word: Techlash*, FIN. TIMES (Dec. 16, 2018), <https://www.ft.com/content/76578fba-fca1-11e8-ac00-57a2a826423e> [https://perma.cc/HW8D-HG93] (defining “[t]echlash” as the “growing public animosity towards large Silicon Valley platform technology companies and their Chinese equivalents”).

handful of AI-related bills were pending in Congress in 2017.<sup>192</sup> Since then, more than 100 distinct pieces of AI-related bills have been introduced in Congress.<sup>193</sup> This trend is paralleled in U.S. state and local jurisdictions (and across the globe).<sup>194</sup>

## 1. Ethical AI in Industry

Ethical AI was not unheard of in 2017.<sup>195</sup> But its embrace as an industry *movement* occurred in 2018.<sup>196</sup> By 2019, a spate of ethical AI frameworks were promulgated or adopted by technology firms, trade groups, and non-government organizations.<sup>197</sup> While the particulars vary, ethical AI principles generally coalesce around a set of values relating to safety, fairness, transparency, accountability, privacy, and human well-being.<sup>198</sup>

---

192. See STAN. INST. FOR HUMAN-CENTERED A.I., ARTIFICIAL INTELLIGENCE INDEX REPORT 172 (2021); see also Yoon Chae, *U.S. AI Regulation Guide: Legislative Overview and Practical Considerations*, 3 J. ROBOTICS, A.I. & L. 17, 17 (2020) (reporting that from 2015–2016, only two bills were introduced that contained the term “artificial intelligence,” which increased to fifty-one bills by the end of 2019).

193. See STAN. INST. FOR HUMAN-CENTERED A.I., *supra* note 192, at 172; see also *AI Legislation Tracker—United States*, CTR. FOR DATA INNOVATION (June 19, 2020), <https://www.datainnovation.org/ai-policy-leadership/ai-legislation-tracker/> [https://perma.cc/E2U7-LEBQ].

194. See *Legislation Related to Artificial Intelligence*, NAT’L CONF. OF STATE LEGISLATURES (Apr. 16, 2021), <https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx> [https://perma.cc/T9US-2FWC] (tracking state AI-related legislation); *State Facial Recognition Policy*, ELEC. PRIV. INFO. CTR., <https://epic.org/state-policy/facialrecognition/> [https://perma.cc/UZL5-XLFL] (tracking state and local laws pertaining to facial recognition); *National AI Policies & Strategies*, OECD.AI (2021), <https://oecd.ai/en/dashboards> [https://perma.cc/AVX7-N75Z] (tracking global AI policies and strategies).

195. In 2016 and 2017, Apple, Amazon, Google, Facebook, Microsoft, IBM, joined to form the Partnership for Artificial Intelligence to Benefit People and Society. See Hern, *supra* note 43; James Vincent, *Apple Joins Research Group for Ethical AI with Fellow Tech Giants*, VERGE (Jan. 27, 2017, 7:02 AM), <https://www.theverge.com/2017/1/27/14411810/apple-joins-partnership-for-ai> [https://perma.cc/6R2V-HFA5]. Then, as now, the consortium’s express purpose is to develop industry best practices for promoting “ethics, fairness and inclusivity; transparency, privacy, and interoperability; collaboration between people and AI systems; and the trustworthiness, reliability and robustness of the technology.” See Hern, *supra* note 43.

196. Cf. STAN. INST. FOR HUMAN-CENTERED A.I., *supra* note 192, at 129 (“In terms of rolling out ethics principles, 2018 was the clear high-water mark for tech companies—including IBM, Google, and Facebook . . .”).

197. *Id.* at 129–30; JESSICA FJELD ET AL., PRINCIPLED ARTIFICIAL INTELLIGENCE: MAPPING CONSENSUS IN ETHICAL AND RIGHTS-BASED APPROACHES TO PRINCIPLES FOR AI, BERKMAN KLEIN CTR. FOR INTERNET & SOC’Y (2020), <http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420> [https://perma.cc/2A6N-N2HX].

198. See generally Jobin et al., *supra* note 40 (mapping and analyzing the corpus of principles and guidelines on ethical AI); FJELD ET AL., *supra* note 197.

The motivations driving the ethical AI movement are ideological and instrumental. No doubt, altruism and corporate social responsibility are playing a part.<sup>199</sup> Just as surely, ethical AI is a political pitch to forestall government regulation,<sup>200</sup> and a market pitch to placate consumers and investors.<sup>201</sup> But it must also be appreciated that ethical AI is a grassroots movement, curated and cultivated in significant part by the high-skilled, and highly in-demand, technology workforce.<sup>202</sup>

In 2018, thousands of technologists signed open letters and staged headline-generating protests, urging corporate leaders to end law enforcement and military contracts with the government.<sup>203</sup> Heeding the

---

199. See Andrew Charlesworth, *Regulating Algorithmic Assemblages: Looking Beyond Corporatist AI Ethics*, in DATA-DRIVEN PERSONALISATION IN MARKETS, POLITICS AND LAW 243, 245–46 (Uta Kohl & Jacob Eisler eds., 2021) (linking the proliferation of ethical AI frameworks in the technology industry to the corporate social responsibility movement). For a discussion of some of these initiatives, see JESSICA CUSSINS NEWMAN, DECISION POINTS IN AI GOVERNANCE (2020); Kathy Baxter, *Ethical Frameworks, Tool Kits, Principles, and Oaths—Oh My!*, SALESFORCE (Oct. 19, 2020), <https://blog.einstein.ai/frameworks-tool-kits-principles-and-oaths-oh-my/> [<https://perma.cc/898E-KLTQ>]. For examples of ethical AI toolkits, see *AI Fairness 360*, IBM RSCH. TRUSTED AI, <https://aif360.mybluemix.net/> [<https://perma.cc/EU5N-3J7S>]; SARAH BIRD ET AL., FAIRLEARN: A TOOLKIT FOR ASSESSING AND IMPROVING FAIRNESS IN AI 1 (2020), [https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn\\_WhitePaper-2020-09-22.pdf](https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf) [<https://perma.cc/77T6-JJAJ>]; Rachel K. E. Bellamy et al., AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. (Oct. 3, 2018) (unpublished manuscript), <https://arxiv.org/abs/1810.01943> [<https://perma.cc/Q4ZH-X4Z8>]; Google Research, *What if Tool* (2019), <https://pair-code.github.io/what-if-tool/> [<https://perma.cc/8UMW-J8R8>].

200. See Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 407–08 (2017) (noting that “ethically motivated ‘self-commitments’ can discourage policymakers from pursuing legally binding principles and constraints”); Orr & Davis, *supra* note 109, at 8 (“[O]rganizations and industry professionals have been careful to set their own standards to avoid control at the hands of non-expert forces.”); cf. Charlesworth, *supra* note 199, at 245 (“The establishment of ethics boards, ethics oversight committees and codes of practice for AI by corporate entities follows a familiar regulatory pattern, well established in the technology sphere, whereby industries seek to head off formal governmental regulatory intervention by providing putatively self-regulatory mechanisms to address the problematic impacts of their services or corporate activities.”).

201. See NEWMAN, *supra* note 199, at 14 (discussing how ethics committees are “viewed with some suspicion, and in some cases have been called out as ‘AI ethics-washing’” (quoting Karen Hao, *In 2020, Let’s Stop AI Ethics-Washing and Actually Do Something*, MIT TECH. REV. (Dec. 27, 2019))).

202. Nataliya Nedzhvetskava & JS Tan, *What We Learned From over a Decade of Tech Activism*, GUARDIAN (Dec. 23, 2019), <https://www.theguardian.com/commentisfree/2019/dec/22/tech-worker-activism-2019-what-we-learned> [<https://perma.cc/M6H3-HQTX>] (discussing the rise of activism within the technology industry).

203. See NEWMAN, *supra* note 199, at 18 (“A group called Microsoft Workers 4 Good, whose mission is ‘to empower every worker to hold Microsoft accountable to their stated values,’ has called on Microsoft leadership to end certain contracts.”); Daisuke Wakabayashi & Scott Shane, *Google Will Not Renew Pentagon Contract That Upset Employees*, N.Y. TIMES (June 1, 2018),

call, Google declined to renew its contract with the Department of Defense (DoD) on Project Maven (which uses AI for drone strikes), and declined to compete for a major DoD cloud-computing contract (which was worth up to \$10 billion).<sup>204</sup> More recently, in the Summer of 2020, three leading technology firms stopped selling facial recognition technology to law enforcement agencies.<sup>205</sup> In a telling letter to Congress, IBM explained that it “will not condone uses of any technology . . . for mass surveillance, racial profiling, violations of basic human rights and freedoms, or any purpose which is not consistent with [IBM’s] values and Principles of Trust and Transparency.”<sup>206</sup>

More than ironic, these corporate displays of social responsibility are instructive here for three related reasons. First, the government is susceptible to techlash, including from technologists. Second, ethical AI speaks loudly in the market, and the government market is no exception. Third, public anxieties around AI systems will not be neatly cabined into government and commercial spheres. Nor should the polity draw sharp distinctions, given that the AI technologies used in the private sector are, by and large, the same technologies deployed for government functions.

## 2. Ethical AI in Government

In 2019, the White House issued an Executive Order that sketched an agenda for “[m]aintaining American leadership” in innovative and trustworthy AI.<sup>207</sup> Soon after, the United States joined with other global leaders to adopt a set of “[p]rinciples for responsible stewardship of

---

<https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html> [<https://perma.cc/3PA7-XPVE>] (“About 4,000 Google employees signed a petition demanding ‘a clear policy stating that neither Google nor its contractors will ever build warfare technology.’ A handful of employees also resigned in protest, while some were openly advocating the company to cancel the Maven contract.”).

204. See Wakabayashi & Shane, *supra* note 203. In the wake of protracted litigation, the DoD cancelled the contract in 2021. See Kate Conger & David E. Sanger, *Pentagon Cancels a \$10 Billion Technology Contract*, N.Y. TIMES (July 6, 2021, 12:52 PM), <https://www.nytimes.com/2021/07/06/technology/JEDI-contract-cancelled.html> [<https://perma.cc/7W3S-PKBC>].

205. See Jay Greene, *Microsoft Won’t Sell Police Its Facial-Recognition Technology, Following Similar Moves by Amazon and IBM*, WASH. POST (June 11, 2020, 2:30 PM), <https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/> [<https://perma.cc/6CJ2-VDKJ>]; *We Are Implementing a One-Year Moratorium on Police Use of Rekognition*, DAY ONE: AMAZON BLOG (June 10, 2020), <https://blog.aboutamazon.com/policy/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition> [<https://perma.cc/Q2U-F5PZH>].

206. Arvind Krishna, *IBM CEO’s Letter to Congress on Racial Justice Reform*, IBM (June 8, 2020), <https://www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms/> [<https://perma.cc/BD5G-5R8R>].

207. Exec. Order No. 13,859, 84 Fed. Reg. 3967, 3967 (Feb. 14, 2019).

trustworthy AI.”<sup>208</sup> These principles, promulgated by the Organization for Economic Cooperation and Development (OECD), were the first intergovernmental standards on AI.<sup>209</sup> Although the OECD framework is not binding on member states, the core ethical AI principles are beginning to germinate in U.S. policy.

In 2020, the DoD and U.S. Intelligence Community formally adopted ethical AI principals.<sup>210</sup> Later that year, the White House issued another Executive Order, with the eponymous aim of “Promoting the Use of Trustworthy [AI] in the Federal Government.”<sup>211</sup> Like the foregoing ethical AI initiatives, this White House directive espouses principles relating to safety, fairness, transparency, and accountability.<sup>212</sup> Moreover, it instructs agencies to “design, develop, acquire, and use AI in a manner that exhibits due respect for our Nation’s values and is consistent with the Constitution and all other applicable laws and policies, including those addressing privacy, civil rights, and civil liberties.”<sup>213</sup>

Thus far, Congress has been slow to act on a multitude of pending AI-related bills.<sup>214</sup> However, the National Defense Authorization Act of 2021 provides an early glimpse of Congress’s wide bipartisan support for responsible AI uses by government and industry alike.<sup>215</sup> Most pertinent here, the Act directs NIST to support the development of technical

---

208. *Recommendation of the Council on Artificial Intelligence*, OECD LEGAL INSTRUMENTS (May 21, 2019), <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> [<https://perma.cc/6MAK-PGG2>] (espousing (1) inclusive growth, sustainable development and well-being; (2) human-centered values and fairness; (3) transparency and explainability; (4) robustness, security and safety; and (5) accountability); *see also* Michael Kratsios, *White House OSTP’s Michael Kratsios Keynote on AI Next Steps*, U.S. MISSION TO ORG. FOR ECON. COOP. & DEV. (May 21, 2019), <https://usoecd.usmission.gov/white-house-ostps-michael-kratsios-keynote-on-ai-next-steps/> [<https://perma.cc/2QZ8-L2MT>] (discussing the principles adopted by the OECD).

209. *Recommendation of the Council on Artificial Intelligence*, *supra* note 208.

210. Press Release, U.S. Dep’t of Def., DOD Adopts Ethical Principles for Artificial Intelligence (Feb. 24, 2020), <https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence> [<https://perma.cc/X82V-SLHD>]; *see also* Hicks, *supra* note 7, at 1 (“As the DoD embraces [AI], it is imperative that we adopt responsible behavior, processes, and outcomes in a manner that reflects the Department’s commitment to its ethical principles, including the protection of privacy and civil liberties.”).

211. *See* Exec. Order No. 13,960, 85 Fed. Reg. 78,939, 78,939 (Dec. 8, 2020).

212. *Id.* at 78,940–41.

213. *Id.* at 78,940. The executive order also establishes a common policy for implementing the principles, instructs agencies to catalog their AI use cases, and directs the General Services Administration and the Office of Personnel Management to enhance AI implementation expertise within the executive branch. *Id.* at 78,941–43.

214. *See supra* notes 191–92 and accompanying text.

215. Pub. L. No. 116–283; *see also* Summary of AI Provisions from the National Defense Authorization Act 2021, STAN. INST. FOR HUMAN-CENTERED A.I., <https://hai.stanford.edu/policy-policy-resources/summary-ai-provisions-national-defense-authorization-act-2021> [<https://perma.cc/LB64-QVWF>] (providing a summary of AI-related provisions from the Act).

standards, guidelines, and risk-management frameworks to promote “trustworthy” AI systems.<sup>216</sup> The Act also creates a new National AI Initiative Office (tasked with coordinating federal AI activities and supporting AI research),<sup>217</sup> and a National AI Advisory Committee (which will advise the President on a range of matters pertaining to ethical AI, including the use of facial recognition by law enforcement authorities).<sup>218</sup> Undoubtedly, this is just the opening salvo of much more federal policymaking and oversight to come.

### III. FROM PRINCIPLES TO PRACTICE

The proliferation of ethical AI principles, along with the government’s high-level support, are generally viewed as steps in the right direction.<sup>219</sup> As this Part explains, however, ethical AI is easier said than done. Most assuredly, the extant frameworks and declarations of ethical AI do not address, much less resolve, an open set of challenges and tradeoffs at the fulcrum of law, society, and technology.<sup>220</sup> This Part spotlights those tensions and their implications for algorithmic governance.

#### A. *The Gap Between Ethical AI Principles and Practice*

This first Section homes in on the impediments to ethical AI in workaday practice. The challenges manifest somewhat differently within industry and across the public/private divide. To tease out some of those differences, the discussion begins with industry before turning to the government. This ordering tracks reality: the government’s AI journey is effectively bootstrapped to industry, and the government’s ethical AI challenges are mostly derivative.

---

216. National Defense Authorization Act of 2021 § 5301.

217. *Id.* §§ 5101–03.

218. *Id.* § 5104.

219. To say that these are steps in the right direction is not to say they are sufficient. See Lee Rainie et al., *Experts Doubt Ethical AI Design Will Be Broadly Adopted as the Norm Within the Next Decade*, PEW RSCH. CTR. (June 16, 2021), <https://www.pewresearch.org/internet/2021/06/16/experts-doubt-ethical-ai-design-will-be-broadly-adopted-as-the-norm-within-the-next-decade/> [https://perma.cc/3ST9-789D]; JONATHAN ROTNER, HOW CAN ETHICS MAKE BETTER AI PRODUCTS? 6 (2020) (“Skeptics might see declarations, frameworks, and toolkits as virtue signaling, resulting in words without action.” (footnote omitted)).

220. See *infra* Section III.A.1 (discussing a range of challenges and gaps between ethical AI principles and practice); Raji et al., *supra* note 96, at 2 (“The AI industry lacks proven methods to translate principles into practice.”); see also *infra* Section III.A.2 (discussing the government’s reliance on industry for translating ethical AI principles into practice).

## 1. Industry Challenges

To start, the voluntary nature of ethical AI allows competing market incentives to dominate.<sup>221</sup> For instance: if the choice is between algorithmic auditing and rushing a product to market, many if not most firms will choose the latter. To be clear, some firms may choose ethical AI principles over profits and growth. But most firms don't because they don't have to.

The principles-to-practice challenge is exacerbated by the AI ecosystem's distributed network of responsibilities and domain expertise.<sup>222</sup> For example, data scientists and software engineers may not anticipate or feel responsible for the social impacts of their digital creations.<sup>223</sup> Meanwhile, social scientists and lawyers may not appreciate the technical challenge of translating nebulous concepts like fairness and nondiscrimination into code.<sup>224</sup> More generally, "the ethical development and deployment of AI systems typically involves decisions that no individual practitioner can make on their own."<sup>225</sup> Consequently, the amount of influence or responsibility that any individual has might be preempted or superseded by others in the AI pipeline. For example,

---

221. See Charlesworth, *supra* note 199, at 2 ("‘AI ethics’ is far removed from ‘AI law’ and broadly captures the idea of self-policing by private corporate actors in their use of AI systems, as sanctioned by government."); Michael A. Madaio et al., *Co-Designing Checklists to Understand Organizational Challenges and Opportunities Around Fairness in AI*, in CHI '20: PROCEEDINGS OF THE 2020 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1, 10 (2020) ("[O]rganizational culture typically prioritizes ‘moving fast’ and shipping products over pausing to consider fairness."); see also Schwartz et al., *supra* note 171, at 4 ("Often a technology is not tested—or not tested extensively—before deployment, and instead deployment may be used as testing for the technology."); Emanuel Moss & Jacob Metcalf, *The Ethical Dilemma at the Heart of Big Tech Companies*, HARV. BUS. REV., Nov. 14, 2019, <https://hbr.org/2019/11/the-ethical-dilemma-at-the-heart-of-big-tech-companies> [https://perma.cc/B95N-EDGC] (highlighting the tension between the race to market and the race to ethical AI).

222. See, e.g., Kenneth Holstein et al., *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?*, in CHI 2019: PROCEEDINGS OF THE 2019 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, at 5–12 (2019), <https://arxiv.org/pdf/1812.05239.pdf> [https://perma.cc/HV4D-Q8AC] (assessing the practical needs of private sector AI practitioners in relation to ethical AI); Orr & Davis, *supra* note 109, at 10 ("[P]ractitioners play the part of (highly skilled) technicians, rather than morally autonomous agents."); see also Michael Veale et al., *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, in CHI 2018: PROCEEDINGS OF THE 2018 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, at 3–7 (2018), <https://arxiv.org/pdf/1802.01029.pdf> [https://perma.cc/8FPA-A8GW] (surveying public sector AI practitioners and cataloguing challenges they face achieving fairness standards).

223. See Orr & Davis, *supra* note 109, at 9 (describing the disconnect between practitioners and those that commission their work).

224. See KEARNS & ROTH, *supra* note 74, at 18 ("[T]he first challenge in asking an algorithm to be fair or private is agreeing on what those words should mean . . . in so precise a manner that they can be ‘explained’ to a machine.").

225. See Madaio et al., *supra* note 221, at 2.

system engineers may have no control over the decisions of subject matter experts, and vice versa. Sales representatives may be informed about system deficiencies but bury those problems in market pitches. Last, but not least, corporate leadership can ignore, marginalize, or terminate ethical AI champions that do not tow the company bottom line.<sup>226</sup>

Even under the right conditions, ethical AI frameworks are too generalized to resolve more specific issues that commonly arise in practice.<sup>227</sup> When ethical AI principles collide, the problem can be particularly acute.<sup>228</sup> For example, ethical AI frameworks generally do not resolve conflicts between AI safety and transparency, transparency and accuracy, accuracy and fairness, and so on.<sup>229</sup> Nor do the frameworks resolve incoherencies within specific principles.<sup>230</sup> Fairness, for example, has upwards of a dozen formulations in the AI field.<sup>231</sup> For virtually all

---

226. Unfortunately, the marginalization of ethical AI voices within industry is reportedly common. See Madaio et al., *supra* note 221, at 5 (reporting that “[i]ndividual advocates [for AI fairness] face both sociocultural barriers to speaking up and structural barriers to having their teams address AI fairness issues”); see also *id.* at 6 (“[T]he disconnect arising from rhetorical support for AI fairness efforts coupled with a lack of organizational incentives that support such efforts is a central challenge for practitioners.”). Recently, Google made national headlines when it ousted the co-leads of its ethical AI team, Timnit Gebru and Melanie Mitchell. See *Google to Change Research Process After Uproar Over Scientists’ Firing*, GUARDIAN (Feb. 26, 2021, 2:32 PM), <https://www.theguardian.com/technology/2021/feb/26/google-timnit-gebru-margaret-mitchell-ai-research> [<https://perma.cc/A7KL-Z737>]. This was an especially shocking display of capitalism cancelling ethical AI. See Alex Hanna & Meredith Whitaker, *Timnit Gebru’s Exit from Google Exposes a Crisis in AI*, WIRED (Dec. 31, 2020, 7:00 AM), <https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/?redirectURL=https%3A%2F%2Fwww.wired.com%2Fstory%2Ftimnit-gebru-exit-google-exposes-crisis-in-ai%2F> [<https://perma.cc/2DFG-5RZD>]; Tom Simonite, *What Really Happened When Google Ousted Timnit Gebru*, WIRED (June 8, 2021, 6:00 AM), <https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/> [<https://perma.cc/F2BG-YAN7>] (providing an in-depth, behind-the-scenes account).

227. Brent Mittelstadt, *Principles Alone Cannot Guarantee Ethical AI*, 1 NATURE MACH. INTEL. 501, 504 (2019) (“Norms and requirements cannot be deduced directly from mid-level principles without accounting for specific elements of the technology, application, context of use, or relevant local norms.”); Raji et al., *supra* note 96, at 2 (noting that “AI principles have been criticized for being vague and providing little to no means of accountability”).

228. Cf. Mittelstadt, *supra* note 227, at 504; Madaio et al., *supra* note 221, at 2 (“AI ethics principles can fail to achieve their intended goal if they are not accompanied by other mechanisms for ensuring that practitioners make ethical decisions.”).

229. Jess Whittlestone et al., *The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions*, in AIES ‘19: PROCEEDINGS OF THE 2019 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 196–97 (2019), <https://dl.acm.org/doi/10.1145/3306618.3314289> [<https://perma.cc/NL4A-9BKA>].

230. See *id.*

231. See Arvind Narayanan, *Translation Tutorial: 21 Fairness Definitions and Their Politics*, YOUTUBE (Mar. 1, 2018), <https://www.youtube.com/watch?v=jIXIuYdnyyk> [<https://perma.cc/ZX3Q-GJV6>]; see also Jacobs & Wallach, *supra* note 145, at 382–83 (discussing and distinguishing conceptions of “individual fairness” and “group fairness”).

data distributions, however, it is mathematically impossible to simultaneously satisfy the three most commonly used fairness metrics.<sup>232</sup> Likewise, transparency and accountability are flexible ideals; there are different conceptions, expressions, and degrees of each.<sup>233</sup> Ethical AI frameworks could be more prescriptive and precise (and arguably should be). The claim here, however, is purely descriptive: ethical AI practice is unavoidably noisy and uneven because of the generality and incoherence of the frameworks themselves.

To some extent, AI systems can be ethically designed at inception with disciplined procedures and protocols.<sup>234</sup> The curation of “datasheets for datasets,”<sup>235</sup> “model cards for model reporting,”<sup>236</sup> and “fairness checklists”<sup>237</sup> are examples of responsible design practices. Moreover,

---

232. See Jon Kleinberg et al., *Inherent Trade-offs in the Fair Determination of Risk Scores*, in PROCEEDINGS OF 8TH INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE (2016), <https://arxiv.org/pdf/1609.05807.pdf> [<https://perma.cc/YHZ5-XSD4>]; see also Kailash Karthik Saravananakumar, *The Impossibility Theorem of Machine Fairness: A Casual Perspective* (Jan. 29, 2021) (preprint), <https://arxiv.org/pdf/2007.06024.pdf> [<https://perma.cc/PK9E-3R2Y>]; KEARNS & ROTH, *supra* note 74, at 85 (“There are certain combinations of fairness criteria that—although they are each individually reasonable—simply cannot be achieved simultaneously, even if we ignore accuracy considerations.”).

233. See, e.g., Engstrom & Ho, *supra* note 117, at 61 (discussing different types and conceptions of transparency in the AI literature); Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 9–11 (2017) (comparing “technical accountability” to its legal and political forms).

234. See INST. OF ELEC. & ELECS. ENG’RS, IEEE P7000: DRAFT STANDARD FOR MODEL PROCESS FOR ADDRESSING ETHICAL CONCERNs DURING SYSTEM DESIGN (2020); IBM, EVERYDAY ETHICS FOR ARTIFICIAL INTELLIGENCE 6 (2019) (“Ethics must be embedded in the design and development process from the very beginning of AI creation.”); WORLD ECON. F., ETHICS BY DESIGN: AN ORGANIZATIONAL APPROACH TO RESPONSIBLE USE OF TECHNOLOGY 6–8 (Dec. 2020) (discussing ethical design principles); see also BATYA FRIEDMAN & DAVID G. HENDRY, VALUE SENSITIVE DESIGN: SHAPING TECHNOLOGY WITH MORAL IMAGINATION 1 (2019) (“[A]ctively engaging with values in the design process offers creative opportunities for technical innovation as well as for improving the human condition.”).

235. Timnit Gebru et al., *Datasheets for Datasets*, ARXIV 6 (2020), <https://arxiv.org/pdf/1803.09010.pdf> [<https://perma.cc/YW29-LGTG>] (proposing the use of these instruments to provide information about the providence of data used to train and develop an AI system).

236. Margaret Mitchell et al., *Model Cards for Model Reporting*, in FACCT ’19: PROCEEDINGS OF THE 2019 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 220 (2019), <https://dl.acm.org/doi/pdf/10.1145/3287560.3287596> [<https://perma.cc/KN6J-9CL5>] (proposing the use of model cards that provide information about the intended use of the model, along with its known limitations and risks); see also Galen Harrison et al., *Towards Supporting and Documenting Algorithmic Fairness in the Data Science Workflow*, WORKSHOP ON TECH. & CONSUMER PROTECTION (May 23, 2019), <https://www.ieee-security.org/TC/SPW2019/ConPro/papers/harrison-conpro19.pdf> [<https://perma.cc/VPZ9-B4UN>] (proposing documentation and visualization of algorithms in data science processes).

237. See generally Madaio et al., *supra* note 221 (proposing the use of structured

some ethical AI concerns may be ameliorated with technical patches and toolkits.<sup>238</sup> But it is a categorical error to treat ethical AI as a suite of problems that can be resolved with procedural protocols and technical solutions.<sup>239</sup> To be sure, computers can help. The point here, however, is that many AI challenges cannot, or should not, be resolved by computers.

For example, software tools that detect AI bias and optimize for fairness are readily available.<sup>240</sup> Still, humans must determine which fairness metrics to utilize in which contexts.<sup>241</sup> Explainable AI software is another example of faux techno-solutionism. This software may provide useful insights to data scientists for purposes of model training and evaluation; however, those insights may be meaningless or unsatisfactory for end users, auditors, adjudicators, and policymakers.<sup>242</sup> Even more concerning, studies show that explainable AI tools can be manipulated or misleading.<sup>243</sup> Needless to say, if these tools or their uses are untrustworthy, then the AI explanations will be untrustworthy too.

The burgeoning field of algorithmic auditing offers additional promise for actualizing and incenting ethical AI.<sup>244</sup> Such audits can be internal or

---

considerations pertaining to AI fairness to instigate dialogue and deliberation during AI development).

238. See, e.g., *supra* note 199 and accompanying text (referencing AI debiasing tools); see also Kroll et al., *supra* note 12, at 636–41 (providing a computer scientist’s perspective on algorithmic accountability and calling for specific tailored solutions); ASS’N FOR COMPUTING MACH. U.S. PUB. POL’Y COUNCIL, STATEMENT ON ALGORITHMIC TRANSPARENCY AND ACCOUNTABILITY 2 (2017), [http://www.acm.org/binaries/content/assets/public-policy/2017\\_usa\\_cm\\_statement\\_algorithms.pdf](http://www.acm.org/binaries/content/assets/public-policy/2017_usa_cm_statement_algorithms.pdf) [<https://perma.cc/7U6F-ETB6>].

239. See Coal. for Critical Tech., *Abolish the #TechToPrisonPipeline: Crime Prediction Technology Reproduces Injustices and Causes Real Harm*, MEDIUM (June 23, 2020), <https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16> [<https://perma.cc/93EQ-3AKD>] (“To date, many efforts to deal with the ethical stakes of algorithmic systems have centered mathematical definitions of fairness that are grounded in narrow notions of bias and accuracy. These efforts give the appearance of rigor, while distracting from more fundamental epistemic problems.”).

240. See *supra* note 199 and accompanying text.

241. See BIRD ET AL., *supra* note 199, at 2 (“Because fairness in AI is a sociotechnical challenge, there is no software tool that will ‘solve’ fairness in all AI systems.”); KEARNS & ROTH, *supra* note 74, at 63 (“Good algorithms can specify a menu of solutions, but people still have to pick one of them.”); see also Coal. for Critical Tech., *supra* note 239.

242. See *Explainable AI*, THE ROYAL SOC’Y 12 (2019), <https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf> [<https://perma.cc/R6XE-3N36>].

243. See Himabindu Lakkaraju & Osbert Bastani, “*How Do I Fool You??: Manipulating User Trust Via Misleading Black Box Explanations*, in AIES ’20: PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 79, 85 (2020), <https://dl.acm.org/doi/pdf/10.1145/3375627.3375833> [<https://perma.cc/WJ6H-HBJ7>] (empirically establishing how user trust in black-box AI models can be manipulated by misleading explanations).

244. See, e.g., Raji et al., *supra* note 96, at 1 (introducing “a framework for algorithmic auditing that supports artificial intelligence system development end-to-end, to be applied

external. In both settings, algorithmic auditing generally entails the inspection of technical and non-technical aspects of AI systems.<sup>245</sup> Under the right conditions and constraints, algorithmic audits can be highly beneficial. But under current conditions, the constraints are neither standardized nor regularized. As such, the reliability and social value of algorithmic audits are highly contingent. Certainly, the opacity and partiality of some audits have prompted justified concern that the process may be exploited to legitimize dubious AI systems.<sup>246</sup> Because the nascent AI auditing industry is unregulated, the audits themselves lack the patina of legitimacy enjoyed in more mature markets.<sup>247</sup>

All told, the disciplined practice of ethical AI requires time, resources, and institutional buy-in that many firms may not have—or feel the need to have—unless compelled by market forces or binding norms.<sup>248</sup> Moreover, to greater and lesser extents, firms will externalize the costs of

---

throughout the internal organization development lifecycle”); Jennifer Cobbe et al., *Reviewable Automated Decision-Making*, COMPUT. L. & SEC. REV., Nov. 2020, at 1 (calling for a “reviewability framework” to promote accountability); MILES BRUNDAGE ET AL., TOWARD TRUSTWORTHY AI DEVELOPMENT: MECHANISMS FOR SUPPORTING VERIFIABLE CLAIMS 8–10 (2020); James Guszcza et al., *Why We Need to Audit Algorithms*, HARV. BUS. REV., Nov. 28, 2018, <https://hbr.org/2018/11/why-we-need-to-audit-algorithms> [https://perma.cc/8ZPJ-9EQW]; Rumman Chowdhury & Narendra Mulani, *Auditing Algorithms for Bias*, HARV. BUS. REV., Oct. 24, 2018, <https://hbr.org/2018/10/auditing-algorithms-for-bias> [https://perma.cc/ULJ5-FSDR] (discussing a fairness tool to audit outcomes developed by Accenture); Bruneis & Goodman, *supra* note 18, at 339 (identifying eight criteria that developers would need to identify for external review, including: the predictive goals of the algorithm and the problem it is meant to solve; the training data considered relevant to reach the predictive goal; the training data excluded and the reasons for excluding it; the actual predictions of the algorithm as opposed to its predictive goals; the analytical techniques used to discover patterns in the data; other policy choices encoded in the algorithm besides data exclusion; validation studies or audits of the algorithm after implementation; and a plain language explanation of how the algorithm makes predictions); *see also* INST. OF INTERNAL AUDITORS, GLOBAL PERSPECTIVES AND INSIGHTS: THE IIA’S ARTIFICIAL INTELLIGENCE AUDITING FRAMEWORK 2–3 (2017), <https://na.theiia.org/periodicals/Public%20Documents/GPI-Artificial-Intelligence-Part-II.pdf> [https://perma.cc/6U59-SMAR].

245. See, e.g., Shea Brown et al., *The Algorithm Audit: Scoring the Algorithms that Score Us*, BIG DATA & SOC’Y, Jan.–June 2021, at 2.

246. See Mona Sloane, *The Algorithmic Auditing Trap*, MEDIUM (Mar. 17, 2021), <https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d> [https://perma.cc/DY9T-SVJD].

247. Cf. Raji et al., *supra* note 96, at 4, 10 (comparing the unsystematized nature of AI audits to the maturity of other auditing systems); BRUNDAGE ET AL., *supra* note 244, at 25 (noting that auditing “standards are not yet established for AI systems”).

248. Cf. Orr & Davis, *supra* note 109, at 8 (“Giving primacy to legal mandates renders ethical considerations a relative luxury—something ‘nice to think about,’ but ultimately subservient to the formal codes and regulations in place.”); Rainie et al., *supra* note 219, at 4 (reporting on widespread concern among AI experts that “main developers and deployers of AI are focused on profit-seeking and social control, and there is no consensus about what ethical AI would look like”).

AI risks to clients, consumers, or communities of (un)interest. If this is a failure, then it is both a market failure and a regulatory failure to fix.

## 2. Government Challenges

The government, for its part, faces many of the same challenges as industry and arguably more. True, the government is relieved of the industry's duty to shareholders and market drivers. Yet the government has sovereign duties that offset the difference along all the relevant dimensions: safety, fairness, transparency, and accountability. That is not to say that technology firms have free rein. But it is to acknowledge the asymmetrical laws and expectations that attach to public and private action.<sup>249</sup> Those distinctions are important, insofar as they entail special government responsibilities.

But the immediate focus here is another asymmetry: namely, the government's market dependency on industry to supply the tools of algorithmic governance.

In theory, when the government makes sourcing decisions, “[it] can either hire and train personnel and assemble the raw materials needed to perform government tasks, or it can contract through the procurement process to buy them.”<sup>250</sup> Far from a typical “build-or-buy” decision, however, the government has nowhere near the in-house capacity to build and field AI systems at scale.<sup>251</sup> While AI prototypes and pilot programs are plentiful in some agencies, the government is in short supply of the technical resources and know-how required for enterprise-level AI ideation, development, integration, deployment, monitoring, and sustainment.<sup>252</sup>

The government's technological debt profoundly affects how agency demand for ethical AI solutions will be fulfilled, or perhaps unfulfilled. As a threshold matter, the industry's commercially oriented research agenda only partially aligns with the government's needs and

---

249. See Crawford & Schultz, *supra* note 29, at 1944; see also *infra* Section II.A.4.

250. See ACUS REPORT, *supra* note 1, at 88; see also Oliver E. Williamson, *Public and Private Bureaucracies: A Transaction Cost Economics Perspective*, 15 J.L. ECON. & ORG. 306, 319 (1999) (discussing make-or-buy sourcing decisions).

251. See NAT'L SEC. COMM'N ON A.I., INTERIM REPORT 22 (2019) [hereinafter NSCAI INTERIM REPORT] (“Despite pockets of excellence, the government lacks wide expertise to envision the promise and implications of AI, translate vision into action, and develop the operating concepts for using AI.”).

252. NSCAI FINAL REPORT, *supra* note 11, at 32 (“Successful development and fielding of AI technologies depends on a number of interrelated elements that can be envisioned as a stack,” the integration of which “can be daunting and historically has been underestimated.”); see also MARK TREVEIL ET AL., INTRODUCING MLOPS: HOW TO SCALE MACHINE LEARNING IN THE ENTERPRISE 4 (2020) (explaining that the “machine learning life cycle in an enterprise setting is much more complex, in terms of needs and tooling”).

responsibilities.<sup>253</sup> Thus, commercially available AI tools and services may not exist, or may be unsuitable for government use. When AI solutions are available, customer agencies may be unaware of the functional limits, biases, and value judgments embedded in acquired AI systems. Indeed, as earlier discussed, information about a vendor's design choices may be legally or contractually insulated from disclosure.<sup>254</sup> Meanwhile, on the supply side, vendors may not fully appreciate the government's legal constraints, institutional protocols, or use contexts in which acquired AI tools will be deployed.

To a considerable extent, the informational asymmetries may be overcome (Part IV makes recommendations for how). Yet, more broadly, the government's market dependencies for AI solutions are structurally entrenched and hard to rectify; certainly not on a timescale that matches the government's projected demand for ethical AI solutions. This predicament is the culmination of decades of structural reforms that are coming home to roost.

Widespread bipartisan support to "shrink" and "reinvent" the federal government in the 1990s was propelled by aspirations to make government more business-like and efficient.<sup>255</sup> These reforms yielded certain successes, but they drained the federal workforce.<sup>256</sup> Decades of federal hiring caps, cuts, and freezes have left the federal government

---

253. See REBECCA GELLES ET AL., CTR. FOR SEC. & EMERGING TECH., MAPPING RESEARCH AGENDAS IN U.S. CORPORATE AI LABORATORIES 3 (2021), <https://cset.georgetown.edu/wp-content/uploads/CSET-Mapping-Research-Agendas-in-U.S.-Corporate-AI-Laboratories.pdf> [https://perma.cc/T6XL-ANHU] (finding a "potential mismatch" between "private research investments and national priorities"); *id.* at 8 ("The major private labs that have invested aggressively in [machine learning] in recent years may not be investing in the specific areas that are most beneficial to the overall U.S. position in the technology.").

254. See *supra* notes 172–78 and accompanying text (discussing transparency challenges relating to commercial trade secrecy).

255. See Guttman, *supra* note 28, at 881–90 (discussing the ideological and political shift toward federal outsourcing in the mid-to-late twentieth century); Steven L. Schooner, *Fear of Oversight: The Fundamental Failure of Businesslike Government*, 50 AM. U. L. REV. 627, 636 (2001) ("The mid-1990s witnessed a tsunami of procurement reforms heralded as the most successful aspect of [Vice President] Gore's reinventing government initiative, which were intended to make the procurement system less bureaucratic and more businesslike."); Steven Kelman, *Strategic Contracting Management*, in MARKET-BASED GOVERNANCE 88, 89–91 (John D. Donahue & Joseph S. Nye Jr. eds., 2002). See generally AL GORE, *CREATING A GOVERNMENT THAT WORKS BETTER & COSTS LESS* (1993) (describing the Clinton administration's plans to reduce regulatory barriers and governmental waste).

256. Dan Guttman, *Governance by Contract: Constitutional Visions; Time for Reflection and Choice*, 33 PUB. CONT. L.J. 321, 324–25 (2004) (discussing the challenges associated with the privatization of the federal workforce); see also Shelly Roberts Econom, *Confronting the Looming Crisis in the Federal Acquisition Workforce*, 35 PUB. CONT. L.J. 171, 189 (2006); PAUL R. VERKUIL, *OUTSOURCING SOVEREIGNTY* 162 (2007).

with little choice but to use contract and grant employees to achieve its goals.<sup>257</sup>

Over the same stretch, federal spending on research and development for new technologies declined precipitously.<sup>258</sup> The technological waves that ushered in home computers, pocket computers, and the internet-of-things, were mostly sourced with private capital, free and clear of government rights.<sup>259</sup> Even with the recent uptick in federally spending on AI research and development, private capital investments still “dwarf” federal funding.<sup>260</sup> Consequently, the government is in “perpetual catch-up mode,” with “limited control over how AI technologies are developed, shared, and used.”<sup>261</sup>

None of this is lost on the government, which remains clear-eyed about its in-house capacity challenges. Recently, the government has established several programs to recruit and build an AI workforce,<sup>262</sup> and

---

257. See PAUL C. LIGHT, THE GOVERNMENT INDUSTRIAL COMPLEX 65–69 (2019) (explaining how increased government services, combined with personnel ceilings, led to a rise in private contracting by the government).

258. Federal R&D dropped from a height of near 1.9% of GDP in 1964 to just 0.62% in 2018. Anne Q. Hoy, *Increases in U.S. Federal R&D Needed in a Global Crisis*, AM. ASS’N FOR ADVANCEMENT SCI. (Aug. 31, 2020), <https://www.aaas.org/news/increases-us-federal-rd-needed-global-crisis> [https://perma.cc/522L-LBU3].

259. See NAT’L RCH. COUNCIL, FUNDING A REVOLUTION: GOVERNMENT SUPPORT FOR COMPUTER RESEARCH 179–81 (1999).

260. NSCAI, INTERIM REPORT, *supra* note 20, at 15–16.

261. *Id.* at 15.

262. For example, the General Services Administration (GSA) offers a variety of programs including: “18F” (“[a] digital consulting office that partners with agencies to help them build or buy digital services”) and “IT Modernization Centers of Excellence” (“[a] centralized team of technical experts that accelerate agency-wide IT modernization”). See *Technology Transformation Services*, GEN. SERVS. ADMIN., <https://www.gsa.gov/about-us/organization/federal-acquisition-service/technology-transformation-services> [https://perma.cc/KB6H-JJH7]. GSA also hosts a Presidential Innovation Fellowship, which is “[a] program that pairs top technologists with civil-servants to spend 12 months tackling some of our nation’s biggest challenges.” *Id.* Tellingly, these capacity building programs lean on private industry. For example, the GSA’s AI Center for Excellence is designed to provide acquisition consulting and assistance to agencies on a government-wide basis. See Kathleen Walch, *How The Federal Government’s AI Center of Excellence Is Impacting Government-Wide Adoption of AI*, FORBES (Aug. 8, 2020, 1:00 PM), <https://www.forbes.com/sites/cognitiveworld/2020/08/08/how-the-federal-governments-ai-center-of-excellence-is-impacting-government-wide-adoption-of-ai/#7da611206660> [https://perma.cc/D5TW-CYYK]. Additionally, “18F” offers a centralized team of private-sector technology experts to consult and work with agencies on specific projects. See *About 18F*, GEN. SERVS. ADMIN., <https://18f.gsa.gov/about/> [https://perma.cc/QD9D-9AKV]. Likewise, DoD’s Joint Artificial Intelligence Center (JAIC) has transitioned from a product building unit into an AI training, acquisition, and platform hub. Jackson Barnett, *“JAIC 2.0” Moves Away From Building Products to Focus on DOD-wide AI Transformation*, FEDSCOOP (Nov. 6, 2020), <https://www.fedscoop.com/jaic-2-0-moving-away-from-products-artificial-intelligence/> [https://perma.cc/U5SS-VM7R]; Joint A.I. Ctr., *JAIC Completes Responsible AI Champions Pilot*, AI IN

is piloting programs for AI acquisition reform.<sup>263</sup> As matters currently stand, however, “the vast majority of IT leaders say their agency is struggling to incorporate AI into overall IT operations.”<sup>264</sup>

### B. The Gap Between Ethical AI and Algorithmic Governance

A rich scholarship has emerged to square AI systems with U.S. legal structures, institutions, and democratic norms that have not yet been coded for algorithmic governance.<sup>265</sup> As Aziz Huq explains, “present doctrinal formulations” do not necessarily mesh with, or address, a range of constitutional values that are implicated “when the focus shifts from human to machine action.”<sup>266</sup> Moreover, as David Engstrom and Daniel Ho explain, there is a dearth of ready-made legal or technological tools to cope with the breadth of governance challenges in the digital era: “[J]udicial review of agency action using AI is unlikely to yield systematic scrutiny,” “a thicket of reviewability and related doctrines largely insulate algorithmic decision making,” and “the current [administrative law] mechanisms remain ill-suited for providing meaningful accountability over rapid advances in AI.”<sup>267</sup>

Legal scholars have proposed a variety of doctrinal and institutional reforms to address the foregoing mismatches and maladaptation of AI for government use. For good reason, much of that prescriptive work is focused on constitutional law, administrative law, and norms of good

---

DEF. BLOG (July 8, 2020), [http://www.ai.mil/blog\\_07\\_08\\_20-jaic\\_completes\\_responsible\\_ai\\_champions\\_pilot.html](http://www.ai.mil/blog_07_08_20-jaic_completes_responsible_ai_champions_pilot.html) [<https://perma.cc/EQ73-GCRR>]; Jackson Barnett, *With \$106M Contract, JAIC Takes Major Step Building Central AI Platform for DOD*, FEDSCOOP (Aug. 13, 2020), <https://www.fedscoop.com/jaic-ai-development-platform-dod-joint-common-foundation-deloitte/> [<https://perma.cc/8VXT-J67B>].

263. See Press Release, JAIC Pub. Affs., Joint Artificial Intelligence Center to Pilot a Responsible AI Procurement Process (July 27, 2021), [https://www.ai.mil/news\\_07\\_27\\_21-jaic\\_to\\_pilot\\_a\\_responsible\\_ai\\_procurement\\_process.html](https://www.ai.mil/news_07_27_21-jaic_to_pilot_a_responsible_ai_procurement_process.html) [<https://perma.cc/333Y-AV45>]; see also *infra* notes 369–72 and accompanying text.

264. See *From Pilots to Proficiency: Operationalizing Federal AI*, MERITALK 17 (2021), <https://www.meritalk.com/wp-content/uploads/2021/05/pilots-to-proficiency.pdf> [<https://perma.cc/P5A4-LTZN>].

265. See *supra* notes 31–37 and accompanying text.

266. Huq, *supra* note 33, at 1881.

267. Engstrom & Ho, *supra* note 37, at 828, 844–45. In a similar vein, Wendy Wagner and Martin Murillo argue that the incentives and doctrines around agency rulemaking not only cut against the grain of current AI best practices but may also “tacitly reward[] agencies for developing and using algorithmic tools that are opaque and potentially biased.” Wagner & Murillo, *supra* note 37, at 3. Moreover, Deirdre Mulligan and Kenneth Bamberger argue that inexplicable AI systems may run afoul of the Administrative Procedure Act’s prohibition against “arbitrary and capricious” agency action. See Mulligan & Bamberger, *supra* note 37, at 773–74 (explaining that AI may violate the arbitrary and capricious agency standard); 5 U.S.C. § 706(2)(A) (providing that courts shall “hold unlawful and set aside [an] agency action” they deem to be “arbitrary [or] capricious”); Motor Vehicle Mfrs. Ass’n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 42–43 (1983).

governance—most notably pertaining to substantive and procedural regularity, transparency, and accountability.<sup>268</sup>

This Article aligns with that larger project: namely, to reconcile the ideals of our constitutional democracy with the sociotechnical challenges inhering in AI systems. But the injunctions and sanctions of constitutional and administrative law can only obliquely address a cache of governance challenges that originate and disseminate through the acquisition gateway. Without procurement law, the reformist agenda is dangerously incomplete.

This Article is not the first to sound the alarm. Deidre Mulligan and Kenneth Bamberger, for example, offer a trenchant account of how a “procurement mindset” can forfeit the government’s responsibility to make important design choices with public input.<sup>269</sup> Yet their proposed solutions are adjacent to procurement law itself. Specifically, they prescribe: (1) the use of “algorithmic impact assessments,” which would allow for public deliberation and input for AI systems that embed certain types of policy decisions; and (2) specialized in-house technical teams to provide consulting and support services to agencies adopting AI tools.<sup>270</sup>

Closer to home, some scholars and advocates have proposed more contract-based solutions. Robert Brauneis and Ellen Goodman, for example, urge procurement officials to use their “contracting powers to insist on appropriate record creation, provision, and disclosure.”<sup>271</sup> Along similar lines, Cary Coglianese and Erik Lampmann argue that “careful drafting of contracts for AI services paired with suitably robust public input over [contract] provisions . . . can allow procurement officers to assure the public that agencies are using AI tools responsibly.”<sup>272</sup> These prescriptions, too, angle in the right direction. Still, they only scratch the surface.

As yet, federal procurement law offers a reservoir of untapped possibilities. Indeed, as elucidated below, the acquisition gateway is primely situated to check and enable ethical algorithmic governance. As importantly, procurement law is uniquely suited for these purposes in ways that other legal frameworks miss.

#### IV. OPERATIONALIZING ETHICAL AI THROUGH PROCUREMENT LAW

This final Part drills into procurement’s positive potential. The recommendations here are keyed to four phases of the procurement pipeline: (1) acquisition planning; (2) market solicitation; (3) bid

---

268. See *supra* notes 36–37, 266–67, and accompanying text (observing the mismatch between machine learning AI systems and legal doctrine).

269. Mulligan & Bamberger, *supra* note 37, at 782.

270. *Id.* at 774.

271. See, e.g., Brauneis & Goodman, *supra* note 18, at 164.

272. Coglianese & Lampmann, *supra* note 39, at 180.

evaluation and source selection; and (4) contract performance. While each recommendation can be adopted in isolation, their full value will accrue in combination. Thematically, the approach here aims to capitalize on the merger of public and private interests around ethical AI. Toward that end, the recommendations exploit the procurement system's monetary and regulatory levers to incent market competition and responsible innovation. By centering ethical AI across the procurement lifecycle, the hope is that federal buyers and their AI suppliers will think more critically and holistically about the AI tools passing through the acquisition gateway for government use.

By way of background, federal procurement is subject to an elaborate body of regulations and practices designed to advance myriad objectives.<sup>273</sup> Most pertinent here, those objectives include market competition, transparency, efficiency, socioeconomic policy, risk avoidance, and best value to the government customer.<sup>274</sup> The Federal Acquisition Regulation (FAR) is "the primary regulation for use by all [federal] executive agencies in their acquisition of supplies and services with appropriated funds."<sup>275</sup> Federal procurement is also governed by agency-specific supplements to the FAR,<sup>276</sup> congressional statutes,<sup>277</sup> presidential Executive Orders,<sup>278</sup> and agency guidance documents.<sup>279</sup>

---

273. Schooner, *supra* note 255, at 634–37 ("The laws, regulations, and policies controlling the award and performance of government contracts present a dense thicket reflective of a large, complex bureaucracy."); *see also infra* note 274 and accompanying text.

274. *See generally* KATE M. MANUEL ET AL., CONG. RSCH. SERV., RS2826, THE FEDERAL ACQUISITION REGULATION (FAR): ANSWERS TO FREQUENTLY ASKED QUESTIONS (2015) (providing an overview of various procurement regulations and the values they serve); Steven L. Schooner, *Desiderata: Objectives for a System of Government Contract Law*, 11 PUB. PROCUREMENT L. REV. 103 (2002) (discussing several goals commonly associated with procurement systems).

275. *Foreword* to FAR (2021), <https://www.acquisition.gov/sites/default/files/current/far/pdf/FAR.pdf> [<https://perma.cc/WQ7W-M93R>].

276. *See* FAR 1.301(a)(1) (2021) (authorizing agencies to issue "agency acquisition regulations that implement or supplement the FAR, and incorporate . . . agency policies, procedures, and contract clauses, solicitation provisions, and forms" that govern the contract); *id.* 1.301(a)(2) (allowing for "internal agency guidance").

277. *See, e.g.*, 41 U.S.C. §§ 3101–06 (governing the procurement of supplies and services for most civilian agencies); 10 U.S.C. §§ 2302–39c (governing the procurement procedures for the DoD, National Aeronautics and Space Administration, and Coast Guard).

278. *See, e.g.*, Exec. Order No. 11,246, 30 Fed. Reg. 12,319, 12,319 (Sept. 28, 1965) (requiring government contractors not to discriminate and to develop affirmative action plans); Exec. Order No. 14,026, 86 Fed. Reg. 22,835, 22,835 (Apr. 27, 2021) (calling for increase in hourly minimum wage paid by the parties that contract with the federal government).

279. *See, e.g.*, FAR 1.301(a)(2) (2021) (allowing for "internal agency guidance"). *See generally* OFF. OF MGMT. & BUDGET, EXEC. OFF. OF THE PRESIDENT, OMB CIRCULAR A-76, PERFORMANCE OF COMMERCIAL ACTIVITIES (1999) (setting forth the procedures for determining

Two caveats before proceeding. First, this Article's recommendations for acquiring ethical AI are mostly agnostic to policymaking form—statutory, regulatory, or otherwise. To be sure, the choice of policymaking form can be consequential. For example, statutory mandates are generally more stable than Executives Orders, and amendments to the FAR would apply more broadly than agency-specific policies and practice. As a first pass, however, the discussion below focuses on substance and leaves questions of policymaking form to future work.<sup>280</sup> This approach allows for variation and experimentation, which can be a virtue, given AI's sociotechnical and contextual sensitivities.

Second, the recommendations below are most directly applicable to FAR-based procurement contracts—which account for the great bulk of federal acquisitions. Other Transaction Agreements (OTAs) are beyond this Article's immediate purview.<sup>281</sup> By regulatory design, a main purpose of OTAs is to provide alternative acquisition pathways for (mostly) nontraditional vendors to conduct research and prototype development for the government, unencumbered by the FAR's regulatory strictures, procedures, and contractual clauses.<sup>282</sup> While many of the recommendations advanced below can be adopted or adapted for OTAs, this Article leaves those questions for future work.

#### A. Acquisition Planning: AI Risk Assessments

Federal procurement begins with acquisition planning.<sup>283</sup> During this phase, the agency's product or service requirements are established, the personnel responsible for the acquisition are coordinated, costs and risks relating to the acquisition are assessed, and an overall acquisition strategy is developed.<sup>284</sup> When acquiring information technology (IT), agency officials must conduct specialized risk assessments pertaining to schedule

---

whether commercial activities should be outsourced or performed in-house using government facilities and personnel); OFF. OF MGMT. & BUDGET, EXEC. OFF. OF THE PRESIDENT, OMB CIRCULAR A-130, MANAGING INFORMATION AS A STRATEGIC RESOURCE (establishing general policy for the planning, budgeting, governance, acquisition, and management of federal IT systems).

280. Thus, insofar as the recommendations below would require agency officials to take certain actions, those mandates could come from Congress in the form of a statute, the White House in the form of an Executive Order, amendment to the FAR, or other forms of binding federal policy. Other recommendations are not intended to be binding, but call for new or modified procurement practice. Those recommendations can be implemented and supported by informal agency guidance.

281. For a useful overview of OTAs, and their use by the DoD in particular, see MOSHE SCHWARTZ & JEODO M. PETERS, CONG. RSCH. SERV., R45521, DEPARTMENT OF DEFENSE USE OF OTHER TRANSACTION AUTHORITY: BACKGROUND, ANALYSIS, AND ISSUES FOR CONGRESS (2019).

282. *Id.* at 2 (“[OTAs] are legally binding contracts that are generally exempt from federal procurement laws and regulations such as the Competition in Contracting Act and the [FAR]”).

283. See generally FAR Part 7 (2021) (governing acquisition planning).

284. See *id.* at 7.105 (2021).

and cost overruns, security and privacy, and interoperability with existing government systems.<sup>285</sup> IT risk assessments, however, are not styled or suited for the unique challenges of acquiring AI.<sup>286</sup> The recommendation here aims to fill that void with mandatory “AI risk assessments.”<sup>287</sup>

Because AI risks are contextually contingent, a one-size-fits-all approach is neither necessary nor advisable. But acquiring AI will always entail certain types of risks that can and should be accounted for during the planning phase. If nothing else, forecasting and logging AI risks will force conversations about whether an AI solution is necessary or appropriate to meet the agency’s needs, and if so, under what conditions and constraints.

By way of illustration, below is a non-exhaustive set of considerations that an AI risk assessment could capture:

- ❖ To what extent, if any, will the agency need to rely on third parties to design, develop, deploy, audit, or monitor the AI system? All else equal, the more the government must rely on third parties for these lifecycle needs, the less control the government will have over a system’s operations—both when the technology is working as intended and not.
- ❖ What are the transparency gaps in the AI system? As earlier explained, AI systems can be more or less transparent for a variety of technical and non-technical reasons.<sup>288</sup> The government should

---

285. See *id.* at 39.102(b) (2021); see also 44 U.S.C. § 3554(b); *id.* § 11331 (delegating authority to the Office of Management and Budget (OMB) and National Institute of Science & Technology (NIST) the authority to “promulgate information security standards pertaining to Federal information systems”); *Appendix IV to OMB Circular No. A-130*, OFF. OF MGMT. & BUDGET,

[https://obamawhitehouse.archives.gov/omb/circulars\\_a130\\_a130appendix\\_iv](https://obamawhitehouse.archives.gov/omb/circulars_a130_a130appendix_iv) [<https://perma.cc/HF67-332Z>] (“Each agency program official must understand the risk to [information] systems under their control.”).

286. See Raji et al., *supra* note 96, at 5 (“[T]he design, prototyping and maintenance of AI systems raises many unique challenges not commonly faced with other kinds of intelligent systems or computing systems more broadly.”); Exec. Order No. 13,960, 85 Fed. Reg. 78,939, 78,941 (Dec. 8, 2020) (observing that “[e]xisting OMB policies currently address many aspects of information and information technology design, development, acquisition, and use that apply, but are not unique, to AI.”). See generally *infra* Sections II.B, Section III.A (mapping the sociotechnical challenges of AI systems).

287. It bears noting that risk-management is considered best practice when private enterprises acquire AI technologies. For frameworks utilized in the private sector, see for example: *AI and Risk Management Innovating with Confidence*, DELOITTE CTR. FOR REGUL. STRATEGY (2018), <https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/deloitte-gx-ai-and-risk-management.pdf> [<https://perma.cc/3Y47-4N38>]; *AI Risk and Controls Matrix*, KPMG LLP (2018), <https://assets.kpmg/content/dam/kpmg/uk/pdf/2018/09/ai-risk-and-controls-matrix.pdf> [<https://perma.cc/XNB7-F9VR>].

288. See *supra* Section II.B.3.

tease out those differences, and assess each risk separately. For example, transparency risks may relate to model interpretability, data provenance, trade secrecy, model versioning, or some combination thereof.<sup>289</sup> Moreover, in many government settings, the interpretability and explainability of AI systems may be operationally or legally required. The costs of mitigating or overcoming these transparency risks will necessarily be context specific. But, at least in some contexts, transparency gaps might render an AI system unusable for its intended purpose.

- ❖ Will there be a human in-the-loop (or on-the-loop)? If so, what roles and responsibilities will be assigned to the human? In high-stakes and sensitive contexts, human validation of system inputs and outputs will generally be necessary before further government action is taken.<sup>290</sup> Moreover, in contexts where human judgment or discretion is required, risks relating to automation bias, automation aversion, model interpretability, etc., must also be forecasted and assessed.<sup>291</sup>
- ❖ Will the AI system be used in contexts that may have a discriminatory effect, or that may inflict special burdens or hardships on marginalized groups? Most public-facing AI use cases will carry these risks. But internal and back-office AI uses can also be risky. For example, AI systems used for government

---

289. See *supra* notes 167–71 and accompanying text; see also JONATHON PHILLIPS ET AL., NAT'L INST. STANDARDS & TECH., FOUR PRINCIPLES OF EXPLAINABLE ARTIFICIAL INTELLIGENCE 2–6 (2021), <https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf> [<https://perma.cc/YPE7-3X43>]; P. Jonathon Phillips et al., Nat'l Inst. Standards & Tech., Four Principles of Explainable Artificial Intelligence 10–11 (Aug. 2020), <https://www.nist.gov/system/files/documents/2020/08/17/NIST%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf> [<https://perma.cc/X26Y-FMGW>] (discussing different dimensions and considerations relating to AI explainability).

290. See Singh et al., *supra* note 99, at 13–14 (“[H]aving a human in the loop represents a clear point for exercising judgement, intervention and control.”). In time-critical contexts, a human in-the-loop might obstruct optimal system performance. For example, in the realms of cybersecurity and military tactical engagement, human oversight of the system (i.e., a human *on-the-loop*) may lead to better outcomes than human validation of AI outputs in real-time. Cf. Joel E. Fischer et al., *In-the-Loop or On-the-Loop? Interactional Arrangements to Support Team Coordination with a Planning Agent*, CONCURRENCY & COMPUTATION PRAC. & EXPERIENCE, Apr. 25, 2021, at 1, <https://onlinelibrary.wiley.com/doi/10.1002/cpe.4082> [<https://perma.cc/4DEW-EMUH>] (distinguishing between humans in-the-loop and on-the-loop, and studying contexts in which one structure might be preferable to others); NSCAI FINAL REPORT, *supra* note 11, at 9 (“Human operators will not be able to keep up with or defend against AI-enabled cyber or disinformation attacks, drone swarms, or missile attacks without the assistance of AI-enabled machines.”).

291. See *supra* notes 109–11 and accompanying text (discussing human-computer interactions as a feature of system design).

hiring and promotion, resource allocation, language translation, text generation, and building security,<sup>292</sup> may not work equally or sufficiently for certain subpopulations.

- ❖ Will the data used or generated by the AI system contain sensitive personal information? If yes, then a slew of considerations relating to data privacy, data integrity, and data security must be assessed and accounted for in the risk portfolio.<sup>293</sup>
- ❖ How might the AI system fail, or drift from its intended uses or performance standards? Relatedly, what protocols exist (or need to exist) to identify and rectify failure modes? As earlier explained, AI systems can fail or drift for myriad reasons—malign and benign, technical and non-technical.<sup>294</sup> Due to network effects, AI failure modes may also infect surrounding systems. To mitigate harm, agencies must be prepared for these contingencies in advance.
- ❖ Will the AI system require frequent updating, and if so, what protocols exist (or need to exist) to ensure the traceability and reliability of model versioning over time? A modified AI system may improve performance along one or more metrics but impair performance in other ways. Moreover, without proper precautions, model versioning can make it impossible to know how a model performed at the point in time that a particular government decision was made.<sup>295</sup> Without that information, an agency may be hard pressed to justify any actions based on the algorithmic output.

---

292. See U.S. GOV’T ACCOUNTABILITY OFF., GAO-21-526, FACIAL RECOGNITION TECHNOLOGY: CURRENT AND PLANNED USES BY FEDERAL AGENCIES 12–13 (2021) (discussing the use of facial recognition by several federal agencies for the purpose of digital access, cybersecurity purposes, and building security).

293. Existing risk-management frameworks for securing data and sensitive personal information could be used and tailored as necessary to capture AI-specific risks. Cf. *NIST Risk Management Framework*, NAT’L INST. STANDARDS & TECH., <https://csrc.nist.gov/Projects/Risk-Management> [https://perma.cc/7FKF-T7Q7] (linking to a suite of NIST standards and guidelines to support implementation of risk management programs to meet the requirements of the Federal Information Security Modernization Act); NIST JOINT TASK FORCE, NAT’L INST. STANDARDS & TECH., SPEC. PUBL’N 800-53 REV. 5, SECURITY AND PRIVACY CONTROLS FOR INFORMATION SYSTEMS AND ORGANIZATIONS (2020), <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf> [https://perma.cc/XA8D-7HU5].

294. See *supra* Section II.B.1.

295. See Cuéllar, *supra* note 36, at 135–36 (cautioning that “heavy reliance on computer programs—particularly adaptive ones that modify themselves over time—may complicate public deliberation about administrative decisions, because few observers may be entirely capable of understanding how a given decision was reached”).

- ❖ In addition to the foregoing, risks relating to system access, change management, compute resources, system interoperability, sustainability, and lifecycle costs should also be assessed (to the extent not already accounted for in other acquisition planning documents).<sup>296</sup>

The value of AI risk assessments will depend, in large measure, on the people responsible for their curation. In general, cross-disciplinary teams will be necessary for *all* procurement phases. At minimum, the AI risk assessment team should include subject matter experts, IT personnel, data scientists, lawyers, and ethical AI champions (who could be specially trained or certified for that function).

Skeptics may question whether this investment in human capital is necessary, but the answer is unequivocally yes. A diversity of experience and expertise mitigates contextual blind spots and cultivates systematic thinking about sociotechnical risks.<sup>297</sup> Skeptics may also worry that AI risk assessments will create bureaucratic drag on AI acquisitions.<sup>298</sup> To some extent, however, that is the point: to carve time and space for critical deliberations that otherwise may not occur or come too late.

Of course, time is a valuable resource that should not be squandered. But the benefits of AI risk assessments are likely to outweigh the costs of conducting them. More importantly, the benefits of curating AI risk assessments are likely to outweigh the costs of forgoing them.<sup>299</sup> Especially under current market conditions,<sup>300</sup> it would be irresponsible for the government to acquire AI solutions without rigorously screening for risks relating to safety, discrimination, privacy, transparency, and accountability. Future AI regulation may mitigate these concerns;

---

296. Even if these risks are captured elsewhere, collecting them in the AI risk assessment may be useful so that they can be managed and mitigated systematically.

297. See Schwartz et al., *supra* note 171, at 8 (“A consistent theme from the literature is the benefit of engaging a variety of stakeholders and maintaining diversity along social lines where bias is a concern (racial diversity, gender diversity, age diversity, diversity of physical ability.”); *id.* (“Technology or datasets that seem non-problematic to one group may be deemed disastrous by others.”).

298. This is a long-running concern in government contracting, especially for rapidly evolving technologies. Cf. Katherine M. John, *Information Technology Procurement in the United States and Canada: Reflecting on the Past with an Eye Toward the Future*, PROCUREMENT L., Summer 2014, at 4, 5 (2013) (“If procurement regimes overemphasize transparency and competition—or otherwise take too long—then end users might end up saddled with technology that is outdated by the time it reaches them.”).

299. Empirically, this may prove not to be the case. For the reasons indicated in the text above, however, it seems fair to assume that the costs of not doing risk assessments will be greater than the costs required to conduct them.

300. See *supra* note 23 and accompanying text.

nevertheless, AI risks will still endure to some significant extent, both in general and in government contexts more specifically.<sup>301</sup>

Last but not least: humans interacting with an AI system may reject it or engage in (risky) compensating behaviors if they do not trust the technology. Done right, AI risk assessments can set the foundations for that trust, inside and outside of government.<sup>302</sup>

### B. Market Solicitations: Calling for Ethical AI

The dividends of AI risk assessments extend beyond the planning phase. Most pertinent here, the government can recast the identified risks as focal points in the government's market solicitations. These solicitations may come in the form of requests for proposals (RFPs),<sup>303</sup> quotations (RFQs),<sup>304</sup> or information (RFIs).<sup>305</sup> Despite their legal and

---

301. See *supra* Section II.B (discussing a cache of latent risks and challenges in machine learning AI systems); KEARNS & ROTH, *supra* note 74, at 64 (“[A]nywhere machine learning is applied, the potential for discrimination and bias is very real—not in spite of the underlying scientific methodology but often because of it.”).

302. It is worth noting that this Article’s prescriptions for AI risk assessments may cohere with, but are different than, “algorithmic impact assessments” (AIAs). See, e.g., DILLON REISMAN ET AL., ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY 5–6 (2018), <https://ainowinstitute.org/aiareport2018.pdf> [<https://perma.cc/U26X-EDMQ>] (arguing for the use of AIAs to promote government accountability and public deliberation); Mulligan & Bamberger, *supra* note 37, at 842–45 (same); see also Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 110, 168 (2017) (same in the context of predictive policing in particular). In general, proposals for AIAs aim “to engage the public and proactively identify concerns, establish expectations, and draw on expertise and understanding from relevant stakeholders.” See REISMAN ET AL., *supra*, at 7. Some proposals for AIAs include opportunities for third-party auditing, as well as mechanism to challenge the agency’s impact assessment, including through judicial review. See *id.* at 10. This Article takes no position on AIAs or their ideal design features. For present purposes, the more important point is that the two instruments can harmonize toward the same general objectives: namely, safe, fair, transparent, and accountable algorithmic governance. Because of this alignment, information curated in AI risk assessments can be incorporated into an AIA covering the same system. And working in reverse, the public-facing requirements of AIAs can incent agencies to undertake robust AI risk assessments. (To the extent that AIAs are intended for agency self-assessments only, and not for public participation and external review, then pre-acquisition AI risk assessments and AIA might be functional equivalents).

303. See FAR 15.203(a) (2021) (“[RFPs] are used in negotiated acquisitions to communicate Government requirements to prospective contractors and to solicit proposals.”).

304. *Id.* at 8.402(d)(1) (explaining how RFQs are used when agencies order goods and services from federal supply schedules).

305. See FAR 15.201(e) (2021) (“RFIs may be used when the Government does not presently intend to award a contract, but wants to obtain price, delivery, other market information, or capabilities for planning purposes. Responses to these notices are not offers and cannot be accepted by the Government to form a binding contract.”). Market participants are not required to respond to RFIs. But, for strategic reasons, they often do. For example, a vendor can hope to draw attention to its products and capabilities, which may influence the requirements on a future government contract.

contextual distinctions, each of these instruments serve important dialogic functions. First, as discussed further below, the government can strategically utilize market solicitations to smooth out information asymmetries. Second, and as importantly, the government can craft these solicitations to drive innovation and commercial competition around ethical AI.

By way of illustration, and with the foregoing objectives in view, the government's solicitations can prompt vendors along the following lines:<sup>306</sup>

- ❖ Describe any training programs that your team members have undergone, and any official policies or protocols adopted by your company that specifically relate to AI safety, fairness, transparency, accountability, or other ethical AI principles.
- ❖ Describe how your developmental protocols or practices enable end-to-end auditability of the proposed AI solution, and any technical or proprietary limitations that may inhibit auditability. In this regard, would you permit auditing by independent third parties? If yes, explain the conditions or limitations you would impose. If such audits would not be allowed, then explain why.
- ❖ Describe any known or foreseeable performance weaknesses and vulnerabilities in your proposed AI solution, and explain the source or causes of those vulnerabilities (e.g., in the data, algorithm, design process, human-computer interface, interoperability with other hardware and software, or otherwise).
- ❖ Describe whether and how your proposed AI solution will be explainable and interpretable to end users, operators, auditors, and other stakeholders, including lay persons, judges, and policymakers.
- ❖ Describe any anticipated data-related limitations and challenges for your proposed AI solution. What strategies or protocols, if any, might you implement or recommend to address those challenges?

---

306. Additional questions and prompts, tailored to specific use cases, can and should be included in the government's solicitations. The World Economic Forum provides useful templates and suggestions that can be tailored for federal acquisitions. *See generally* SABINE GERDON ET AL., WORLD ECON. F., AI PROCUREMENT IN A BOX: WORKBOOK (2020), [http://www3.weforum.org/docs/WEF\\_AI\\_Procurement\\_in\\_a\\_Box\\_Workbook\\_2020.pdf](http://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_Workbook_2020.pdf) [<https://perma.cc/LE73-NGR8>].

- ❖ Describe your privacy and cybersecurity approach to the proposed AI solution, including but not limited to how the data and model will be protected from adversarial attack and human error.

Agencies have little (or nothing) to lose and much to gain from this information exchange. To start, vendor responses may shed light on previously unidentified AI risks. In such cases, the government can and should modify the AI risk assessment to capture those additional concerns. As importantly, vendor responses will enable the agency to make side-by-side comparisons of the risks and capabilities associated with a particular vendor (or AI solution) relative to the field.

Beyond obtaining information, the government can use these instruments to provide information about the government's requirements, constraints, and AI use contexts. More generally, however, centering ethical AI in market solicitations will signal to prospective vendors that they will need to compete on the field of ethical AI to win federal contracts. That signaling is important for three related reasons. First, strategic and innovative vendors may embrace ethical AI as a competitive differentiator. Second, the government's ethical voicing may draw responsible and innovative firms *into* the government market.<sup>307</sup> Third, as discussed below, the inclusion of ethical AI criteria in contract solicitations carries legal significance.

### C. Evaluation and Source Selection: Requiring Ethical AI

Under existing regulations, agencies must evaluate vendor proposals solely on the criteria pre-specified in the relevant contract solicitation.<sup>308</sup> Thus, to capitalize on this opportunity, agency officials will *need* to include ethical AI requirements in contract solicitations.<sup>309</sup> Separately, or

---

307. See *supra* notes 202–06 and accompanying text (discussing the technology industry's activism and the reticence of some firms to partner with the government in the areas of national security and law enforcement).

308. See FAR 15.305(a) (2021) ("An agency shall evaluate competitive proposals and then assess their relative qualities solely on the factors and subfactors specified in the solicitation."); *id.* at 15.304(d) (2021) ("All factors and significant subfactors that will affect contract award and their relative importance shall be stated clearly in the solicitation."); *id.* at 13.106-1(a)(2)(i) ("When soliciting quotations or offers [for simplified acquisitions,] the contracting officer shall notify potential quoters or offerors of the basis on which award will be made (price alone or price and other factors, e.g., past performance and quality)."); *id.* at 13.106-2(a)(2) ("Quotations or offers shall be evaluated on the basis established in the solicitation."); *see also* Antarctic Support Assocs. v. United States, 46 Fed. Cl. 145, 155 (2000) (noting that contractual awards must be consistent with stated evaluation criteria).

309. Advocacy groups and organizations have made similar recommendations as a matter of best practice, but not as a matter of law. See, e.g., AM. COUNCIL FOR TECH.-INDUS. ADVISORY COUNCIL, AI PLAYBOOK FOR THE U.S. FEDERAL GOVERNMENT 15, 22, 29, 35 (2020); World Econ. Forum, *AI Procurement in a Box: AI Government Procurement Guidelines*, WORLD ECONOMIC

additionally, ethical AI considerations could be factored into pre-award “responsibility” determinations of prospective vendors.<sup>310</sup> The discussion below elaborates on these recommendations and situates them within existing procurement policy.

### 1. Evaluation Criteria

As prefaced above, agencies must “evaluate competitive proposals and then assess their relative qualities solely on the factors and subfactors specified in the solicitation.”<sup>311</sup> This regulatory constraint promotes competition by steadyng the target for prospective vendors. Moreover, this constraint promotes the integrity and transparency of the acquisition process by committing agency officials to the specified evaluative criteria. When crafting solicitations, agencies have discretion over which evaluation criteria to include and prioritize.<sup>312</sup> But certain evaluative criteria, such as price and vendor past performance, generally must be included as a matter of law in competitive procurements.<sup>313</sup> The recommendation here is to create a similar requirement for ethical AI when the government acquires AI solutions.

Specifically, under this proposal, agency officials would be legally required to evaluate vendor proposals on ethical AI grounds. Discretionary waivers of this general rule could be allowed in exceptional circumstances or in specific contexts where AI risks are negligible. In such cases, however, contracting officials should be required to justify the waiver in writing.<sup>314</sup>

Like price and past performance, ethical AI principles will almost always be relevant in AI acquisitions. And, like price and past performance, the relative weight afforded to ethical AI can be determined on a contract-by-contract or contextual basis.<sup>315</sup> To be clear, ethical AI need not be paramount. But including ethical AI among the evaluative criteria will be necessary if the government intends to award contracts even partly on that basis.<sup>316</sup>

---

FORUM (June 11, 2020), <https://www.weforum.org/reports/ai-procurement-in-a-box/ai-government-procurement-guidelines#report-nav> [https://perma.cc/7L9L-DGFM].

310. See FAR 9.103(a) (“Purchases shall be made from, and contracts shall be awarded to, responsible prospective contractors only.”); see also *infra* notes 326–334 and accompanying text (discussing the regulatory framework for responsibility determinations).

311. FAR 15.305(a) (2021).

312. *Id.* at 15.304(c) (2021).

313. See *id.* at 15.304(c)(1), (2) (2021); see also *id.* at 13.106-1(a)(2).

314. Requiring a written justification for norm deviations has a pedigree in procurement law. See, e.g., FAR 6.303 (2021) (requiring written justifications under certain circumstances); *id.* at 13.501 (2021) (same). Without such a requirement, there is a real concern that contracting officials will not include ethical AI criteria in a systematic way and in contexts when they should.

315. See FAR 15.101–1 (2021).

316. See *supra* note 308 and accompanying text.

Currently, this is not the government's general practice—far from it. However, there are some encouraging signs of positive change. In 2021, for example, the DoD's Joint Artificial Intelligence Center (JAIC)<sup>317</sup> issued an RFP "to form multiple Blanket Purchase Agreements" with vendors who can provide AI testing and evaluation services to support the DoD and entire U.S. government.<sup>318</sup> As one of the first publicly available RFPs that even *mentions* ethical AI, it provides a useful baseline and template to build upon.

The Performance of Work Statement for this RFP plainly indicates that vendor solutions must account for the "DoD's AI Ethical Principles."<sup>319</sup> Moreover, the RFP explains that blanket purchase agreements will be awarded to a pool of the most highly qualified offerors based on their responses to the accompanying questionnaire.<sup>320</sup> In turn, that questionnaire asks vendors to describe their AI capabilities and developmental processes, along with examples of past performance to support their claims.<sup>321</sup> So far so good.

As pertains to ethical AI, however, the questionnaire contains only one question. To wit: "Is your company willing to incorporate responsible AI methodologies, such as the Department of Defense's AI Ethical Principles . . . into your company's testing and evaluation approach. (Yes or No)."<sup>322</sup> For this question, prospective vendors are instructed that "the Government will consider 'Yes' answers Acceptable and 'No' answers Unacceptable. The purpose of the [yes/no] question is to build awareness

---

317. The JAIC "serves as the DoD's acquisition hub and coordinator for the development and implementation of the Department's 'Responsible AI' strategy, guidance, and policy." Memorandum from Kathleen H. Hicks, *supra* note 7, at 2; *see also* Barnett, *supra* note 262 (discussing the JAIC's transition from development to acquisition and coordination).

318. JOINT AI CTR., U.S. DEP'T OF DEF., NOTICE I.D. W52P1J21R0029, JOINT ARTIFICIAL INTELLIGENCE CENTER TEST AND EVALUATION BLANKET PURCHASE AGREEMENT REQUEST FOR PROPOSAL (Feb. 11, 2021), [https://sam.gov/opp/93bc03aa061e43c0b5567ae8e33e9c2b/view?keywords=\[https://perma.cc/4LEU-AJAC\]](https://sam.gov/opp/93bc03aa061e43c0b5567ae8e33e9c2b/view?keywords=[https://perma.cc/4LEU-AJAC]).

319. See JOINT AI CTR., U.S. DEP'T OF DEF., NOTICE I.D. W52P1J21R0029, PERFORMANCE WORK STATEMENT §§ 1.2.2, 3.4.2, 4 (Feb. 11, 2021), [https://sam.gov/opp/93bc03aa061e43c0b5567ae8e33e9c2b/view?keywords=\[https://perma.cc/4LEU-AJAC\]](https://sam.gov/opp/93bc03aa061e43c0b5567ae8e33e9c2b/view?keywords=[https://perma.cc/4LEU-AJAC]) (downloadable as "Attachment 0002-JAIC and E BPA PWS 11Feb21.pdf") (describing compliance in testing and evaluation, quality control, and tasks).

320. JOINT AI CTR., U.S. DEP'T OF DEF., NOTICE I.D. W52P1J21R0029, INSTRUCTIONS AND EVALUATION CRITERIA 5–7 (Feb. 11, 2021) [hereinafter "JAIC INSTRUCTIONS"], [https://sam.gov/opp/93bc03aa061e43c0b5567ae8e33e9c2b/view?keywords=\[https://perma.cc/4LEU-AJAC\]](https://sam.gov/opp/93bc03aa061e43c0b5567ae8e33e9c2b/view?keywords=[https://perma.cc/4LEU-AJAC]) (downloadable as "Attachment 0003-JAIC TE BPA Instructions and Eval Criteria 11Feb21.pdf").

321. JOINT AI CTR., U.S. DEP'T OF DEF., NOTICE I.D. W52P1J21R0029, TEST & EVALUATION OF AI BLANKET PURCHASE AGREEMENT QUESTIONNAIRE 1–3 (Feb. 11, 2021) [hereinafter "JAIC QUESTIONNAIRE"], [https://sam.gov/opp/93bc03aa061e43c0b5567ae8e33e9c2b/view?keywords=\[https://perma.cc/4LEU-AJAC\]](https://sam.gov/opp/93bc03aa061e43c0b5567ae8e33e9c2b/view?keywords=[https://perma.cc/4LEU-AJAC]) (downloadable as "Attachment 0001-JAIC TE BPA Questionnaire 11Feb21.docx").

322. *Id.* at 3.

for DoD's AI Ethical Principles and to begin incorporating aspects of the principles in future call orders.”<sup>323</sup>

The JAIC, of course, is acutely aware that ethical AI cannot be captured in “Yes or No” terms. In fairness, perhaps the industrial base is not yet prepared to compete for government contracts on ethical AI grounds. Or perhaps the government is not ready to evaluate vendor proposals on those grounds. Either way, this early snapshot exposes the current gap between ethical AI in principle and practice, which this Article’s recommendations aim to bridge.<sup>324</sup>

As is, the government’s needs may be underserved. Worse still, the reduction of ethical AI to yes/no box ticking might be self-defeating if prospective vendors perceive, rightly or wrongly, that the government is not treating ethical AI as a *differentiator* in sourcing decisions. Per the RFP instructions, answering “No” to the ethical AI prompt will automatically disqualify prospective vendors from consideration.<sup>325</sup> Presumably, therefore, all responsive vendors checked “Yes.” But without more particulars, it is far from clear how the JAIC can or will differentiate among competing proposals on ethical AI grounds when awarding blanket purchase agreements under the RFP. Nor is it clear how customer agencies can or will do so at the call-order level. At minimum, however, the government will need to ensure that vendor proposals that reflect the costs of ethical AI will not be competitively *disadvantaged* (which potentially could occur, for example, if price is treated as an evaluative criteria, but ethical AI is not). Surely, this is not what the government intends.

Emphatically, the JAIC’s leadership in acquiring ethical AI is commendable, and the constructive critique here is not meant to suggest otherwise. Rather, the point is that even the government leaders in this space have a long haul ahead. Requiring contracting officials to include ethical AI among the evaluation criteria, and differentiating vendors on that basis, will be pivotal to progress. Acquiring ethical AI is something

---

323. See JAIC INSTRUCTIONS, *supra* note 320, at 6–7.

324. In May 2021—after the RFP for blanket purchase agreements was issued—the Deputy Secretary of Defense issued a memorandum titled “Implementing Responsible [AI] in the Department of Defense.” See generally Memorandum from Kathleen H. Hicks, *supra* note 7. Among other things, the memorandum calls for the incorporation of ethical AI principles into the DoD’s AI requirements and acquisition processes. To that end, the memorandum directs the JAIC to take specific steps, including the establishment of a Responsible AI Working Council that will provide recommendations to integrate ethical AI into the acquisition lifecycle. See *id.* at 3 (instructing the RAI Working Council to “provide recommendations on the integration of RAI into the AI acquisition requirements, on process, and on any policy modifications to enable RAI considerations within existing supply chain risk management practices.”). As of this writing, those recommendations are pending.

325. See JAIC INSTRUCTIONS, *supra* note 320, at 6–7.

that the government should insist upon. Otherwise, vendors cannot be expected to compete upon that basis.

## 2. Responsibility Determination

Additionally, or alternatively, ethical AI criteria could be integrated into contracting officials' pre-award responsibility determinations of prospective vendors. Like all the forgoing recommendations, this one builds upon pre-existing regulatory structure.

By way of background, longstanding procurement law requires vendors to satisfy a set of "responsibility" requirements,<sup>326</sup> which fall into three general categories. *First*, contracting officials must assess whether prospective vendors can fulfill the contract in a timely and satisfactory manner.<sup>327</sup> Toward those ends, a prospective vendor must demonstrate that it: has adequate financial resources; can meet the delivery schedule; has a satisfactory record of past performance; has a satisfactory record of business integrity and ethics; and has the necessary organization, technical skills, and production capabilities to perform the contract.<sup>328</sup> *Second*, prospective vendors must be "otherwise qualified and eligible to receive an award under applicable laws and regulations."<sup>329</sup> This general requirement, in turn, incorporates a range of socioeconomic policies effectuated through procurement law.<sup>330</sup> For example, a potential awardee must be deemed ineligible if it has not complied with federal equal employment opportunity requirements,<sup>331</sup> or fails to agree to an acceptable subcontracting plan with small businesses under the contract.<sup>332</sup> *Third*, the government may establish "special standards of responsibility" in contract solicitations when "necessary for a particular acquisition or class of acquisitions."<sup>333</sup> Per regulation, special standards "may be particularly desirable when experience has demonstrated that unusual expertise" is needed "for adequate contract performance."<sup>334</sup>

---

326. FAR 9.103(a) ("Purchases shall be made from, and contracts shall be awarded to, responsible prospective contractors only."). A vendor's failure to meet the responsibility threshold is disqualifying as a matter of law. *Id.* at 9.103(b) ("No purchase or award shall be made unless the contracting officer makes an affirmative determination of responsibility."); *id.* at 9.103(c) ("A prospective contractor must affirmatively demonstrate its responsibility . . .").

327. See KATE M. MANUEL, CONG. RSCH. SERV., R40633, RESPONSIBILITY DETERMINATIONS UNDER THE FEDERAL ACQUISITION REGULATION: LEGAL STANDARDS AND PROCEDURES 6–13 (2013) (providing explanations of the FAR's responsibility standards and processes); Ryan Co. v. United States, 43 Fed. Cl. 646, 651 (1999).

328. FAR 9.104-1 (2021).

329. *Id.* at 9.104-1(g) (2021).

330. See MANUEL, *supra* note 327, at 5.

331. See FAR 22.802 (2021); *id.* at 52.222–26 (2021); 41 C.F.R. § 60-1.1 (2021).

332. See 15 U.S.C. § 637(d)(4)(C); see also MANUEL, *supra* note 327, at 9 (listing these and other collateral responsibility requirements).

333. FAR 9.104-2 (2021).

334. *Id.*

The foregoing responsibility framework offers several points of ingress for ethical AI. Here are some possibilities, keyed to the typology above. *First*, as a general performance standard, ethical AI could be factored into a prospective vendor's "necessary organization, experience . . . [and] technical skills" to perform the contract, or the vendor's "record of integrity and business ethics."<sup>335</sup> *Second*, or alternatively, ethical AI standards could be established (e.g., by NIST), and then be required by law for vendors doing business with the federal government.<sup>336</sup> *Third*, ethical AI can be the basis for special standards of responsibility in connection with a particular contract or class of acquisitions.

For example, prospective vendors could be required to allow independent third-party auditing of their proposed AI solutions. Vendors could also be required to waive trade-secrecy claims under certain conditions or in certain contexts (e.g., in adjudicatory settings where the government must provide an explanation for an AI output). Furthermore, to support a diverse and robust AI ecosystem, prime contractors could be required to agree to subcontracting plans that include small businesses, or socioeconomically disadvantaged businesses, with ethical AI expertise.<sup>337</sup>

The foregoing suggestions come with important qualifiers and caveats. To start, intellectual property (IP) rights can be a major sticking point for AI vendors.<sup>338</sup> Indeed, for many small businesses and startups, trade secrets are their most valuable assets.<sup>339</sup> Thus, responsibility requirements related to IP should be limited to what is foreseeable

---

335. See *id.* at 9.104-1 (2021); see also *supra* notes 327–28 and accompanying text (discussing general performance standards).

336. See *supra* notes 329–44 and accompanying text.

337. Federal law has an established program to promote contracting with any "small business which is unconditionally owned and controlled by one or more socially and economically disadvantaged individuals who are of good character and citizens of and residing in the United States, and which demonstrates potential for success." 13 C.F.R. § 124.101 (2020); see also 15 U.S.C. § 637(a)(5) (defining "[s]ocially disadvantaged individuals" under the Small Business Act as "those who have been subjected to racial or ethnic prejudice or cultural bias because of their identity as a member of a group without regard to their individual qualities"); FAR 19.15 (2019) (discussing Woman-Owned Small Business Program); Exec. Order No. 13,985 § 7, 86 Fed. Reg. 7009, 7011 (Jan. 25, 2021) ("Government contracting and procurement opportunities should be available on an equal basis to all eligible providers of goods and services.").

338. See generally Rob Kitchin, *Thinking Critically About and Researching Algorithms*, 20 INFO. COMM'C'N. & SOC'Y 14, 20 (2016) ("[I]t is often a company's algorithms that provide it with a competitive advantage and they are reluctant to expose their intellectual property even with non-disclosure agreements in place."); Nancy O. Dix et al., *Fear and Loathing of Federal Contracting: Are Commercial Companies Really Afraid to Do Business with the Federal Government? Should They Be?*, 33 PUB. CONT. L.J. 5 (2003) (providing a review of the relevant contracting requirements and industry concerns around IP provisions in government contracts that depart from general commercial terms).

339. See Kitchin, *supra* note 338, at 20.

necessary.<sup>340</sup> If the government has no better options, it may need to pay vendors for their trade secrets, whether upfront, on a contingency basis, or otherwise. But, so long as a critical mass of innovative and responsible market participants are available to compete for the work, the disinclination of *some* vendors to meet the government's legal and operational needs is a poor reason to lower the bar for *all*.

Another potential friction is the government's use of procurement requirements to advance collateral socioeconomic policies. This is a longstanding and generalized concern, with varying degrees of intensity depending on context.<sup>341</sup> However, in this context, objections of this sort miss the mark. Procurement requirements pertaining to ethical AI are not, in the main, collateral socioeconomic policies detached from a vendor's ability to execute the contract. Rather, ethical AI requirements are foremost directed at meeting the government's operational and legal needs. The fact that ethical AI requirements may also promote socioeconomic objectives is a testament to procurement law's breadth of purpose and regulatory value. Lest there be any doubt, the procurement system's express and overarching objective is to "deliver . . . the *best value* product or service to the [government] customer, while maintaining the *public's trust* and fulfilling *public policy* objectives."<sup>342</sup> Ethical AI strikes all of those notes.<sup>343</sup> Conversely, spending many millions (or billions) of taxpayer dollars for ethically agnostic AI solutions would be antithetical to best value, could undermine public trust, and leave public policy objectives unfilled.

If a prospective vendor is unable or unwilling to satisfy responsibility thresholds relating to ethical AI, then the government should be required to select a competitor that will. And if no vendor will, then the

---

340. There are well-established industry practices (e.g., nondisclosure agreements with liability provisions) and federal trade secrecy laws that can be utilized to safeguard vendors against trade secrecy misappropriation. *See, e.g.*, Defend Trade Secrets Act of 2016, Pub. L. No. 114-153, 130 Stat. 376 (codified at 34 U.S.C. § 41310 (2017)).

341. The use of procurement law to influence socioeconomic policy has been the subject of controversy, at various times at to various degrees. *See, e.g.*, Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 211 (1995) (analyzing the validity of the Federal Government's practice of providing general contractors on federal projects with incentives to hire subcontractors operated by socially and economically disadvantaged individuals); OFF. OF SEN. ELIZABETH WARREN, BREACH OF CONTRACT: HOW FEDERAL CONTRACTORS FAIL AMERICAN WORKERS ON THE TAXPAYER'S DIME 2 (2017). For generations, the government has leveraged the procurement system to advance national policy objectives, and has generally been able to do so because of its special relationship with federal contractors and raw spending power. *See* Schooner, *supra* note 274, at 108–09 (explaining that "government spending can influence behav[ior] and infuse growth in communities and economic sectors").

342. FAR 1.102(a) (2021) (emphasis added).

343. *See* Exec. Order No. 13,960, 85 Fed. Reg. 78,939, 78,940 (Dec. 8, 2020) (instructing agencies to abide to ethical AI principles when acquiring AI); *see also supra* Section II.C.2 (discussing ethical AI initiatives within the federal government).

government should rethink whether a market solution is appropriate, endeavor to fill the need in-house, or seek other (non-AI) solutions.<sup>344</sup>

#### D. Contract Performance: Pathways and Pitfalls

Thus far, the discussion has focused on acquisition planning, market solicitation, proposal evaluation, and contract award. This final section turns to contract performance.<sup>345</sup> Before proceeding, it must be emphasized that the challenges and opportunities for acquiring ethical AI will depend on what transpires during the preceding phases. But, even if the recommendations above are duly implemented, contract performance will be key to mission success.

The ethical AI challenges during this phase depend on countless variables. One important organizing distinction is between (1) commercial off-the-shelf (COTS) and (2) customized AI solutions. Although necessarily partial, this dichotomy provides useful framing to address various challenges that may arise or manifest after a contract is awarded.

##### 1. COTS AI Solutions

COTS acquisitions aspire to transactions in the commercial market.<sup>346</sup> For that reason, COTS items are offered to the government “without modification.”<sup>347</sup> Moreover, COTS vendors are relieved of certain regulatory terms and conditions unique to government contracting.<sup>348</sup> This lowers the market barriers for commercial vendors that otherwise might not do business with the government. And by reducing red tape, COTS acquisitions are generally more efficient for the government as well.

The regulatory pretenses around COTS acquisitions are fundamentally pragmatic: if the product is good enough for commercial

---

344. The use of RFIs, discussed above, is one way the government can gauge whether sufficient competition exists in the market to meet the government’s legal, operational, and sustainment needs. *See generally supra* Section IV.B (discussing market solicitations and offering illustrations).

345. *See generally* FAR Part 42 (2021) (governing contract administration and audit services).

346. *See* Christopher F. Corr & Kristina Zisis, *Convergence and Opportunity: The WTO Government Procurement Agreement and U.S. Procurement Reform*, 18 N.Y.L. SCH. J. INT’L & COMP. L. 303, 314–15 (1999) (discussing the statutory genesis and motivations behind COTS acquisitions).

347. *See* FAR 2.201 (2021) (defining “[c]ommercially available off-the-shelf item”).

348. *See id.* at 12.503–04 (2021) (listing laws that are not applicable to federal contracts and subcontracts for commercial items); *id.* at 12.505 (2021) (listing additional laws that are not applicable to COTS items).

consumption, then it should be good enough for the government too.<sup>349</sup> Through an ethical AI lens, however, these pretenses are precarious. Unlike virtually all other COTS solutions, the risks of AI product failure and related harms have *not* been meditated by regulatory or market forces. Quite the contrary, the design and development of commercially available plug-and-play AI solutions are virtually unregulated.

In these routinized transactions, moreover, the government will “acquire only the technical data and the rights in that data customarily provided to the public with a commercial item or process.”<sup>350</sup> Because those data rights are generally quite limited, the government may forfeit crucial opportunities to address its ethical AI needs, both at the time of purchase and thereafter. Although COTS products can be configured to agency needs, doing so can lead to long term support and maintenance challenges, insofar as customized functionality is not supported by the COTS vendor.<sup>351</sup>

## 2. Customized AI Solutions

Customized AI systems may be better suited for the government’s missions and lifecycle needs but give rise to different challenges. At the outset, traditional “waterfall” acquisition pathways are not viable for custom AI solutions. Under a typical waterfall approach, the government’s technical and design requirements are fixed at the time of contracting.<sup>352</sup> Given the complexity of the AI development process, however, it may be impossible for the agency to specify conditions of performance at the time of contracting. Even if possible, front-loading decisions about system features and configurations is antithetical to the trial-and-error nature of AI development.

For these reasons, “agile” acquisition methodologies are better suited for custom AI solutions. Agile methodologies are characterized by incremental, modular, and iterative processes in which software is produced in close collaboration with the end user.<sup>353</sup> Information obtained during these frequent iterations allow developers to respond quickly to feedback from agency customers, thus potentially reducing sociotechnical, legal, and programmatic risk. Moreover, modular

---

349. Cf. Laura Gerhardt et al., *When to Use Commercial Off-the-Shelf (COTS) Technology*, 18F BLOG (Mar. 26, 2019), <https://18f.gsa.gov/2019/03/26/when-to-use-COTS/> [https://perma.cc/SN2N-DN2Q] (outlining some general considerations for agencies to consider when choosing between COTS and customized software solutions).

350. FAR 12.211 (2021) (technical data rights); *see also id.* at 12.212 (2021) (computer software documentation).

351. *See* Gerhardt et al., *supra* note 349.

352. *See* U.S. GOV’T ACCOUNTABILITY OFF., GAO-20-590G, AGILE ASSESSMENT GUIDE: BEST PRACTICES FOR AGILE ADOPTION AND IMPLEMENTATION 7 (2020), <https://www.gao.gov/products/GAO-20-590G> [https://perma.cc/KZ7S-F3V6].

353. *See id.*

contracting vehicles “provide an opportunity for subsequent increments to take advantage of any evolution in technology or needs that occur during implementation,” and can “reduce risk of potential adverse consequences on the overall project by isolating and avoiding custom-designed components of the system.”<sup>354</sup>

The built-in flexibilities of agile processes may be well suited for AI development—certainly more so than a waterfall approach. But better is not sufficient. Like so much else in AI’s domain, agile methodologies will need to be retrofitted and retooled for the unique challenges of acquiring ethically designed AI systems.

This is no small matter. Even for conventional (non-AI) software acquisitions, agile approaches require skillsets, resources, and institutional buy-in that many agencies currently lack. In 2020, the Government Accountability Office (GAO) chronicled an array of challenges that agencies have experienced using agile acquisition processes in the past.<sup>355</sup> For example, “teams reported difficulty collaborating closely or transitioning to self-directed work due to constraints in organization commitment and collaboration.”<sup>356</sup> Moreover, GAO reported that some agency organizations “did not have trust in iterative solutions and that teams had difficulty managing iterative requirements.”<sup>357</sup>

This GAO report does not directly address the unique challenges of AI systems, much less AI ethics. Nor do other recently issued government best-practices guides.<sup>358</sup> But it must be assumed that the government’s

---

354. 48 C.F.R. § 39.103 (2021).

355. See GAO-20-590G, *supra* note 352, at 14–16.

356. *Id.* at 14.

357. *Id.*; see also U.S. GOV’T ACCOUNTABILITY OFF., GAO-16-467, IMMIGRATION BENEFITS SYSTEM: US CITIZENSHIP AND IMMIGRATION SERVICES CAN IMPROVE PROGRAM MANAGEMENT 24 (2016), <https://www.gao.gov/products/GAO-16-467> [https://perma.cc/R279-REQW] (reporting that the United States Citizenship and Immigration Service Transformation program was not setting outcomes for Agile software development); U.S. GOV’T ACCOUNTABILITY OFF., GAO-18-46, TSA MODERNIZATION: USE OF SOUND PROGRAM MANAGEMENT AND OVERSIGHT PRACTICES IS NEEDED TO AVOID REPEATING PAST PROBLEMS 57 (2017), <https://www.gao.gov/products/GAO-18-46> [https://perma.cc/CS84-J42U] (reporting that the Transportation Security Administration’s Technology Infrastructure Modernization (TIM) program did not define key Agile roles, prioritize system requirements, or implement automated capabilities).

358. The U.S. Digital Service and the General Services Administration (GSA) have likewise championed agile methodologies for acquisitions of customized software. See, e.g., OFF. OF MGMT. & BUDGET, CONTRACTING GUIDANCE TO SUPPORT MODULAR DEVELOPMENT 7, 12 (2012), <https://obamawhitehouse.archives.gov/sites/default/files/omb/procurement/guidance/modular-approaches-for-information-technology.pdf> [https://perma.cc/4S9G-JWBC]; HANDBOOK FOR PROCURING DIGITAL SERVICES USING AGILE PROCESSES, at i, 1 (2014), [https://playbook.cio.gov/assets/TechFAR%20Handbook\\_2014-08-07.pdf](https://playbook.cio.gov/assets/TechFAR%20Handbook_2014-08-07.pdf) [https://perma.cc/R3YE-FBZJ]; GEN. SERVS. ADMIN., DE-RISKING GOVERNMENT TECHNOLOGY: FEDERAL AGENCY FIELD GUIDE 7, 10–12

agility challenges will be amplified in the AI context, given the data-centric dynamics, value-laden judgments, cross-disciplinarity, procedural discipline, and institutional buy-in required to acquire ethical AI solutions.<sup>359</sup> Indeed, under the status quo, agile methodologies may cut against the grain of ethical AI. As Joshua Kroll explains, agile workflows and development sprints can lead to path-dependent and shallow thinking about the social implications of technical designs.<sup>360</sup> Moreover, Deb Raji et al. note that agile AI development presents unique auditing challenges, especially if the developers are not scrupulous about managing the data and documenting key decisions throughout the iterative process.<sup>361</sup>

To be sure, synching ethical AI with agile methodologies is an ongoing challenge. While some promising approaches to this challenge exist,<sup>362</sup> none can lay claim to standard industry practice. Regardless, what may work in the private sector may not translate in the government sector. The government's general struggles with agile methodologies, and its AI talent shortages, are again of relevant concern. But, in addition, the government's agility is limited by regulatory constraints. For example, the government cannot lawfully outsource "inherently governmental functions."<sup>363</sup> Although this limitation is notoriously fuzzy and forgiving, it could—and arguably should—prevent the government from devolving policy choices to vendors in the AI development process.<sup>364</sup> Moreover, the government is generally prohibited from entering into "personal services contracts."<sup>365</sup> This is a fuzzy and forgiving standard as well, but it generally forbids the government from

---

(2020), <https://derisking-guide.18f.gov/assets/federal-field-guide-a245c3a7dcd0a24f619b458fd51e1e490f2299023fd1bd13fddc87318e67cf03.pdf> [<https://perma.cc/DE6S-QJUA>]. These best-practices guides are generalized for software. Thus, they do not speak to the special challenges of AI development.

359. Raji et al., *supra* note 96, at 5 ("The design, prototyping and maintenance of AI systems raises many unique challenges not commonly faced with other kinds of intelligent systems or computing systems more broadly.").

360. Joshua A. Kroll, *The Fallacy of Inscrutability*, 376 PHIL. TRANSACTIONS ROYAL SOC'Y 1, 5 (2018), <http://dx.doi.org/10.1098/rsta.2018.0084> [<https://perma.cc/7EC4-UT6Z>].

361. Raji et al., *supra* note 96, at 4.

362. See, e.g., Dorian Peters, Karina Vold, Diana Robinson & Rafael A. Calvo, *Responsible AI—Two Frameworks for Ethical Design Practice*, 1 IEEE TRANSACTIONS ON TECH. & SOC'Y 34 (2020).

363. See FAR 7.503 (2021); see also *id.* at 37.104(c)(2) ("Each contract arrangement must be judged in the light of its own facts and circumstances, the key question always being: Will the Government exercise relatively continuous supervision and control over the contractor personnel performing the contract.").

364. See generally KATE M. MANUEL, DEFINITIONS OF "INHERENTLY GOVERNMENTAL FUNCTION" IN FEDERAL PROCUREMENT LAW AND GUIDANCE (2014), <https://sgp.fas.org/crs/misc/R42325.pdf> [<https://perma.cc/QD7J-E7YW>] (analyzing the bounds of these constraints).

365. See *id.* at 37.104 (2021).

micromanaging vendors during contract performance.<sup>366</sup> These regulatory bounds are seldom a problem for government acquisitions. Yet they may prove to be in this context, whether because of the value judgments embedded in AI technologies, or the process by which those decisions are made.

To put it mildly, it is too soon to know—and dangerous to assume—that industry approaches for integrating AI ethics into agile workflows will be effective, scalable, and suitable for government contexts. To set the right conditions, agencies will first need to be much more attentive to the risks inherent in ethics-agnostic agile methodologies, and then foster market competition around that particular challenge.

With those related aims, the government's market solicitations should contain prompts and requirements that explicitly tether ethical AI to agile methodologies. Just for example, the government could ask potential vendors to:

- ❖ Identify agile methodologies that your company has used that incorporate ethical AI principles, and provide two (or three) examples that demonstrate capability in those methodologies.
- ❖ Describe any challenges or lessons learned from those past experiences, and any anticipated challenges, strategies or solutions that your company might implement in future work.

These prompts are no panacea; they are merely preludes to contract performance. But if the government expects ethical AI *and* agile methodologies, the government must recognize that the sum of the two is greater than its parts. Moreover, the government should make sourcing decisions that account for the difference, and secure funding to pay for that difference. Otherwise, the government will have plenty of industry partners that are fluent in agile AI development, and “willing to incorporate” ethical AI principles,<sup>367</sup> but that have no plans or protocols to synchronize these ambitions.

---

366. See *id.* at 37.104(c)(2) (“Each contract arrangement must be judged in the light of its own facts and circumstances, the key question always being: Will the Government exercise relatively continuous supervision and control over the contractor personnel performing the contract.”); see also ASI Gov’t, *A COR’s Guide to Personal Services Contracts* (2011), [https://www.navsup.navy.mil/site/public/fleph/documents/contracting/cor\\_guides/A\\_CORs\\_Guide\\_to\\_Personal\\_Services\\_Contracts.pdf](https://www.navsup.navy.mil/site/public/fleph/documents/contracting/cor_guides/A_CORs_Guide_to_Personal_Services_Contracts.pdf) [<https://perma.cc/62BX-DH3U>] (providing guidance in navigating these murky constraints).

367. See JAIC QUESTIONNAIRE, *supra* note 321, at 3 (requiring such willingness as a condition of a procurement proposal).

\*\*\*

The foregoing discussion has provided a blueprint for acquiring ethical AI which spans the procurement process. Still, and again, it must be emphasized that parchment policies are not enough: acquiring ethical AI requires intentionality, additional resources, industry collaboration, and government coordination. Moreover, robust implementation of this Article's recommendations will depend on a cadre of talented and dedicated personnel that can execute the mission. It is one thing if the government does not have the in-house capacity to meet its demand for ethical AI solutions. It is quite another matter, however, if the government does not have a federal acquisition workforce with the skills, resources, and cultural commitment to *acquire* ethical AI from the private market. In this regard, the NSCAI has admonished that agencies which "rely solely on contractors for digital expertise will become incapable of understanding the underlying technology well enough to make successful acquisition decisions independent of contractors."<sup>368</sup>

As this Article nears completion, there are some promising signs of meaningful government progress toward acquiring ethical AI. Of special note, the JAIC has recently launched a program to align AI acquisitions with ethical AI as "part of a holistic approach that focuses not only on the technology, but on organizational operating structures and culture to advance Responsible AI within the DoD."<sup>369</sup> The GSA's AI Center of Excellence, which provides governmentwide acquisition and development support, is also championing ethical AI in its offerings.<sup>370</sup> More generally, a recently introduced Senate bill, titled the "Artificial Intelligence Training for the Acquisition Workforce Act,"<sup>371</sup> would establish an AI training program to ensure that acquisition personnel,

---

368. NSCAI FINAL REPORT, *supra* note 11, at 123.

369. Press Release, JAIC Pub. Affs., Joint Artificial Intelligence Center to Pilot a Responsible AI Procurement Process (July 27, 2021), [https://www.ai.mil/news\\_07\\_27\\_21-jaic\\_to\\_pilot\\_a\\_responsible\\_ai\\_procurement\\_process.html](https://www.ai.mil/news_07_27_21-jaic_to_pilot_a_responsible_ai_procurement_process.html) [https://perma.cc/333Y-AV45]. Moreover, the program aspires to "establish clear guidance and expectations for those who are interested in working with DoD to ensure that they are providing AI systems designed, developed, deployed, and used responsibly." *Id.*

370. See A.I. CTR. OF EXCELLENCE, GEN. SERVS. ADMIN., COE GUIDE TO AI ETHICS 1, <https://coe.gsa.gov/docs/CoE%20Guide%20to%20AI%20Ethics.pdf> [https://perma.cc/3TJ6-ZFVX]; *Accelerate Adoption of Artificial Intelligence to Discover Insights at Machine Speed*, A.I. CTR. OF EXCELLENCE, <https://coe.gsa.gov/coe/artificial-intelligence.html> - coe-updates [https://perma.cc/5MLP-EVEL]; A.I. CTR. OF EXCELLENCE, GEN. SERVS. ADMIN., PROMOTE ADOPTION OF MODERN PRACTICES TO INCREASE SUCCESS OF IMMEDIATE & FUTURE MODERNIZATION EFFORTS 1 (2021), [https://coe.gsa.gov/docs/2020/CoE\\_Innovation\\_Adoption\\_Service\\_Catalog\\_2021.pdf](https://coe.gsa.gov/docs/2020/CoE_Innovation_Adoption_Service_Catalog_2021.pdf) [https://perma.cc/P76Z-8MDE]; see also Dave Nyczepir, *IT Centers of Excellence Program Is Signed into Law*, FEDSCOOP (Dec. 3, 2020), <https://www.fedscoop.com/gsa-centers-of-excellence-codified/> [https://perma.cc/X26N-828P].

371. S. 2551, 117th Cong. (2021).

program managers, system evaluators, and others agency officials, have “knowledge of the capabilities and risks associated with AI.”<sup>372</sup> Time will tell, but initiatives of this sort are precisely what is needed across government.

## CONCLUSION

It is encouraging that the United States has committed to ethical AI principles and the protection of “civil liberties, privacy, and American values . . . in order to fully realize the potential of AI technologies for the American people.”<sup>373</sup> But proselytizing is not actualizing. If federal officials are truly committed to ethical algorithmic governance, then the federal procurement system must be recalibrated for that purpose. This Article has provided a principled and pragmatic agenda for acquiring ethical AI that future work can utilize and build upon.

---

372. *Id.* § 2(b)(2)–(3).

373. Exec. Order No. 13,859, 84 Fed. Reg. 3967, 3967 (Feb. 14, 2019).

# Federal Procurement of Artificial Intelligence:

## Perils and Possibilities

REPORT BY **DAVID S. RUBENSTEIN**  
DECEMBER 2020

THE  
**GREAT**  
**DEMOCRACY**  
INITIATIVE

# THE GREAT DEMOCRACY INITIATIVE

## ABOUT THE GREAT DEMOCRACY INITIATIVE

The Great Democracy Initiative develops policy blueprints that offer solutions to the most pressing problems of our time. From taming the concentration of power in our economy to fundamentally reforming our broken government, GDI aims to generate policy ideas that confront the forces that have rigged our society in favor of the powerful and connected.

## ABOUT THE AUTHOR

**David S. Rubenstein** is James R. Ahrens Chair in Constitutional Law, and Director of the Robert J. Dole Center for Law & Government, at Washburn University School of Law. Prior to teaching, Professor Rubenstein clerked for The Honorable Sonia Sotomayor when she was a judge on the United States Court of Appeals for the Second Circuit, and for The Honorable Barbara Jones in the United States District Court for the Southern District of New York. Prior to clerking, he served for three years as Assistant United States Attorney in the Southern District of New York, and was a litigation associate for five years at King & Spalding LLP.

## ACKNOWLEDGEMENTS

The author is grateful to Daniel Ho, Raj Nayak, Suzanne Kahn, and Eric Jacobs for their incisive comments and feedback; to Matt Hughes for tight editing; to Anna Smith for logistical support; and to Kaitlyn Bull, Ande Davis, Penny Fell, Barbara Ginsberg, Creighton Miller, and Zach Smith, for invaluable research assistance; and to Leah for everything always. All errors and omissions are the author's alone.

# I. Introduction

Artificial intelligence (AI) is transforming how government operates.<sup>1</sup> For example, the Federal Bureau of Investigation uses AI in law enforcement; the Social Security Administration uses AI to adjudicate benefits claims; the Food and Drug Administration uses AI to support its rulemaking processes;<sup>2</sup> the Department of Homeland Security uses AI to regulate immigration;<sup>3</sup> and countless other agencies are experimenting with AI for the delivery of government services, customer support, research, and regulatory analysis.<sup>4</sup> This small sampling presages a new era of “algorithmic governance,” in which government tasks assigned to humans will increasingly migrate to machines.<sup>5</sup>

Algorithmic governance brims with promise and peril. Under the right conditions, AI systems can solve complex problems, reduce administrative burdens, and optimize resources. Under the wrong conditions, AI systems can lead to widespread discrimination, invasions of privacy, dangerous concentrations of power, and the erosion of democratic norms.

The possibilities and perils of AI’s social disruptions have led the United States and institutions across the globe to propagate principles of trustworthy and ethical AI.<sup>6</sup> Although the particulars vary, ethical AI envisages a cluster of principles relating to transparency, accountability, fairness, privacy, and security.<sup>7</sup> Translating these principles into practice is the next step. Currently, dozens of pending congressional

---

<sup>1</sup> AI has no singular definition. Here, I use the term AI to describe a range of computer-enabled abilities and methods to perform tasks that would otherwise require human intelligence, such as learning, adaptation, reasoning, prediction, optimization, and sensory understanding. See *infra* Subpart I.A (explaining the definitional problem and offering some refinements).

<sup>2</sup> See generally DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY, MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (2020) [hereafter GOVERNMENT BY ALGORITHM], <https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf>.

<sup>3</sup> See Aaron Boyd, *CBP Is Upgrading to a New Facial Recognition Algorithm in March*, NEXTGOV.COM (Feb. 7, 2020), <https://www.nextgov.com/emerging-tech/2020/02/cbp-upgrading-new-facial-recognition-algorithm-march/162959/>.

<sup>4</sup> See, e.g., GOVERNMENT BY ALGORITHM, *supra* note 2; Cheryl Ingstad, *A Message from Leadership*, AI AT DOE NEWSLETTER (Aug. 2020), <https://www.energy.gov/sites/prod/files/2020/10/f80/AI%20Newsletter%202020%2002%2011.pdf>; PR Newswire, Geospark Analytics Awarded FEMA Contract for Use of Hyperion and AI-Driven Risk Models, AP NEWS (May 28, 2020), <https://apnews.com/PR%20Newswire/29397df99616a5ea6413cc70caf7ca68>.

<sup>5</sup> See Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 636 (2017); see also Will Hurd and Robin Kelly, *Rise of the Machines: Artificial Intelligence and its Growing Impact on U.S. Policy* (2018), <https://www.hsl.org/?view&id=816362>.

<sup>6</sup> See Anna Jobin et al., *The Global Landscape of AI Ethics Guidelines*, 1 NATURE MACHINE LEARNING 389 (2019), <https://www.nature.com/articles/s42256-019-0088-2.pdf> (mapping and analyzing the corpus of principles and guidelines on ethical AI).

<sup>7</sup> *Id.*; see also Jessica Fjeld et al., Berkman Klein Ctr., Res. Publ’n No. 2020-1, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (2020), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3518482](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482).

bills would regulate private and governmental uses of AI, the data that fuels this technology, and the infrastructures needed to sustain it.<sup>8</sup>

Meanwhile, the executive branch is revving its engines. In December 2020, President Trump issued an Executive Order to “promote the innovation and use of AI” in government operations “in a manner that fosters public trust, builds confidence in AI, protects our Nation’s values, and remains consistent with all applicable laws, including those relating to privacy, civil rights, and civil liberties.”<sup>9</sup> This builds on a February 2019 Executive Order, which projects that AI will affect the missions of nearly all executive departments and agencies, and sketches a plan for “maintaining American leadership” in innovative and trustworthy AI.<sup>10</sup>

Wide swaths of law and public administration will need retrofitting to accommodate algorithmic governance.<sup>11</sup> This report focuses critical attention on one regulatory domain that requires immediate attention: *federal procurement law*.

For a variety of reasons, the government’s pent-up demand for AI systems far exceeds its in-house capacity to design, develop, field, and monitor this powerful technology.<sup>12</sup> Accordingly, many (if not most) of the tools and services of algorithmic governance will be procured by contract from the technology industry. This is highly concerning, in part, because AI systems are virtually unregulated in the private market.<sup>13</sup> Without intervention, the government will be acquiring unregulated technology for government functions. Moreover, when procured from the private market, AI systems

---

<sup>8</sup> See Center for Data Innovation, *AI Legislation Tracker—United States*, <https://www.datainnovation.org/ai-policy-leadership/ai-legislation-tracker/> (last visited July 17, 2020); Yoon Chae, *U.S. AI Regulation Guide: Legislative Overview and Practical Considerations*, ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW (Jan.–Feb. 2020) (reporting that in 2015–2016, only two introduced bills contained the term “artificial intelligence”; that increased to 51 bills by the end of 2019).

<sup>9</sup> Exec. Order No. 13,960, 85 Fed. Reg. 78,939 (Dec. 8, 2020). By sheer happenstance, the executive order was issued the same day this report was finalized. For that reason, time and space did not permit further discussion here. It is worth noting, however, that the executive order aligns in many ways with the normative and prescriptive thrust of this report.

<sup>10</sup> Exec. Order No. 13,859, 84 Fed. Reg. 3,967 (Feb. 14, 2019).

<sup>11</sup> See Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399 (2017) (outlining several policymaking issues that will need to be addressed in the near term).

<sup>12</sup> As recently reported by the National Security Commission on Artificial Intelligence (NSCAI): “[T]here is a severe shortage of AI knowledge in [Department of Defense] and other parts of government . . . Current initiatives are helpful, but only work around the edges, and are insufficient to meet the government’s needs.” Nat’l Sec. Comm’n on Artificial Intelligence, Second Quarter Recommendations 34 (2020) [hereafter NSCAI, Second Quarter Recommendations], <https://drive.google.com/file/d/1hgIA38FcyFcVQOJhsycz0Ami4Q6VLVEU/view>. Cf. GOVERNMENT BY ALGORITHM, *supra* note 2, at 18, 89 (finding that approximately half of AI applications currently in use were developed in-house by federal agency personnel, but acknowledging the government’s in-house capacity challenges); Rudy Mehrbani, Tess Byars, Louis Katz, *A Time to Serve: Proposals for Renewing the Civil Service*, GREAT DEMOCRACY INITIATIVE (Aug. 2020), <https://greatdemocracyinitiative.org/wp-content/uploads/2020/08/Personnel-Policy-Final-Copy.pdf> (arguing for the “need to change hiring practices to create more and better pathway into government for diverse and talented workers,” including by “modernizing the civil service system”).

<sup>13</sup> See Russell T. Vought, Office of Mgmt. & Budget, Guidance for Regulation of Artificial Intelligence Applications (2020), <https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf>.

may be shrouded in trade secrecy protection, which can impede public transparency and accountability.<sup>14</sup>

Beyond these concerns lies another. Acquiring AI is not business as usual: It often entails the procurement of policy choices from nongovernmental actors. AI systems are embedded with value-laden tradeoffs between what is technically feasible, socially acceptable, economically viable, and legally permissible. Thus, without intervention, the government will be acquiring technology with embedded policies from private actors whose financial motivations and legal sensitivities may not align with the government or the people it serves.

## AI systems are embedded with value-laden tradeoffs between what is technically feasible, socially acceptable, economically viable, and legally permissible.

Of course, the risks of harm are contextually contingent. It is one thing when an AI system misclassifies emails as spam or recommends purchasing more office supplies than needed. It is quite another when an AI system mistakenly deprives individuals of unemployment benefits,<sup>15</sup> automates the illegal seizure of tax refunds,<sup>16</sup> encroaches on personal privacy,<sup>17</sup> leads to wrongful arrest,<sup>18</sup> perpetuates racial and gender biases,<sup>19</sup> deprives access to government food programs,<sup>20</sup> impedes the right to travel,<sup>21</sup> and so on.

---

<sup>14</sup> See David S. Levine, *Secrecy and Unaccountability: Trade Secrets in Our Public Infrastructure*, 59 FLA. L. REV. 135 (2007); Sonia Katyal, *The Paradox of Source Code Secrecy*, 104 CORNELL L. REV. 1183, 1186–87 (2019) (explaining how “source code that underlies and governs automated decision making is hidden from public view, comprising an unregulated ‘black box’ that is privately owned and operated”).

<sup>15</sup> See Stephanie Wykstra & Undark, *It Was Supposed to Detect Fraud. It Wrongfully Accused Thousands Instead: How Michigan’s Attempt to Automate its Unemployment System Went Terribly Wrong*, THE ATLANTIC (June 7, 2020), <https://www.theatlantic.com/technology/archive/2020/06/michigan-unemployment-fraud-automation/612721/>.

<sup>16</sup> *Id.*

<sup>17</sup> See, e.g., Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1716–27 (2010) (showing that an individual’s identity may be reverse-engineered from a small number of data points); Tristan Greene, *Horrific AI Surveillance Experiment Uses Convicted Felons as Human Guinea Pigs*, TNW (Aug. 14, 2020, 5:40 PM), <https://thenextweb.com/neural/2020/08/14/horrific-ai-surveillance-experiment-uses-convicted-felons-as-human-guinea-pigs/>.

<sup>18</sup> Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. TIMES (Jun. 24, 2020), <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html>.

<sup>19</sup> See, e.g., Rashida Richardson, Jason Schultz, Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 NYU L. REV. ONLINE 15 (2019); SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Josh Feast, *4 Ways to Address Gender Bias in AI*, HARV. BUS. REV. (Nov. 20, 2019), <https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai>; Amazon Ditched AI Recruiting Tool that Favored Men for Technical Jobs, THE GUARDIAN (Oct. 10, 2018), <https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>.

<sup>20</sup> See Florangela Davila, *USDA Disqualifies Three Somalian Markets from Accepting Federal Food Stamps*, SEATTLE TIMES (Apr. 10, 2002), <http://community.seattletimes.nwsource.com/archive/?date=20020410&slug=somalis10m>.

<sup>21</sup> See generally JEFFREY KAHN, MRS. SHIPLEY’S GHOST: THE RIGHT TO TRAVEL AND TERRORIST WATCHLISTS (2013); see also Latif v. Holder, 28 F. Supp. 3d 1134, 1153 (D. Or. 2014) (ordering the agency to “fashion new procedures that provide plaintiffs with the requisite due process . . . without jeopardizing national security”).

More than a marketplace, the acquisition gateway must be reimagined as a policymaking space for promoting trustworthy and ethical AI. Toward that objective, this report offers a set of legal prescriptions that aim to align federal procurement law with the imperatives of ethical algorithmic governance.

**More than a marketplace, the acquisition gateway must be reimagined as a policymaking space for promoting trustworthy and ethical AI.**

First, federal lawmakers should mandate the creation of a government-wide inventory report that includes clear information on each AI system used by federal agencies. Currently, policymakers and stakeholders are wrangling about algorithmic governance, including whether AI tools such as facial recognition should even be permitted.<sup>22</sup> But an informed policy debate is impossible without knowledge about which AI tools have already been adopted by which agencies, for what purposes, from which vendors, and at what cost.

Second, federal lawmakers should require that agencies prepare “AI risk assessment” reports prior to the government’s acquisition of AI tools and services. These risk assessments would foreground several challenges and vulnerabilities that inhere in AI systems—most notably, relating to transparency, accountability, fairness, privacy, and safety.

Third, federal lawmakers should integrate ethical AI considerations into existing regulations for source selection and contractual award. Currently, nothing prevents federal contracting officials from soliciting and evaluating competing bids with an eye toward ethical AI. That is not the general practice, however, and it should be required as matter of law. Doing so will force agency officials and vendors to think more critically—and competitively—about the AI systems passing through the acquisition gateway. Less directly, yet as importantly, the government’s purchasing power and virtue signaling can spur market innovation and galvanize public trust in AI technologies.

---

<sup>22</sup> The use of facial recognition AI technology in law enforcement, for example, is arguably inappropriate because of technological and human limitations. Recent proposals in Congress would create a moratorium on the use of such technology by law enforcement. See Facial Recognition and Biometric Technology Moratorium Act of 2020, S. 4084, 116th Cong. (2020) (as referred to S. Comm. on the Judiciary, June 25, 2020).

## II. AI Today

AI has no universally accepted definition.<sup>23</sup> That dissensus owes in part to the wide cache of technologies that AI envisages. AI's definitional problem also reveals something about the concept itself: AI sweeps across fields of computer science, mathematics, psychology, sociology, neuroscience, and philosophy, and intersects with countless more. While disagreement persists about what AI means, there is wide consensus that civilization as we know it will never be the same.<sup>24</sup> Whether for better or worse is not the question. Instead, the questions are *whose lives* will be better and worse, in *which ways*, and under *what rules or conditions*.<sup>25</sup>

**While disagreement persists about what AI means, there is wide consensus that civilization as we know it will never be the same. Whether for better or worse is not the question. Instead, the questions are whose lives will be better and worse, in which ways, and under what rules or conditions.**

AI is disrupting every major market and facet of society. The technology is used in our phones, homes, cars, police stations, schools, social platforms, news feeds, satellites, workplaces, voting booths, and weapons systems. The unprecedented growth and dissemination of AI over the past decade owes to the conflation of several sociotechnical

---

<sup>23</sup> See U.S. Gov't Accountability Office, GAO-18-142SP, Artificial Intelligence: Emerging Opportunities, Challenges, and Implications (2018) [hereafter GAO, Artificial Intelligence] (observing "there is no single universally accepted definition of AI, but rather differing definitions and taxonomies"). One provision of U.S. law broadly defines AI to include the following:

- (1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.
- (2) An artificial system developed in computer software, physical hardware, or another context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
- (3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
- (4) A set of techniques, including machine learning, designed to approximate a cognitive task.
- (5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting.

John S. McCain National Defense Authorization Act for Fiscal Year 2019, Pub. L. No. 115-232, 132 Stat. 1636, 1695 (Aug. 13, 2018) (codified at 10 U.S.C. § 2358, note).

<sup>24</sup> See, e.g., Organization for Economic Cooperation and Development, Artificial Intelligence in Society 122 (2019), [https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society\\_eedfee77-en](https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en) [hereafter OECD, Artificial Intelligence] (providing a comprehensive survey of the many ways that AI is projected to transform social structures and power dynamics across markets and borders); GAO, Artificial Intelligence, *supra* note 23.

<sup>25</sup> See VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 180–88 (2018); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 29–31 (2016).

developments: the availability of exponentially more data; increases in computing power; the democratization of the “internet of things” (mobile phones, tablets, etc.); and breakthroughs in “machine learning” research and development.

A full exposition of machine learning is beyond the scope of this report. But it will be important to unpack some of the key attributes of machine learning for the project ahead because, in many ways, the promises and perils of algorithmic governance are anchored to how machine learning systems are designed and operationalized.<sup>26</sup>

Stripped to its essentials, machine learning is (i) a statistical technique (ii) that learns from data (iii) to make classifications or predictions (iv) for new data inputs. For example, if the objective is to train a machine that distinguishes between pictures of cats and dogs, the machine can be fed thousands of labeled pictures of cats and dogs, learn the difference between them by finding correlations in the data, and generate an algorithmic model that can then be used to identify cats and dogs in real-world settings. Or, to predict home values, datasets of past home sales can be used to train an algorithmic model to predict the sales value of homes that are not in the training data.

**Stripped to its essentials, machine learning  
is (i) a statistical technique (ii) that learns  
from data (iii) to make classifications or  
predictions (iv) for new data inputs.**

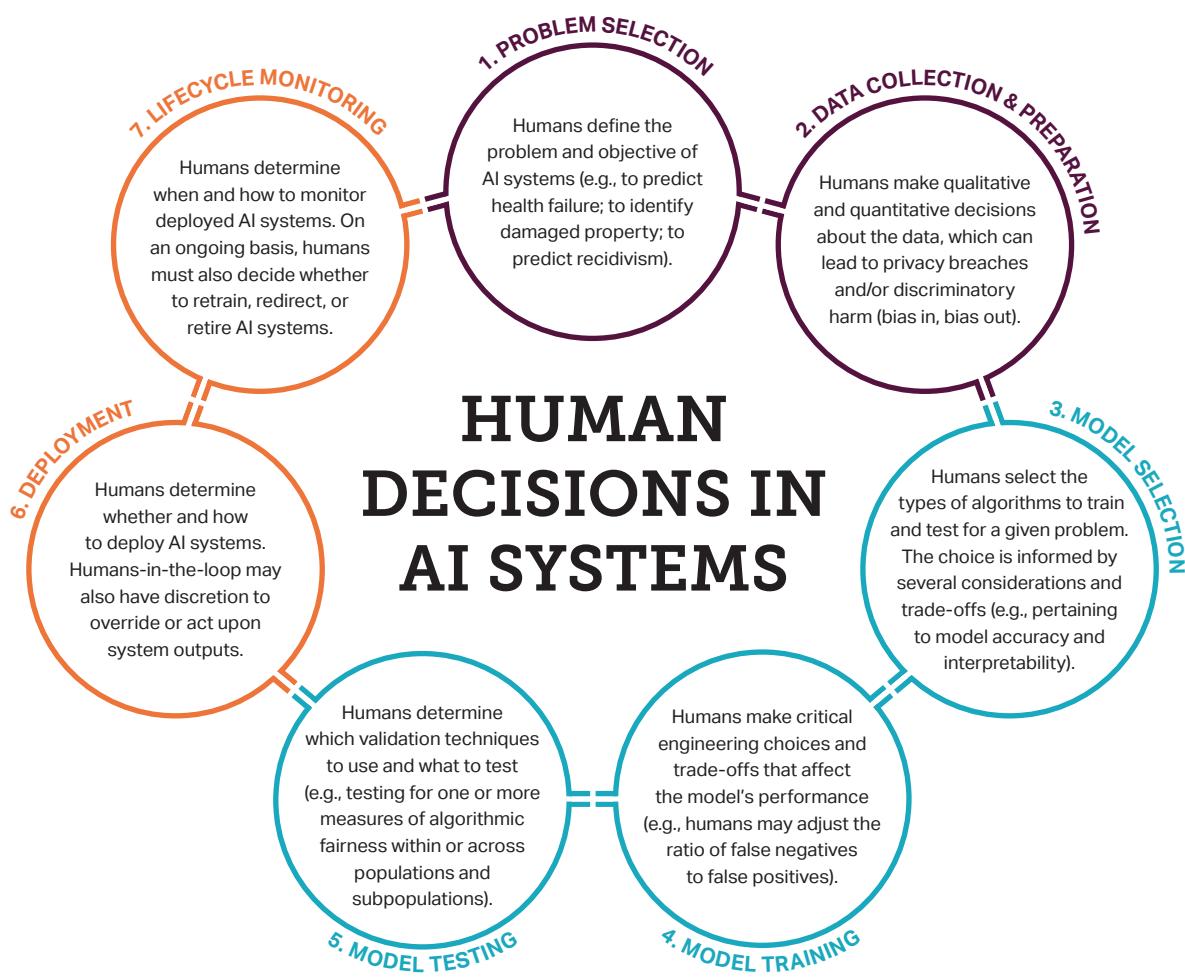
As these examples illustrate, machine learning is not a free-floating enterprise. Rather, it is part of a larger ecosystem comprised of data, humans, and human-computer interactions. For instance, humans generally select and clean the data, train and optimize machine learning algorithms, and deploy the algorithms in real-world or virtual settings. Moreover, humans make a wide variety of choices and trade-offs throughout the process. Just to name a few, humans must make choices about which datasets to include and exclude to train the model, which algorithmic model or models to use for a given task, which validation techniques to use, and which performance metrics to test for.<sup>27</sup>

---

<sup>26</sup> There are several different approaches to machine learning. For a short overview of the approaches, see Jatinder Singh et al., *Responsibility & Machine Learning: Part of a Process*, SSRN 4–9 (2016), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2860048](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2860048).

<sup>27</sup> *Id.*; see also Deirdre K. Mulligan & Kenneth A. Bamberger, *Procurement as Policy: Administrative Process for Machine Learning*, 34 BERKELEY TECH. L.J. 773, 778 (2019) (discussing embedded values choices in AI system design); see also NSCAI, Second Quarter Recommendations, *supra* note 12, at 129–31 (discussing “trade-off decisions for AI systems [that] must be made about internal representations, policies of usage and controls, run-time execution monitoring, and thresholds”).

Moreover, once a model is deployed, humans may be required to select and input new data. For example, an AI system designed to predict whether a criminal defendant is a high-risk recidivist will require a human to input features about the defendant (e.g., criminal history, age, home address).<sup>28</sup> Humans may also decide what to do, if anything, with the AI model's prediction or classification. In such systems, there is a so-called "human-in-the-loop." The human might be a judge, for example, who takes an algorithm's risk assessment into account when setting bail for a criminal defendant. Other AI systems, such as email spam filters, are called "autonomous" because human input is not required after the tool is deployed. In both types of systems, however, humans are responsible for many critical and consequential decisions throughout the AI lifecycle.



<sup>28</sup> AI systems are currently used throughout the country for this and other functions in the criminal justice system. See, e.g., Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018); Natalie Ram, *Innovating Criminal Justice*, 112 NW. U. L. REV. 659 (2018).

### III. Challenges of Algorithmic Governance

Under the right conditions, AI can make government more efficient and effective across a range of regulatory domains: from law enforcement to adjudication; from health care to building maintenance; from environmental protection to national defense; from policymaking to ministerial tasks.<sup>29</sup> Moreover, AI systems could potentially enhance government transparency, accountability, and fairness.<sup>30</sup> For example, AI decision-making systems might be more transparent and accountable than government officials who might conceal, or who might not be aware of, their decisional biases.<sup>31</sup> AI systems may also enable faster, more accurate, and more consistent decisions than humans, in contexts like social security, veterans benefits, immigration, and more. Furthermore, algorithmic governance may result in dramatic cost savings. According to one recent estimate, the federal government might save upwards of \$500 billion over the next decade by “automating repetitive tasks” and by “augmenting” the capabilities of public-sector workers.<sup>32</sup>

But glaring challenges persist. As discussed in more detail below, AI tools are inherently risky, irrespective of whether public or private actors are utilizing them. Yet the risks are exacerbated when AI tools are wielded by the federal government—not only because of the types of harms that can occur, but because expectations and legal requirements differ between public and private action. Federal action is governed by the Constitution, administrative law, freedom of information laws, and federal procurement laws, in ways that do not apply to private action.<sup>33</sup> The point is not that private actors have free rein;

---

<sup>29</sup> See GOVERNMENT BY ALGORITHM, *supra* note 2, at 6 (“Rapid developments in AI have the potential to reduce the cost of core governance functions, improve the quality of decisions, and unleash the power of administrative data, thereby making government performance more efficient and effective.”); Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 6 (2019) (describing how machine learning algorithms produce “unparalleled accuracy” compared to other statistical methods and human judgment).

<sup>30</sup> See David Freeman Engstrom & Daniel E. Ho, *Artificially Intelligent Government: A Review and Agenda*, in BIG DATA LAW (Roland Vogl ed., forthcoming 2020) (manuscript at 10) (“The perhaps counter-intuitive result is that the displacement of enforcement discretion by algorithm might, on net, yield an enforcement apparatus that is less opaque and more legible to agency heads and reviewing courts alike than the existing system.”); Kroll et al., *supra* note 5, at 656–77 (explaining how, through proper design, AI systems can be made more transparent and accountable).

<sup>31</sup> See Daniel Castro, *Data Detractors Are Wrong: The Rise of Algorithms is a Cause for Hope and Optimism*, CTR. FOR DATA INNOVATION (Oct. 25, 2016), <http://www.datainnovation.org/2016/10/data-detractors-are-wrong-the-rise-of-algorithms-is-a-cause-for-hope-and-optimism/>.

<sup>32</sup> See Christina Bone et al., *The Coming of AI Productivity Boom: And How Federal Agencies Can Make the Most of It*, ACCENTURE, 1, 4 (2020).

<sup>33</sup> See Daniel Guttman, *Public Purpose and Private Service: The Twentieth Century Culture of Contracting Out and the Evolving Law of Diffused Sovereignty*, 52 ADMIN. L. REV. 859, 862, 881–90 (2000) (explaining that “in practice, two different sets of regulations have come to govern those doing the basic work of government”: those that apply to federal officials, on the one hand, and those that apply to federal contractors, on the other). For an incisive treatment of the constitutional state action doctrine as applied to private AI vendors, see Kate Crawford & Jason Schultz, *AI Systems as State Actors*, 119 COLUM. L. REV. 1941, 1943–44 (2019) (arguing that courts should adopt a version of the state action doctrine to apply to vendors who supply AI systems for government decision-making).

they do not. Rather, the point is that legal norms around public and private action differ in ways that matter for algorithmic governance, especially as pertains to questions of transparency, accountability, privacy, safety, and fairness (broadly defined).<sup>34</sup>

## A. RISK OF HARM

Unlike calculators, algorithmic classifications and predictions can be wrong. Of course, human classifications and predictions can be wrong too. But the efficiencies and scalability of AI systems make them especially risky.<sup>35</sup> One coding error, engineering choice, unfounded assumption, or lapse in human oversight can cause widespread harm. Moreover, when exposed to real-world elements, AI systems make mistakes that most humans never would.<sup>36</sup> Compounding the risk, AI systems may interact with other technologies, humans, or environmental conditions in ways that can negatively bear on those surrounding systems.<sup>37</sup>

## B. TRANSPARENCY

AI systems raise a host of transparency challenges for algorithmic governance.<sup>38</sup> Technologically, some algorithms and design configurations are more scrutable than others. Machine learning “neural networks,” which are some of the most powerful, sophisticated, and useful, are also the most difficult for humans to comprehend. The inputs and outputs can be known, but the so-called neural network that turns inputs into outputs can entail millions of data correlations, at scales that the smartest minds on earth cannot understand, much less accurately explain to anyone else. Machine learning

---

<sup>34</sup> The term “fairness” is borrowed here from the AI field and has no agreed-upon meaning. See Abigail Z. Jacobs & Hannah Wallach, *Measurement and Fairness 1* (Microsoft, Working Draft No. 1912.05511, 2019), <https://arxiv.org/pdf/1912.05511.pdf>. A concept like “justice” may work just as well or better. For present purposes, what matters is the breadth of concerns that fairness (or justice) captures, including nondiscrimination, privacy, procedural due process, human rights, and more. See *infra* Subpart II.D.

<sup>35</sup> See Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J. L. & TECH. 103, 129 (2018) (“The ability of these algorithmic processes to scale, and therefore to influence decisions uniformly and comprehensively, magnifies any error or bias that they embody”); O’NEIL, *supra* note 25, at 29–31 (discussing the “scalability of algorithms”). For example, an efficient AI system that makes 100,000 predictions at a 10 percent error rate may negatively affect 1,000 individuals. An inefficient human who makes 100 recommendations at a 20 percent error rate may negatively affect 20 individuals.

<sup>36</sup> See Fabio Kepler, *Why AI Fails in the Wild*, UNBABEL (Nov. 15, 2019), <https://unbabel.com/blog/artificial-intelligence-fails/>; see also Colin Smith et al., *Hazard Contribution Modes of Machine Learning Components*, 2020 AAAI WORKSHOP ON ARTIFICIAL INTELLIGENCE SAFETY 4 (2020) (discussing unexpected performance, for example, “through unanticipated feature interaction . . . that was also not previously observed during model validation”), <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20200001851.pdf>.

<sup>37</sup> Singh et al., *supra* note 26, at 15.

<sup>38</sup> See Jenna Burrell, *How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC’Y 1, 3–5 (2016).

models of this type are metaphorical “black boxes.” While they can perform well on the dimensions of efficiency and functionality, the complexity of the algorithmic models can eclipse human-scale reasoning.

Worth emphasizing here, model inscrutability is a feature of complex AI systems, not a bug. After all, the ambition and promise of machine learning is not to think as humans do, but rather to find statistical correlations in big datasets beyond human cognition or intuition. In some settings, the inscrutability of the model’s decisional pathway may be of little concern. But for consequential governmental decisions, the inscrutability of AI systems is highly concerning. “[T]he algorithm made me do it” will not satisfy expectations for a human-centric explanation.<sup>39</sup>

## For consequential governmental decisions, the inscrutability of AI systems is highly concerning.

Without clarity about how or why an AI system makes a prediction, the government may fail in its responsibility to provide a legally or morally sufficient reason for acting on it. Thus, the government may want or need to deploy “interpretable” and “explainable” AI tools.<sup>40</sup> Its ability to do so will depend on a variety of dynamic—and yet unsettled—legal, technological, and sourcing contingencies.

Trade secrecy can also interfere with algorithmic transparency.<sup>41</sup> In some contexts, the government may not have access to a vendor’s trade secrets—for example, when the government purchases the technology as a “commercial item off the shelf.”<sup>42</sup> In other contexts, the government may have access to a vendor’s trade secrets, but federal law

---

<sup>39</sup> See, e.g., Fairness, Accountability, and Transparency in Machine Learning, [www.fatml.org](http://www.fatml.org) (“[T]here is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to ‘the algorithm made me do it.’”); Victoria Burton-Harris & Philip Mayor, *Wrongfully Arrested Because Face Recognition Can’t Tell Black People Apart*, ACLU (June 24, 2020) (“One officer responded, ‘The computer must have gotten it wrong.’”), <https://www.aclu.org/news/privacy-technology/wrongfully-arrested-because-face-recognition-can-tell-black-people-apart>.

<sup>40</sup> See, e.g., P. Jonathan Phillips et al., Nat’l Ins. Sci. & Tech, *Four Principles of Explainable Artificial Intelligence* (Aug. 2020), <https://www.nist.gov/system/files/documents/2020/08/17/NIST%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf>; GOOGLE LLC, AI Explanations Whitepaper 1–28 (2019), <https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf>; The Royal Society, Explainable AI: The Basics Policy Brief (2019), <https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf>.

<sup>41</sup> See FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Levine, *supra* note 14, at 180-81.

<sup>42</sup> See, e.g., 48 C.F.R. § 12.212 (providing, for the acquisition of commercially available computer software, that vendors generally “shall not be required” to “[r]elinquish to, or otherwise provide, the Government rights to use, modify, reproduce, release, perform, display, or disclose commercial computer software or commercial computer software documentation except as mutually agreed to by the parties”).

or nondisclosure agreements may prevent the government from revealing the secrets to third parties.<sup>43</sup> As a result, information about an AI system may be withheld from individuals affected by the system, government watchdogs, and perhaps even judges and lawmakers.<sup>44</sup>

The right to privacy is another friction point.<sup>45</sup> Most notably, the data ingested by AI systems may contain sensitive personal information that can be traced to individuals (even if names and other identifying attributes are scrubbed from the data).<sup>46</sup> AI systems may thus run headlong into privacy laws, in a wide variety of contexts where the government is provided personal information for specific purposes that may not be used or disclosed for other purposes.<sup>47</sup>

Moreover, in law enforcement and national security contexts, full transparency about AI systems may be self-defeating or yield bad outcomes<sup>48</sup> because AI systems can be “gamed” or “hacked” by adversarial actors in ways that humans cannot.<sup>49</sup> Of course, humans can be manipulated, bribed, or spied upon in ways that undermine or endanger the public interest. The rub, however, is that AI systems have similar human vulnerabilities (e.g., data scientists, computer programmers, technical engineers,

---

<sup>43</sup> See *id.*; Katherine Fink, *Opening the Government’s Black Boxes: Freedom of Information and Algorithmic Accountability*, 21 INFO., COMM. & SOC’Y 1453 (2017) (reviewing current state of law and practice with respect to whether algorithms would be considered “records” under the Freedom of Information Act (FOIA) and reviewing agency bases for withholding algorithms and source code under FOIA requests).

<sup>44</sup> See Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM L. REV. 1265, 1299–1302 (2020) (discussing impediments to algorithmic transparency under FOIA and trade secrecy laws, which hinder public interest groups and watchdogs from obtaining information about AI tools used by government).

<sup>45</sup> See Engstrom & Ho, *Artificially Intelligent Government*, *supra* note 30, at 11–12 (canvassing a range of federal privacy laws to explain how “privacy and data security constraints, while designed to safeguard privacy and minimize public burdens, can also impose significant costs on agencies, reduce the efficacy of algorithmic tools, and stymie agency innovation”).

<sup>46</sup> See Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 111 (2008). This is a major concern, especially in light of malicious threats to information security. See, e.g., Zolan Kanno-Youngs & David E. Sanger, *Border Agency’s Images of Travelers Stolen in Hack*, N.Y. TIMES (June 10, 2019), <https://www.nytimes.com/2019/06/10/us/politics/customs-data-breach.html>; Julie Hirschfield Davis, *Hacking of Government Computers Exposed 21.5 Million People*, N.Y. TIMES (July 9, 2015), <https://www.nytimes.com/2015/07/10/us/office-of-personnel-management-hackers-got-data-of-millions.html>.

<sup>47</sup> See, e.g., 5 U.S.C. §§ 552a(b) & (e)(3) (prohibiting disclosure of records without the prior written consent of the person whom the records pertain to, excepting for reasons such as routine use for, *inter alia*, census purposes, matters of the House of Congress or any of its committees or subcommittees, etc.); Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (1996) (codified as amended in scattered sections of 18, 26, 29 and 42 U.S.C. (1996) (setting forth privacy and security standards for protecting personal health information).

<sup>48</sup> See Engstrom & Ho, *supra* note 30 (explaining that “in enforcement context[s] . . . transparency facilitates strategic action that can drain [AI] tools of their value”); see also 5 U.S.C. § 552(b)(7)(E) (exempting “records or information compiled for law enforcement purposes” whose disclosure “could reasonably be expected to risk circumvention of the law”). But cf. Ignacio N. Cafone & Katherine J. Strandburg, *Strategic Games and Algorithmic Secrecy*, 64 MCGILL L.J. (forthcoming 2020) (arguing that the range of situations in which people are able to game decision-making algorithms is narrow, even when there is substantial disclosure), [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3440878](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3440878).

<sup>49</sup> In machine learning literature, the gaming problem is associated with “adversarial learning”—the problem of developing models when it is anticipated from the beginning that adversaries will try to defeat them. Guofu Li et al., *Security Matters: A Survey on Adversarial Machine Learning*, ARXIV (2018), <https://arxiv.org/abs/1810.07339>.

domain experts) *plus* technical vulnerabilities in the code, cloud, or hardware that can be exploited in ways that elide easy detection. For similar reasons, cybersecurity concerns will generally trump algorithmic transparency around critical infrastructure: in energy, transportation, telecommunication, voting systems, waterways, etc.

**AI systems have similar human vulnerabilities  
(e.g., data scientists, computer programmers,  
technical engineers, domain experts) plus technical  
vulnerabilities in the code, cloud, or hardware that  
can be exploited in ways that elide easy detection.**

As these examples demonstrate, there is a lot of algorithmic governance that could be off-limits to full transparency. Yet AI systems are already being utilized across all of these domains, with many more in the works.

## C. ACCOUNTABILITY

The transparency concerns in algorithmic governance are directly related to a set of accountability concerns. The less stakeholders know, the more difficult it becomes to ascertain whether the human inputs and machine outputs are accurate, fair, and legal. And without those determinants, stakeholders cannot know which actors, if any, should be held accountable for any resulting harms.

One way that our system holds government actors accountable is through judicial review. The use of AI systems in government decision-making, however, can stymie a court's ability to know or understand the reasons for an agency's action if the agency (or its vendors) cannot adequately explain why an AI tool made a particular prediction or classification that led to a government decision.<sup>50</sup> Beyond judicial settings, government watchdogs, journalists, and stakeholders are similarly constrained in their ability to "look under the hood" of AI tools affecting the polity's rights and interests.<sup>51</sup> This is highly

---

<sup>50</sup> See Rebecca Wexler, *When a Computer Program Keeps You in Jail*, N.Y. TIMES (June 13, 2017) ("The root of the problem is that automated criminal justice technologies are largely privately owned and sold for profit."), <https://www.nytimes.com/2017/06/13/opinion/howcomputers-are-harming-criminal-justice.html>; see also Sonia K. Katyal, *The Paradox of Source Code Secrecy*, 104 CORNELL L. REV. 1183, 1186–87 (2019) (expounding on this concern beyond litigation settings).

<sup>51</sup> See Bloch-Wehba, *supra* note 44 (emphasizing the importance of open records laws for third-party stakeholders to hold government accountable outside of litigation settings); Brauneis & Goodman, *supra* note 35, at 159 (complaining that "the information allegedly protected by trade secret law may lie at the heart of essential public functions and constitute political judgments long open to scrutiny").

problematic because it shutters out stakeholder input and breeds public distrust in government uses of the technology.

## D. FAIRNESS

The transparency and accountability gaps in algorithmic governance are especially concerning because the use of AI may not be procedurally or substantively fair.

Procedurally, AI systems raise due process concerns that, to date, have gone unmet in several high-profile cases.<sup>52</sup> In groundbreaking work, Danielle Citron provides an account of the “automated administrative state” using software to determine whether someone should receive “Medicaid, food stamp, and welfare” benefits, be on a no-fly list, or be identified as owing child support.<sup>53</sup> As she cogently explains, “[a]utomation jeopardizes the due process safeguards owed individuals and destroys the twentieth-century assumption that policymaking will be channeled through participatory procedures that significantly reduce the risk that an arbitrary rule will be adopted.”<sup>54</sup> These concerns, aired more than a decade ago, have only intensified during the intervening years because machine learning algorithms, which are on the rise, are generally more complex and less explainable.<sup>55</sup>

**Without proper precautions, and even with precautions, technical bias can reify social biases in the analog world.**

Beyond procedural unfairness, the use of AI tools can be unfair for technological, legal, or moral reasons.<sup>56</sup> Algorithmic models are statistical simplifications that cannot consider all possible relevant facts about subjects. Generalizations and profiling thus typify AI systems.<sup>57</sup> While AI-generated predictions are not inherently unfair,

---

<sup>52</sup> See Rashida Richardson, Jason M. Schultz, Vincent M. Southerland, *Litigating Algorithms*, AI NOW INSTITUTE (2019), <https://ainowinstitute.org/litigatingalgorithms-2019-us.pdf>.

<sup>53</sup> Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1256 (2008).

<sup>54</sup> *Id.* at 1256–57.

<sup>55</sup> Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. (forthcoming 2020), [https://papers.ssrn.com/abstract\\_id=3553590](https://papers.ssrn.com/abstract_id=3553590) (“In the decade since the publication of *Technological Due Process*, governments have doubled down on automation despite its widening problems.”); see also Mulligan & Bamberger, *supra* note 27, at 814–18 (explaining how the machine learning tools of today are even more problematic from a legal standpoint than the AI tools that were the focus of Citron’s original study).

<sup>56</sup> See, e.g., Niraesh Mehrabi et al., *A Survey on Bias and Fairness in Machine Learning*, ARXIV (2019), <https://arxiv.org/pdf/1908.09635.pdf>.

<sup>57</sup> See Mulligan & Bamberger, *supra* note 27, at 787 (“Predictive algorithms are essentially autonomous profiling by a machine-learning system.”).

they will always be biased, in a technical sense, because the predictions are based on generalizations mined from data. Without proper precautions, and even with precautions, technical bias can refract and reify social biases in the analog world.

For example, an algorithm trained on historical data to predict criminal recidivism may exhibit higher false positive rates for Black people if the training data is an artifact of past discriminatory policing against that population.<sup>58</sup> Likewise, an AI system designed to predict a “good hire” for a government position may make predictions based on promotion rates and employee evaluations of past government hires. If those input variables in the training data are biased toward men, then algorithmic predictions for future “good hires” will be biased against women.<sup>59</sup>

Separately, but relatedly, choices in model engineering can result in discrimination. Of particular concern are choices about feature selection in the training and testing data.<sup>60</sup> For example, an AI model may be explicitly trained to account for race or sex in decision-making contexts where it is illegal to discriminate on those grounds. Even if an algorithm is trained to ignore race or sex, proxies for those attributes might intentionally or unintentionally be extracted from the data and lead to the same results. For example, zip codes are a well-known proxy for race.<sup>61</sup>

For sensitive government decisions, judges and civil servants may be expected to exercise human judgment as a check on algorithmic predictions. In practice, however, studies reveal the risk of “automation bias,” whereby humans-in-the-loop exhibit overconfidence in algorithmic predictions and classifications.<sup>62</sup> Automation bias raises the stakes of any unfairness baked into the algorithm itself because humans may not correct for errors. But risks also run in the opposite direction: When a human-in-the-loop compensates for known algorithmic biases, the human interventions may not

---

<sup>58</sup> See, e.g., Richardson et al., *supra* note 19; Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), <https://perma.cc/JRR9-5D29>; cf. Avi Feller et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear*, WASH. POST: MONKEY CAGE (Oct. 17, 2016), <https://perma.cc/7M7V-GPKL>.

<sup>59</sup> Cf. *Amazon Ditched AI Recruiting Tool that Favored Men for Technical Jobs*, THE GUARDIAN (Oct. 10, 2018), <https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>.

<sup>60</sup> See Nicholas Diakopoulos, *Algorithmic Accountability: Journalistic Investigation of Computational Power Structures*, 3 DIGITAL JOURNALISM 398, 400–02 (2015) (discussing the value choices embedded in data prioritization, classification, association, and filtering).

<sup>61</sup> Rhema Vaithianathan et al., *Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation*, CTR. FOR SOC. DATA ANALYTICS 12 (2017) (discussing zip codes and other proxies for race), <https://www.allegenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf>.

<sup>62</sup> See Citron, *supra* note 53, at 1271–72 (discussing “automation bias”); Kate Goddard, Abdul Roudsari, Jeremy C. Wyatt, *Automation Bias: Empirical Results Assessing Influencing Factors*, 83 INT'L J. MED. INFORMATICS 368 (2014).

be fair or legal. For instance, if an algorithm used for government hiring exhibits bias toward hiring men, to what extent (if any) can the algorithmic prediction be favorably adjusted toward hiring women (without running afoul of equal-protection principles)?<sup>63</sup> As this example illustrates, fairness depends on definitions of fairness—of which there are many—and legal principles that do not map easily on AI systems.<sup>64</sup>

Another source of algorithmic bias stems from the lack of diversity in the technology industry.<sup>65</sup> A notorious example is facial recognition technology that was trained on the faces familiar to the design engineers, which were mostly white.<sup>66</sup> Consequently, the facial recognition software had a greater propensity to misidentify dark-skinned faces.<sup>67</sup> Needless to say, the disparity is especially concerning in high-stakes contexts such as law enforcement. Recently, police in Detroit wrongfully arrested a Black man at home in front of his family, including his young daughters.<sup>68</sup> The charges were subsequently dismissed after the “officers-in-the-loop” acknowledged the misidentification, but the damage was already done.<sup>69</sup>

DHS and the FBI use similar facial recognition technology, in conjunction with other AI-enabled surveillance tools, in immigration and other law enforcement.<sup>70</sup> Even if the technology can be made more accurate—for example, by correcting for known biases in the training data—that would not address the overarching concerns relating to power, autonomy, and liberty. Having the technical capability for highly accurate surveillance says nothing about whether, or for what purposes, that capability should be wielded.

---

<sup>63</sup> See Solon Barocas & Andrew D. Selbst, *Big Data’s Disparate Impact*, 104 CALIF. L. REV. 671, 694–714 (2016) (noting the ways in which algorithmic data mining techniques can lead to unintentional discrimination against historically prejudiced groups).

<sup>64</sup> See Deirdre K. Mulligan et al., *This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology*, PROC. 2019 ACM ON HUMAN-COMPUTER INTERACTION 3, 119 (2019). For example, fairness can be operationalized around group metrics or individual metrics, of which there are many types of each. For a discussion, see *id.*

<sup>65</sup> See RUHA BENJAMIN, RACE AFTER TECHNOLOGY: ABOLITIONIST TOOLS FOR THE NEW JIM CODE (2019) (arguing that human social bias is engineered into automated technology because (overwhelmingly white and male) programmers fail to recognize how their understanding of technology is informed by their identities, and the raw data on which robots of all types are trained are products of racist, sexist, and classist societies).

<sup>66</sup> See, e.g., Jacob Snow, *Amazon’s Face Recognition Falsely Matched 28 Members of Congress with Mugshots*, ACLU (July 26, 2018), <https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28>.

<sup>67</sup> See *id.*

<sup>68</sup> See Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. TIMES (Jun. 24, 2020), <https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html>.

<sup>69</sup> *Id.* (“[T]he Wayne County prosecutor’s office said that . . . Williams could have the case and his fingerprint data expunged. ‘We apologize,’ . . . ‘This does not in any way make up for the hours that Mr. Williams spent in jail.’”).

<sup>70</sup> Aaron Boyd, *CBP Is Upgrading to a New Facial Recognition Algorithm in March*, NEXTGOV.COM (Feb. 7, 2020), <https://www.nextgov.com/emerging-tech/2020/02/cbp-upgrading-new-facial-recognition-algorithm-march/162959/>; Kimberly J. Del Greco, Deputy Assistant Dir., Crim. Justice Info. Serv. Div., FBI, Statement Before the House Oversight and Reform Committee, Facial Recognition Technology: Ensuring Transparency in Government Use (Jun. 4, 2019).

More generally, AI may simply be an inappropriate solution for many government problems. This is especially true for AI systems trained to predict “unobservable theoretical constructs,” which by definition are neither observable nor verifiable.<sup>71</sup> Instead, such constructs must be inferred from observable properties that a system designer thinks are closely (enough) related.<sup>72</sup> To illustrate, compare the constructs of “crime” and “criminality.” Both are government problems, but the use of AI to address them differs dramatically. Using AI to detect *crime* is not inherently problematic because humans can verify whether, in fact, criminal activity has occurred or is occurring.

The same cannot be said about a person’s *criminality*. In a recent academic paper, the authors boasted of an AI tool “capable of predicting whether someone is likely going to be a criminal” with “80 percent accuracy and with no racial bias.”<sup>73</sup> In response, thousands of AI experts and practitioners published an open letter that condemned the paper and urged the journal not to publish it.<sup>74</sup> As the open letter explains, criminality based on a person’s facial features “reproduces injustice and causes real harm.”<sup>75</sup> Criminality itself is a racialized construct, and to infer it based on immutable characteristics of a person’s face is an affront to moral justice.<sup>76</sup> One can easily imagine how an AI tool to predict criminality based on a person’s appearance can be used inappropriately by government actors to surveil, detain, or deny rights and access to government services.<sup>77</sup>

---

<sup>71</sup> See Abigail Z. Jacobs & Hannah Wallach, *Measurement and Fairness* 1 (Microsoft, Working Draft No. 1912.05511, 2019), <https://arxiv.org/pdf/1912.05511.pdf>.

<sup>72</sup> *Id.*

<sup>73</sup> *HU Facial Recognition Software Predicts Criminality*, HARRISBURG UNIV. OF SCI. AND TECH. (May 5, 2020), <https://web.archive.org/web/20200506013352/https://harrisburgu.edu/hu-facial-recognition-software-identifies-potential-criminals/>.

<sup>74</sup> See Coal. for Critical Tech., *Abolish the #TechToPrisonPipeline: Crime Prediction Technology Reproduces Injustices and Causes Real Harm*, MEDIUM (June 23, 2020), <https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16>.

<sup>75</sup> *Id.*

<sup>76</sup> *Id.* (“Data generated by the criminal justice system cannot be used to ‘identify criminals’ or predict criminal behavior. Ever.”). Following the backlash, the authors later deleted the paper, and their publisher, Springer, confirmed that it had been rejected. Sidney Fussell, *An Algorithm That ‘Predicts’ Criminality Based on a Face Sparks a Furor*, WIRED (June 24, 2020, 7:00 AM), <https://www.wired.com/story/algorithm-predicts-criminality-based-face-sparks-furor/>.

<sup>77</sup> Conceivably, and assuming accuracy, an AI tool like this could be used for putatively beneficent reasons—for instance, to offer community support for those with facial features that bespeak criminality. Still, that type of profiling would almost certainly be an unwelcome (and unlawful) affront to privacy and dignity.

## IV. Recommendations

The proliferation of AI use by federal agencies has raised urgent questions about how algorithmic governance itself should be regulated. As noted earlier, several bills recently introduced in Congress relate, in some way, to the government's use of AI technologies. Moreover, a range of policy initiatives are underway or under consideration within the executive branch. To greater and lesser extents, the prescriptions on the table speak to the core challenges of algorithmic governance: risk of harm, transparency, accountability, privacy, security, and fairness.

By no means is federal procurement law the sole solution to these complex problems. But procurement must be part of the solution. Currently, the government is investing huge sums of taxpayer dollars to acquire AI systems that may be unusable, either because they are not trustworthy, or because they fail to pass legal muster. If the government cannot explain how an AI system works, for example, then it may violate constitutional due process<sup>78</sup> or administrative law.<sup>79</sup> Even if an AI system clears those hurdles, it may still violate federal antidiscrimination laws, privacy laws, and domain-specific laws and regulations. Litigation will no doubt surface these risks and harms. But much of that screening can occur, *ex ante*, through the acquisition gateway.

**Currently, the government is investing huge sums of taxpayer dollars to acquire AI systems that may be unusable, either because they are not trustworthy, or because they fail to pass legal muster.**

In a recent report, the National Security Commission on Artificial Intelligence recommended a set of best practices "in support of core American values" for the "responsible development and fielding AI technologies" by the government.<sup>80</sup> Of

---

<sup>78</sup> See, e.g., Citron, *supra* note 53; Aziz Z. Huq, *Constitutional Rights in the Machine Learning State*, 105 CORNELL L. REV. (forthcoming 2020); Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327 (2015); Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871 (2016).

<sup>79</sup> See, e.g., David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. REG. 800 (2020); Mulligan & Bamberger, *supra* note 28; Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147 (2017).

<sup>80</sup> NSCAI, Second Quarter Report, *supra* note 12, at 120–55 (Appendix 2). Those American values include the "rule of law," "[v]values established in the U.S. Constitution, and further operationalized in legislation, include freedoms of speech and assembly, the rights to due process, inclusion, fairness, nondiscrimination (including equal protection), and privacy (including protection from unwarranted government interference in one's private affairs), as well as international human rights and dignity." *Id.* at 125. See also *id.* at 126–27 (discussing additional values in the context of the military and warfighting).

particular note here, the commission highlighted the challenges around ethical AI, and stressed that systems acquired from contractors “should be subjected to the same rigorous standards and practices—whether in the acquisitions or acceptance processes.”<sup>81</sup> Clearly, there is a demand for ethical AI in government. Just as surely, federal procurement law can be harnessed in service of that critical mission.

## A. TAKING STOCK

Currently, there is no easy way to identify which agencies are using which AI technologies, for what purposes, from which vendors, and under what constraints. Without such information, policymakers, watchdogs, and stakeholders cannot know whether the systems in use are transparent, accountable, fair, secure, testable, and so on.<sup>82</sup>

As a first step, federal lawmakers should mandate the creation of a government-wide inventory report that includes clear information about each AI system that is currently in production or use by the federal government. The Office of Management and Budget, the General Services Administration, or some other centralized agency should be assigned responsibility for the oversight and reporting of the inventory, which should be publicly available and updated annually. Exceptions can be made as needed for national security, intelligence agencies, or otherwise.

As a general rule, however, the disclosure requirements should include information about all AI systems that the government is currently using in adjudicatory, rulemaking, or law enforcement settings that affect a person’s rights, duties, or access to public benefits and services, or that may otherwise cause property damage or personal injury.<sup>83</sup> The disclosure requirements should also include AI systems that are used for certain internal operations, such as hiring, resource allocation, and other nontrivial ministerial tasks. Too often, internal agency operations get overlooked in discussions about algorithmic governance because they generally are not judicially reviewable under framework statutes such as the Administrative Procedure Act. But judicial reviewability is not the purpose of

---

<sup>81</sup> *Id.* at 124.

<sup>82</sup> Along these lines, the Artificial Intelligence Reporting Act of 2018, H.R. 6090, calls for annual congressional reporting “on the use and status of unclassified machine learning and artificial intelligence applications across the Federal Government,” from the subgroup on Machine Learning and Artificial Intelligence of the Committee on Technology of the National Science and Technology Council. This bill has not progressed beyond the House Committee on Science, Space, and Technology.

<sup>83</sup> For examples of pending legislation that offer definitions of “high risk” automated decision-making systems, see Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. § 2 (2019) (defining the term in connection with the regulation of commercial uses of AI technologies).

the inventory. And, in any event, the triggers for judicial review do not capture the scope of agency activity that policymakers and the public have an interest to know about. For example, if an AI system is biased in hiring decisions, or in how government resources are allocated, those internal managerial tools can surely have a detrimental effect on government performance and public policy.

For any nonexempt systems, the minimum reporting requirements should include:

1. A list of all vendors or nongovernmental partners that have participated in the design, development, testing, auditing, deployment, or use of the AI system. This should include a description of each vendor or partner role, and the total amount of federal funds paid or contractually committed for the work.
2. A description of the AI system's general capabilities and intended purposes, including reasonably foreseeable capabilities and purposes outside the scope of the agency's intended use.
3. A description of the training and testing data of an AI system; how that data was generated, collected, and processed; the type or types of data that the system is reasonably likely to generate; and whether the government, vendors, or third parties have access to or control over that generated data.
4. A description of whether the algorithmic model used in an AI system can be altered by its users; and if so, by whom, under what circumstances and with what disclosures, and with what safeguards to protect the integrity of the model and traceability of any alterations.
5. A description of whether the AI system has been tested for bias (discriminatory or otherwise); whether the testing was performed by a third party; the date and methodology employed for such testing; whether the AI system has any known biases, and if so, how those biases are accounted for, corrected, tracked, or managed.
6. A description of whether the AI system gives notice to impacted individuals when it is used to collect personal information about those individuals or is used in decision-making that involves them.
7. For any piece of information above that is not disclosed, provide specific reason(s) why the information either cannot or will not be disclosed for purposes of this reporting, and whether the information may be available in other forms or by other means.

This inventory report would not be limited to AI systems acquired from nongovernmental vendors. But it would capture those systems that have already

passed through the acquisition gateway, which is notoriously opaque to all but the contracting community (and murky, too, for many within that community).<sup>84</sup> By making this information publicly available, and on a rolling basis, interested stakeholders can help agencies bridge information gaps and identify sociotechnical blind spots in the government's uses of AI systems. As earlier noted, the government does not have the in-house capacity needed to assess, anticipate, and address the challenges of algorithmic governance. By voluntarily sharing this information, the government can leverage the benefit of public input from scientists, academics, laypersons, interest groups, journalists, and members of government who do not have ready access to the information that an inventory report could provide.

## B. FEDERAL ACQUISITION POLICY AND PRACTICE

There are many ways that procurement law may be retrofitted to meet the challenges of algorithmic governance. By way of background, the Federal Acquisition Regulation (FAR) is "the primary regulation for use by all Federal Executive agencies in their acquisition of supplies and services with appropriated funds."<sup>85</sup> FAR addresses various aspects of the procurement process, including acquisition planning, contract formation, source selection, auditing, and contractual management. Moreover, FAR captures a range of objectives, including market competition, integrity, transparency, efficiency, government satisfaction, best value, wealth distribution, and risk avoidance.<sup>86</sup> These pluralistic values do not always align; trade-offs among them are necessary and inevitable. FAR makes those trade-offs through a range of procurement processes, contractual provisions, performance incentives, accountability mechanisms, and a mix of delegated discretion and nonnegotiable directives.<sup>87</sup>

In addition to FAR, the federal acquisition system is governed by congressional statutes, agency regulations, guidance documents, and presidential executive orders. Thus, adjustments to federal procurement law—including those recommended here—can come from Congress, the White House, or designated agencies. Moreover, the

---

<sup>84</sup> Cf. Jarrod McAdoo, *How New Initiatives Might Make Federal Sales Easier*, WASH. TECH. (Jun. 2, 2020) ("The procurement process creates a jungle of barriers including crushing complexity, awkward communications and significant expense just to try and compete for business.").

<sup>85</sup> Gen. Servs. Admin. et al., *Foreword to Federal Acquisition Regulation*, at v (2019), <https://www.acquisition.gov/sites/default/files/current/far/pdf/FAR.pdf>.

<sup>86</sup> See Steven L. Schooner, *Desiderata: Objectives for a System of Government Contract Law*, 11 PUB. PROCUREMENT L. REV. 103 (2002).

<sup>87</sup> See Steven L. Schooner, *Fear of Oversight: The Fundamental Failure of Businesslike Government*, 59 AM. U. L. REV. 627 (2001).

recommendations below can be taken up separately or in combination. Regardless, they are designed to be cohesive and interoperable with other lawmaking initiatives around algorithmic governance.

## a. Pre-Acquisition AI Risk Assessment

Federal law requires that agencies conduct safety risk assessments pertaining to information systems.<sup>88</sup> This requirement, and others like it, are designed to focus agency attention on risk factors in sensitive government contexts so that risks can be addressed and mitigated on an ongoing basis.<sup>89</sup> Currently, no such requirement exists for AI systems in particular. But, given the challenges that inhere in algorithmic governance, federal lawmakers should require agencies to develop and utilize AI-specific risk assessments as part of the acquisition process.<sup>90</sup>

This requirement could fit neatly within current procurement law and practice. FAR already mandates that agencies engage in acquisition planning, which incorporates special considerations beyond mere dollars and cents. For example, agency planners are required to comply with pre-established "Guiding Principles" for green-energy building construction and renovation.<sup>91</sup> Likewise, procurement law can explicitly require that agencies prepare AI risk assessments tailored for the unique challenges of algorithmic governance.

Importantly, the risk assessment should be conducted by a multidisciplinary team that includes agency acquisition and IT personnel, domain experts, legal experts, sociotechnical ethicists, and data specialists. Moreover, as much as possible, the team

---

<sup>88</sup> 44 U.S.C. § 3554(b) ("Each agency shall develop, document, and implement an agency-wide information program to provide periodic assessments of the risk and magnitude of the harm that could result from the unauthorized access, use, disclosure, disruption, modification, or destruction of information and information systems that support the operations and assets of the agency . . ."); see also 40 U.S.C. § 11331 (delegating to the Office of Management and Budget (OMB) and National Institute of Science & Technology (NIST) the authority to "promulgate information security standards pertaining to Federal information systems"); Office of Mgmt. & Budget, Exec. Office of the President, OMB Circular A-130, Appendix IV: Analysis of Key Sections (2016), [https://obamawhitehouse.archives.gov/omb/circulars\\_a130\\_a130appendix\\_iv](https://obamawhitehouse.archives.gov/omb/circulars_a130_a130appendix_iv) ("Each agency program official must understand the risk to [information] systems under their control.").

<sup>89</sup> See Nat'l Inst. of Standards and Tech., U.S. Dept. of Comm., NIST Guide for Conducting Risk Assessments, Special Publication 800-30 (2012), <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf> (providing comprehensive risk assessment guidance for federal information systems).

<sup>90</sup> It bears noting that risk management is considered best practice when private enterprises acquire AI technologies. For a private sector framework, see for example: Deloitte Ctr. for Regulatory Strategy, AI and Risk Management Innovating with Confidence (2018), <https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovate-lu-ai-and-risk-management.pdf>.

<sup>91</sup> See 48 C.F.R. § 7.103(p)(3); see also U.S. Envtl. Prot. Agency, *Guiding Principles for Sustainable Federal Buildings*, EPA (2016), <https://www.epa.gov/greeningepa/guiding-principles-federal-leadership-high-performance-and-sustainable-buildings> (last visited Aug. 12, 2020).

should be comprised of individuals with diverse backgrounds and perspectives, which can help to address potential blind spots relating to bias and other social concerns.

Whether these multidisciplinary teams should be centralized for government-wide deployment or decentralized within agencies is an important question. There are advantages and disadvantages to both approaches.<sup>92</sup> This report does not directly take up this question, except to note that the government's options may be limited by circumstance. Under existing conditions, the government's in-house capacity challenges may necessitate centralizing expertise to provide consulting, advice, and support for AI acquisitions on a government-wide basis. Currently, the General Services Administration (e.g., "18F" and Centers of Excellence)<sup>93</sup> and the Office of Management and Budget (e.g., US Digital Service) are serving these roles.<sup>94</sup>

Substantively, an AI risk assessment should address the issues outlined below and, to promote accountability, should be signed by the agency official overseeing the acquisition.

## **1. Will the AI tool impact individuals, businesses, and communities in high-risk or sensitive contexts?**

If yes, then thorough consideration must be given to issues of transparency, accountability, fairness, performance, privacy, and safety (which are separately addressed below).

## **2. To what extent, if any, will the agency need to rely on vendors and outside consultants to design, develop, implement, audit, and monitor the system?**

Acquiring AI systems is not a one-time decision; designing the system and testing it over its lifecycle are necessary to ensure that the tool is performing as intended, accurately, and within legal bounds. If the government's reliance on vendors and consultants is more than nominal, then considerations about cost, security, and control of the system should be explicitly accounted for in the risk assessment. Moreover, depending on the degree of the government's anticipated reliance on third parties, this risk factor could itself be a reason to forego acquiring an AI

---

<sup>92</sup> The advantages of centralization include coordination, concentrated expertise, and the development and deployment of best practices on a government-wide basis. See Mulligan & Bamberger, *supra* note 27, at 830–32 (favoring a centralized approach). The disadvantages of centralization include cultural and informational gaps between a centralized group of AI experts and the agency personnel who will, post-procurement, need to use, understand, and trust the technology. Cf. Aaron Boyd, *SBA Spent \$30M on a Digital Service-Built App that Doesn't Work*, NEXTGOV (Aug. 3, 2020) (reporting that "confusion over leadership and culture clashes led to a runaway project with little oversight").

<sup>93</sup> See Gen. Servs. Admin., Technology Transformation Services, <https://www.gsa.gov/about-us/organization/federal-acquisition-service/technology-transformation-services> (last visited Sept. 6, 2020).

<sup>94</sup> See *An inside look at USDS*, U.S. Digital Service, <https://www.usds.gov/news-and-blog> (last visited Sept. 6, 2020).

solution.<sup>95</sup> If the acquiring agency cannot assemble and maintain an appropriately dedicated multidisciplinary team to oversee and manage the acquisition, that risk will exacerbate all the others, and should probably lead to a no-go decision on acquiring an AI system for use in high-risk or sensitive contexts.

### **3. Will the data used or generated by the AI system contain sensitive personal information?**

If yes, then additional questions relating to data privacy, data integrity, and data security should be considered and addressed.

### **4. Will there be a human-in-the-loop, and if so, at what point in the operational workflow?**

If the AI system is used in high-risk or sensitive contexts, then human validation of the inputs and outputs will likely be necessary. Moreover, depending on context, a human-in-the-loop may be required to exercise discretionary judgment, in which case, risks about computer-human interactions must also be accounted for. Those risks include biased *disregard* of an algorithmic prediction, which may occur if the human is predisposed to contrary outcomes, or does not understand (or trust) how the system operates. Risks at the human-computer interface, however, also include a system operator's *overconfidence* in AI predictions, which may result from a lack of appreciation for how the system operates, resource constraints, or the accretion of domain expertise and human judgment over time.

### **5. How transparent is the AI system?**

As earlier explained, AI systems can be more or less transparent for a variety of reasons. Thus, agencies should not treat transparency as a monolithic concern, but rather as a compendium of system features that should separately be accounted for in the AI risk assessment. Below are some common transparency concerns that should be specifically addressed:

- **"Interpretability" (and "Explainability"):**<sup>96</sup> Numerous efforts are underway to make complex algorithmic models more interpretable and explainable to humans.<sup>97</sup> As yet, there are no industry or government standards—in part

---

<sup>95</sup> Cf. NSCAI, Second Quarter Recommendations, *supra* note 12, at 89 ("Despite pockets of excellence, the government lacks wide expertise to envision the promise and implications of AI, translate vision into action, and develop the operating concepts for using AI.").

<sup>96</sup> See Gabriel Nicholas, 4 GEO. L. TECH. REV. 711, 715 (2020) (noting disagreement in the literature over the nomenclature around "interpretability" and "explainability"). For present purposes, I use the terms interchangeably to refer to the ability of humans to understand how an AI system generates outputs from inputs.

<sup>97</sup> See *supra* note 40.

because algorithmic explainability carries its own set of risks. The more the model's decisional pathway is reduced for human comprehension (i.e., dumbed down), the more the explanation will depart from the truth of how a decision was actually made. Moreover, the explaining algorithms are vulnerable to gaming and adversarial attack. Consequently, computer-generated explanations can be unintentionally or intentionally misleading,<sup>98</sup> creating risk in both directions.

- **Data:** Because an AI model learns from data, access to information about training data (or lack thereof) is critical for evaluating an AI system's intended purposes and functionality. Not all data is created equal; there are gradients of data integrity, data bias, and associated risks. Moreover, vendors may claim trade secrecy protection over datasets or the process by which data was assembled and used to train the model. The lack of transparency around data can be highly risky.
- **Model Versioning (e.g., Version 1.0, 1.5, 2.0):** Throughout the lifecycle of an AI system, algorithmic models may be updated to account for new data, new workflows, new technology, and so on. An AI system that is modified may improve performance, but may cause the system to perform *differently* than anticipated or originally conceived. Moreover, without proper precautions, model versioning can make it impossible to know how a model performed when a particular decision was made. If that knowledge is needed in a court setting, but is unavailable, an agency will be hard-pressed to legally justify the specific output that is the subject of litigation.

## b. Dividends of AI Risk Assessment

Apart from good practice, pre-acquisition AI risk assessments can serve a variety of useful functions across the lifecycle of an AI system. Most obviously, the deliberation and documentation of known and foreseeable risks will force conversations about whether an AI solution is appropriate at all, and if so, under what conditions. Along similar lines, the risk assessment can inform decisions about data needs, whether data is available on the market, and costs relating to data curation, labeling, scrubbing, enrichment, etc.

Further dividends from risk assessments can accrue during the market solicitation phase. More specifically, contracting officials can use the catalogued risks as focal points for market competition in connection with the agency's requests for information (RFIs),<sup>99</sup>

---

<sup>98</sup> See Dylan Slack et al., *Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods*, In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), <https://doi.org/10.1145/3375627.3375830>.

<sup>99</sup> See 48 C.F.R. § 15.201(e) ("RFIs may be used when the Government does not presently intend to award a contract, but wants to obtain price, delivery, other market information, or capabilities for planning purposes. Responses to these notices are not offers and cannot be accepted by the Government to form a binding contract.").

requests for proposals (RFPs),<sup>100</sup> and requests for quotations (RFQs).<sup>101</sup> By way of illustration, contracting officials can solicit information and proposals from vendors along the following lines:

1. Describe your privacy and cybersecurity approach to the proposed AI system, including but not limited to how the data will be protected.
2. Explain your testing and validation processes for performance, fairness, and discrimination, including any special expertise or innovative approaches that you might use to evaluate the AI system throughout its lifecycle.
3. Describe any anticipated data limitations and challenges, and any strategies or solutions that you might implement to address them.
4. Describe any training programs that your team members have undergone, and any official policies or protocols adopted by your organization, that specifically relate to transparency, accountability, fairness, privacy, or other ethical AI principles.
5. Describe how you enable end-to-end auditability of the system by government personnel, and any technical limitations for such auditing. In this regard, would you permit independent third-party audits of the AI system? If yes, explain the conditions or limitations you would impose on independent auditors. If you would not permit third-party audits, then explain why.
6. Describe whether, and to what extent, the AI system outputs will be explainable and interpretable, and by what means, to (i) third-party computer engineers; (ii) agency personnel trained to use the system; and (iii) other stakeholders, including laypersons, judges, and policymakers.
7. Describe any known or foreseeable (i) biases in the AI system; (ii) performance weaknesses; and (iii) vulnerabilities. Further, describe the known and foreseeable sources or causes of those biases, performance weakness, and vulnerabilities (e.g., in the data, algorithm, design process, human-computer interactions, interoperability with other hardware and software, or otherwise).
8. Explain how you will ensure or test that the AI system does not drift from its intended purposes or outcomes.
9. Explain how you will ensure or facilitate usability for government personnel (e.g., training programs, written materials, access to source code, and otherwise).

---

<sup>100</sup> See *id.* § 15.203(a) (“[RFPs] are used in negotiated acquisitions to communicate Government requirements to prospective contractors and to solicit proposals.”).

<sup>101</sup> See *id.* § 8.402 (RFQs are used when agencies order goods and services from federal supply schedules).

Incorporating AI risk-related questions in market solicitations, along the lines above, can yield several direct and indirect benefits. Most directly, the answers by market participants will enable the agency to compare the types and degrees of risk associated with a *particular vendor* relative to the field. Anticipating this, strategic and innovative vendors will compete for an ethical edge. In some instances, the agency might even find opportunities for collaboration—for example, between two or more start-up enterprises—to mitigate the overall risk based on their respective strengths and weaknesses. Further, vendor responses to the questions above may shed light on additional risks not previously identified. In such cases, the AI risk assessment can and should be modified accordingly.

## c. Source Selection and Contractual Award

Whether independently or in conjunction with the foregoing recommendations, federal procurement law should require that agencies pay due regard for AI principles in source selection and contractual award. For purposes of this discussion, I will mostly bracket the issue of cybersecurity, which is already receiving sustained attention by policymakers. But it is worth pausing to note some recent cybersecurity regulatory interventions because they point to the need, and feasibility, of acquiring ethical AI tools through the procurement process.

Federal law, for example, now prohibits executive agencies from contracting with entities that use any equipment, system, or service that utilizes covered telecommunications equipment from blacklisted companies (such as Huawei and ZTE Corporation).<sup>102</sup> Moreover, the Department of Defense (DoD) will be phasing in a new Cybersecurity Maturity Model Certification (CMMC) program that will require DoD vendors to have adequate controls in place to protect sensitive information and data.<sup>103</sup> Under this program, vendors will be required to obtain CMMC certification from third-party auditors; vendors that fall short of prescribed security levels cannot even compete for those contracts.<sup>104</sup>

---

<sup>102</sup> John S. McCain National Defense Authorization Act for Fiscal Year 2019, Pub. L. No. 115-232, § 889(a)(1), 132 Stat. 1636, 1917 (2018); see also Federal Acquisition Regulation: Prohibition on Contracting with Entities Using Certain Telecommunications and Video Surveillance Services or Equipment, 85 Fed. Reg. 42,665 (July 14, 2020).

<sup>103</sup> Off. of the Under Sec'y of Def. for Acquisition & Sustainment Cybersecurity Maturity Model Certification, CMMC FAQ's, <https://www.acq.osd.mil/cmmc/faq.html>.

<sup>104</sup> See Peter Baldwin & Jason G. Weiss, *DoD's Cybersecurity Maturity Model Certification Is Here: What Your Business Needs to Do to Prepare*, NAT'L L. REV. (Apr. 15, 2020), <https://www.natlawreview.com/article/dod-s-cybersecurity-maturity-model-certification-here-what-your-business-needs-to-do>.

A certification system for ethical AI is within the realm of future possibility but is premature at this time.<sup>105</sup> Standardizing transparency and fairness metrics, for example, could have the unintended consequence of entrenching market incumbents and stifling innovation. Instead, the recommendation here is to require agencies to incorporate ethical AI considerations into the source selection process, and/or to fold such considerations into the prerequisites for contractual award. The discussion below elaborates on these alternatives.

## 1. Evaluative Criteria

By way of background, the express and overarching purpose of the Federal Acquisition Regulation (FAR) is to “deliver . . . the best value product or service to the customer, while maintaining the public’s trust and fulfilling public policy objectives.”<sup>106</sup> To those ends, FAR instructs contracting officials to examine the strengths and weaknesses of all relevant factors (such as cost, performance, quality, and schedule) and to make tradeoffs between cost and non-cost factors.<sup>107</sup> FAR instructs that “cost or price may play a dominant role in source selection” in acquisitions where the requirements for performance are “clearly definable” and the “risk of unsuccessful contract performance is minimal.”<sup>108</sup> The same provision, however, also explains that non-price considerations—such as “technical or past performance”—may “play a dominant role in source selection” when the performance requirements are less specified and the risk of unsuccessful performance is greater.<sup>109</sup>

Contracting officials have wide discretion in how to achieve “best value” under these standards. Yet FAR does erect some important guardrails. Pertinent here, contracting officials must pre-disclose how they intend to evaluate cost and non-cost criteria, and then must adhere to those criteria when awarding contracts.<sup>110</sup> Currently, nothing prevents contracting officials from incorporating ethical AI considerations into the source selection

---

<sup>105</sup> Cf. Institute of Electrical and Electronics Engineers (IEEE) P7000 (proposing a set of standards “for addressing ethical concerns during [AI] system design”), <https://standards.ieee.org/project/7000.html>.

<sup>106</sup> 48 C.F.R. § 1.102(a).

<sup>107</sup> *Id.* § 15.101-1.

<sup>108</sup> *Id.* § 15.101.

<sup>109</sup> *Id.*

<sup>110</sup> See *id.* § 15.305(a) (“An agency shall evaluate competitive proposals and then assess their relative qualities solely on the factors and subfactors specified in the solicitation.”); see also *Antarctic Support Assocs. v. United States*, 46 Fed. Cl. 145, 155 (2000) (noting that reviewing courts must afford contracting officials great deference; contractual awards must only be reasonable and consistent with stated evaluation criteria).

process.<sup>111</sup> But doing so can and should be required as a matter of law. Issues relating to transparency, accountability, privacy, security, and fairness will always be relevant in the procurement of AI systems. How relevant will be context-specific. Thus, considerations of ethical AI principles can be weighed as appropriate depending on the circumstances.

**Nothing prevents contracting officials from incorporating ethical AI considerations into the source selection process. But doing so can and should be required as a matter of law.**

One virtue of this approach is its flexibility. Little or nothing would be sacrificed by requiring contracting officials to consider ethical AI principles in conjunction with price, technical feasibility, and other criteria. Meanwhile, much could be gained from explicitly making ethical AI part of the calculus. Most importantly, doing so can mitigate the special risks that inhere in AI acquisitions. But it will also signal to agency officials, industry players, and stakeholders that AI procurement is not business as usual: The social and governance implications extend beyond technical criteria and model performance. Challenges around transparency, accountability, privacy, security, and fairness may (or may not) be addressed with technical patches.<sup>112</sup> Still, it is a categorical error to conceive of the challenges of algorithmic governance as mere technical problems that can be fixed with better data, mathematical proofs, or more software to “explain” the computations and configurations of black-box AI systems.<sup>113</sup>

**AI procurement is not business as usual: The social and governance implications extend beyond technical criteria and model performance.**

---

<sup>111</sup> Nonprofit organizations have made similar recommendations as a matter of best practice, but not as a matter of law. See, e.g., Am. Council for Tech.-Indus. Advisory Council, AI Playbook for the U.S. Federal Government, at 15, 22, 29, 35 (2020); World Econ. Forum, AI Procurement in a Box: AI Government Procurement Guidelines (June 11, 2020), <https://www.weforum.org/reports/ai-procurement-in-a-box/ai-government-procurement-guidelines#report-nav>.

<sup>112</sup> See, e.g., Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J. L. & TECH. 1 (2017) (providing a computer scientist’s perspective on algorithmic accountability and calling for specific, tailored solutions).

<sup>113</sup> See Coal. for Critical Tech., *supra* note 74, at 2 (“To date, many efforts to deal with the ethical stakes of algorithmic systems have centered mathematical definitions of fairness that are grounded in narrow notions of bias and accuracy. These efforts give the appearance of rigor, while distracting from more fundamental epistemic problems.”); Brent Mittelstadt, Chris Russell, Sandra Wachter, *Explaining Explanations in AI*, in *Proceedings of Fairness, Accountability, and Transparency* (2019) (“[T]he bulk of methods currently occupying AI researchers lies in the . . . building of approximate models that are not intended to capture the full behaviour of physical systems but rather to provide coarse approximations of how the systems behave.”), <https://arxiv.org/pdf/1811.01439.pdf>.

## 2. Vendor Responsibility

For much the same reasons, an additional (or alternative) requirement should be incorporated into a contracting officer's "responsibility" determination. Currently, FAR directs contracting officials to determine whether a vendor satisfies a set of responsibility standards prior to awarding any procurement contract.<sup>114</sup> To meet the standards, a potential awardee must demonstrate, among other things, that it has "adequate financial resources;" a "satisfactory past performance record;" "the necessary organization, experience, facilities," and "technical skills" to perform the contract; and a satisfactory "record of integrity" and "business ethics."<sup>115</sup> These performance standards are designed, in part, to mitigate the risk of vendor misfeasance and contractual nonperformance.<sup>116</sup>

Beyond these performance standards, a responsibility determination also requires that a potential awardee be "otherwise qualified and eligible."<sup>117</sup> These criteria, in turn, encompass a set of collateral requirements. For example, potential awardees must be disqualified if they do not comply with federal equal employment opportunity requirements or if they fail to agree to an acceptable plan for subcontracting with small businesses. Unlike performance standards, which assess whether prospective vendors can be expected to fulfill the contract in a timely and satisfactory manner, collateral requirements ensure that the government's dealings with contractors promote federal socioeconomic goals.

Ethical AI principles could fold neatly into FAR's responsibility framework, either as a performance standard or collateral requirement. As a performance standard, ethical AI principles can be factored into the criteria for "necessary organization, experience, facilities" and "technical skills" to perform the contract, as well as a prospective vendor's "record of integrity and business ethics."<sup>118</sup> As a collateral requirement, and perhaps to greater effect, the imperative of ethical AI could be leveraged to gain important concessions from vendors.

---

<sup>114</sup> 48 C.F.R. § 9.103(a) ("Purchases shall be made from, and contracts shall be awarded to, responsible prospective contractors only"); *id.* § 9.103(b) ("No purchase or award shall be made unless the contracting officer makes an affirmative determination of responsibility.").

<sup>115</sup> 48 C.F.R. § 9.104-1; see also Kate M. Manuel, *Responsibility Determinations Under the Federal Acquisition Regulation: Legal Standards and Procedures*, Congressional Research Service, No. R40633 5–13 (Jan. 4, 2013) (providing explanations of FAR's responsibility standards and processes); Orca Northwest Real Est. Servs. v. U.S., 65 Fed. Cl. 1, 6 (2005), on reconsideration, 65 Fed. Cl. 419 (2005) (discussing criteria for responsibility).

<sup>116</sup> 48 C.F.R. § 9.103; Ryan Co. v. U.S., 43 Fed. Cl. 646, 651 (1999).

<sup>117</sup> 48 C.F.R. § 9.104-1(g).

<sup>118</sup> *Id.* § 9.104-1; see also Manuel, *supra* note 115, at 5–13; Orca Northwest Real Est. Servs., 65 Fed. Cl. at 6.

For example:

1. Prime contractors could be required to subcontract with “nontraditional” government vendors (to create opportunities for innovative small businesses and start-ups that might otherwise be reluctant to enter the government market).<sup>119</sup> Along similar lines, prime contractors could be required to subcontract with women-owned and “socially and economically disadvantaged” small-business enterprises.<sup>120</sup> Doing so could promote diversity in AI system design, development, and testing.<sup>121</sup>
2. Vendors could be required to share trade secrets with the government for specified purposes and under certain constraints. Intellectual property (IP) rights can be a major sticking point in government contracting, especially in high-tech fields like AI.<sup>122</sup> Indeed, trade secrets are often the most valuable assets for many small and start-up enterprises.<sup>123</sup> Thus, a responsibility requirement tethered to IP rights should be narrowly tailored. It should call for no more than is *foreseeably necessary* and should not substitute for other contractually negotiated IP terms. To be sure, there is always a risk that requiring IP concessions will scare off market participants.<sup>124</sup> But just as surely, the government should not be permitted to spend tax dollars on AI technologies that the government cannot fully use, audit, or explain. The minimally necessary criteria for purposes of a responsibility determination will be context-specific. For example, potential awardees should not qualify as responsible if they are not willing to waive trade secrets in adjudicatory contexts, where the invocation of trade secrecy would effectively preclude the government from utilizing an AI tool, and thus negate the purpose of acquiring the tool in the first place. If a vendor will not agree, then the government should be required to select a competitor that will. And, if none will agree, then the government should rethink whether a privately developed AI solution is appropriate for the government task.

---

<sup>119</sup> Cf. Steve Kelman, *Non-Traditional IT Contractors Unite to Watch Each Other’s Backs*, FCW, Lectern Blog (Feb. 12, 2018) (discussing how nontraditional IT vendors in the federal marketplace deliver innovative services), <https://fcw.com/blogs/lectern/2018/02/kelman-digital-services-coalition.aspx>.

<sup>120</sup> Federal law has an established program to promote contracting with any “small business which is unconditionally owned and controlled by one or more socially and economically disadvantaged individuals who are of good character and citizens of and residing in the United States, and which demonstrates potential for success.” 13 C.F.R. § 124.101; 15 U.S.C. § 637 (defining “socially and economically disadvantaged” individuals under the Small Business Act as “[s]ocially disadvantaged individuals are those who have been subjected to racial or ethnic prejudice or cultural bias because of their identity as a member of a group without regard to their individual qualities”); see also 48 C.F.R. § 19.15 (Woman-Owned Small Business Program).

<sup>121</sup> Cf. Bi-Partisan Center, *AI and the Workforce*, at 2 (Jul. 2020) (reporting on the nation’s AI talent gap and stating that “AI talent should ideally have a multi-disciplinary” skill set that includes ethics).

<sup>122</sup> See Nancy O. Dix et al., *Fear and Loathing of Federal Contracting: Are Commercial Companies Really Afraid to Do Business with the Federal Government? Should They Be?* 33 PUB. CONT. L.J. 5, 8–9 (2003) (providing a review of the relevant contracting requirements and industry concerns around IP provisions in government contracts that depart from general commercial terms).

<sup>123</sup> See generally Rob Kitchin, *Thinking Critically About and Researching Algorithms*, 20 INFO. COMM. & SOC’Y 14, 20 (2016) (“[I]t is often a company’s algorithms that provide it with a competitive advantage and they are reluctant to expose their intellectual property even with non-disclosure agreements in place.”).

<sup>124</sup> See Dix et al, *supra* note 122, at 9.

**3.** Vendors could be required to allow independent third-party auditing of an AI system—if for no other reason, because the government may not have the expertise or resources to do so itself. Given the many ways that AI systems can cause harm, a vendor that will not permit third-party auditing of its AI system should be disqualified from being awarded a government procurement contract. Of course, vendors will need adequate assurances that their trade secrets will not be compromised by third-party auditing. But there are well-established industry practices (for example, nondisclosure agreements with liability provisions) and federal laws<sup>125</sup> that can be used to safeguard vendors against trade secrecy misappropriation.

---

<sup>125</sup> See, e.g., 18 U.S.C. § 1836 (allowing the owner of a trade secret to sue in federal court when its trade secrets have been misappropriated); see also 18 U.S.C. § 1905 (imposing criminal penalties for any government employee who discloses information that “concerns or relates to trade secrets”).

## V. Objections

Anticipated objections will come from those who think the forgoing recommendations go too far, and from those who think the recommendations fall short.

For those concerned about overreach, it is likely for one of two reasons. First, it may be objected that federal acquisition procedures are already too cumbersome, especially for technology that can become outdated before they are put to use.<sup>126</sup> The recommendations here are sensitive to this concern. Requiring the government to create an AI inventory report is retrospective-facing and thus should create little or no friction in prospective government acquisitions. Moreover, risk assessments are already undertaken for IT systems in the normal course of procurement planning. The recommendation here would simply formalize and standardize that practice to address the *special risks* associated with AI systems. For example, risks relating to bias, model explainability, and data integrity are not the types of risks normally accounted for in technology acquisitions but can make or break the success of AI projects.

A second objection—and a major one—is that raising the bar on vendor responsibility and contractual award will discourage or disqualify innovative vendors from working with the government.<sup>127</sup> Although the industry is generally wary of more procurement regulations, the prescriptions advanced here seize upon areas of shared interest. For the government and industry alike, AI innovation is a complex ambition that scopes well beyond technological capability. Innovation also entails the responsible development and deployment of AI tools. Currently, every major technology company has teams of high-skilled workers and mounds of investment capital dedicated to ethical AI. And the government, for its part, is pouring huge amounts of tax dollars into related research and development.

Despite motivational differences, public and private interests around trustworthy AI merge in the acquisition gateway. That shared reality is a foundation for principled and pragmatic regulatory compromise, which this report aims to advance. Indeed, the

---

<sup>126</sup> See, e.g., Katherine M. John, *Information Technology Procurement in the United States and Canada: Reflecting on the Past With an Eye Toward the Future*, 48 PROCUREMENT LAWYER 4, 5 (2013) ("If procurement regimes overemphasize transparency and competition—or otherwise take too long—then end users might end up saddled with technology that is outdated by the time it reaches them.").

<sup>127</sup> Cf. L. Elaine Halchin, Other Transaction Authority, Cong. Res. Serv., RL34760 (Jul. 15, 2011) ("The Government is finding that not only can it not acquire many of the technologies it needs, but also many corporations will not even accept government dollars to help develop new technologies.").

government's demand for ethically designed AI systems may attract innovative talent to the government market, given that many technology companies—big and small—have expressed serious concerns about the government's ethical use of this technology (especially as relates to law enforcement and warfighting).<sup>128</sup>

## **Public and private interests around trustworthy AI merge in the acquisition gateway.**

It remains to be seen whether the government can swing these market dynamics more in its favor. But it is easy to see how these dynamics can become measurably worse. Without intervention, the government will only grow more dependent on private industry to provide the tools of algorithmic government. In a world where industry is more responsible and ethical than the government, perhaps that is a good thing. In the meantime, however, the widely shared hope is that the government and industry will be partners in the journey ahead. Suffusing the procurement process with ethical AI is not the only path. But it's the one America should insist upon.

**Without intervention, the government will only grow more dependent on private industry to provide the tools of algorithmic government.**

Coming from the opposite direction, policymakers and stakeholders may think the recommendations in this report do not go far enough. Emphatically, I agree. In due course, more can and should be done to meet the significant challenges of algorithmic governance. Crafting concrete regulatory mechanisms that meaningfully fulfill the government's legal and moral obligations is an all-hands-on-deck effort. The

---

<sup>128</sup> See, e.g., Arvind Krishna, *IBM CEO's Letter to Congress on Racial Justice Reform*, IBM (June 8, 2020) (announcing that IBM "will not condone uses of any technology . . . for mass surveillance, racial profiling, violations of basic human rights and freedoms, or any purpose which is not consistent with our values and Principles of Trust and Transparency."), <https://www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms/>; Brad Smith, *Facial Recognition: It's Time for Action*, Microsoft on the Issues: The Official Microsoft Blog (Dec. 6, 2018) ("[W]e don't believe that the world will be best served by a commercial race to the bottom, with tech companies forced to choose between social responsibility and market success."), <https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/>; *We Are Implementing a One-Year Moratorium on Police Use of Recognition*, Day One: The Amazon Blog (June 10, 2020), <https://blog.aboutamazon.com/policy/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition> (taking their product off the market for a year to give lawmakers time to "put in place stronger regulations to govern the ethical use of facial recognition technology"); Rosalie Chan, *Google Drops Out of Contention for a \$10 Billion Defense Contract Because It Could Conflict with Its Corporate Values*, BUSINESS INSIDER (Oct. 8, 2018), <https://www.businessinsider.com/google-drops-out-of-10-billion-jedi-contract-bid-2018-10>.

recommendations offered here are part of that larger project. And, constructively, nothing suggested here would preclude additional regulatory interventions—whether in procurement law, administrative law, privacy law, or in domain-specific contexts.

Regarding procurement law, in particular, federal lawmakers should consider whether special acquisition pathways for AI are necessary and appropriate. Given the complexity of the AI development process, it may be impossible or ill-advised for the agency to specify AI performance metrics at the time of contracting. Moreover, many if not most of the important design choices may happen after an award is made. These characteristics of AI development make traditional “waterfall” acquisition pathways quite unsuitable, and risky, because agencies may get locked into contracts for dead-end solutions with no easy exit ramp.<sup>129</sup>

To meet these challenges, the Government Accountability Office (GAO) recently issued a best-practices guide touting the use of “agile” development pathways for federal software acquisition.<sup>130</sup> The Office of Management and Budget and General Services Administration have likewise championed the use of agile acquisition frameworks.<sup>131</sup> Unlike the front-loaded waterfall approach, agile methods are characterized by incremental and iterative processes in which software is produced in close collaboration with the acquiring agency. When properly planned and managed, agile boasts of improved investment manageability, lowers the risk of project failure, shortens the time to realize value, and allows agencies to better adapt to changing needs. Information obtained during these frequent iterations allow developers to respond quickly to feedback from agency customers, thus potentially reducing sociotechnical, legal, and programmatic risk.

Arguably, FAR contains enough flexibility to accommodate agile acquisition pathways. For example, pursuant to FAR, agencies may use modular contracting vehicles for “delivery, implementation, and testing of workable systems or solutions in discrete increments, each of which comprises a system or solution that is not dependent on any subsequent increment in order to perform its principal functions.”<sup>132</sup> Moreover,

---

<sup>129</sup> In a typical waterfall acquisition, the technical and design requirements are fixed at the time of contracting.

<sup>130</sup> GAO-20-590G, Agile Assessment Guide: Best Practices for Agile Adoption and Implementation (Sept. 2020) [hereafter, GAO-20-590G, Agile Best Practices], <https://www.gao.gov/products/GAO-20-590G>.

<sup>131</sup> See, e.g., Office of Management and Budget, Office of Federal Procurement Policy (OFPP), Contracting Guidance to Support Modular Development (Jun. 2012), <https://obamawhitehouse.archives.gov/sites/default/files/omb/procurement/guidance/modular-approaches-for-information-technology.pdf>; General Services Administration, De-Risking Government Technology, Federal Agency Field Guide (Sept. 2020), <https://derisking-guide.18f.gov/assets/federal-field-guide-a245c3a7dcd0a24f619b458fd51e1e490f2299023fd1bd13fddc87318e67cf03.pdf>.

<sup>132</sup> 48 CFR § 39.103.

modular contracts “provide an opportunity for subsequent increments to take advantage of any evolution in technology or needs that occur during implementation and use of the earlier increments,” and can “reduce risk of potential adverse consequences on the overall project by isolating and avoiding custom-designed components of the system.”<sup>133</sup>

Agile acquisition pathways, however, are no panacea for AI systems. These approaches require skill sets, resources, and institutional buy-in that many agencies are currently without. Indeed, GAO’s best-practices guide chronicles an array of challenges that agencies have experienced using agile processes. For example, “teams reported difficulty collaborating closely or transitioning to self-directed work due to constraints in organization commitment and collaboration.” Moreover, “some organizations reported that they did not have trust in iterative solutions and that teams had difficulty managing iterative requirements.”<sup>134</sup>

AI acquisitions would give rise to all the same challenges—and more—given the complexities, decisions, and ongoing monitoring that these systems require. Still, there is reason to hope the federal government will gain the experience and resources to responsibly manage the acquisition of ethically designed AI systems, through modular contracting or otherwise.

---

<sup>133</sup> *Id.*

<sup>134</sup> See GAO-20-590G, Agile Best Practices, *supra* note 130, at 14-16; see also GAO-16-467, Immigration Benefits System: US Citizenship and Immigration Services Can Improve Program Management (Jul. 2016) (reporting that the United States Citizenship and Immigration Service Transformation program was not setting outcomes for agile software development), <https://www.gao.gov/products/GAO-16-467>; GAO-18-46, TSA Modernization: Use of Sound Program Management and Oversight Practices is Needed to Avoid Repeating Past Problems (Oct. 2017) (reporting that the Transportation Security Administration’s Technology Infrastructure Modernization (TIM) program did not define key agile roles, prioritize system requirements, or implement automated capabilities), <https://www.gao.gov/products/GAO-18-46>.

## VI. Conclusion

It is encouraging that the United States has committed to “AI principles” and the protection of “civil liberties, privacy, and American values . . . in order to fully realize the potential of AI technologies for the American people.”<sup>135</sup> But proselytizing is not actualizing.<sup>136</sup> If federal policymakers are truly committed to ethical algorithmic governance, then the acquisition gateway is a prime place to start. This report has shined critical light on the need for procurement reform and has offered a set of recommendations that future work can build upon.

---

<sup>135</sup> See Office of Sci. and Tech. Policy, Exec. Office of the President, American Artificial Intelligence Initiative: Year One Annual Report (2020).

<sup>136</sup> Cf. NIST, Second Quarter Report, supra note 12, at 123 (acknowledging that “[t]here is often a gap between articulating high-level goals around responsible AI and operationalizing them”).

THE  
GREAT  
DEMOCRACY  
INITIATIVE



[Home](#)

Guidance

# Data ethics and AI guidance landscape

This page collates the various existing ethical principles for data and AI, developed by government and public sector bodies. It intends to provide clarity and guidance for public servants working with data and/or AI.

---

From: [Department for Science, Innovation and Technology \(/government/organisations/department-for-science-innovation-and-technology\)](#) and [Department for Digital, Culture, Media & Sport \(/government/organisations/department-for-digital-culture-media-sport\)](#)

Published 21 July 2020

## Product

---

[Data Ethics Framework \(<https://www.gov.uk/government/framework/data-ethics-framework>\)](#)

## Product

---

[A guide to using AI in the public sector](https://www.gov.uk/government/publications/a-guide-to-using-ai-in-the-public-sector) (<https://www.gov.uk/government/publications/a-guide-to-using-ai-in-the-public-sector>)

---

[AI procurement guidelines](https://www.gov.uk/government/policies/ai-procurement-guidelines) (<https://www.gov.uk/government/policies/ai-procurement-guidelines>)

---

## Product

---

---

[UKSA Self-assessment \(<https://uksa.statisticsauthority.gov.uk/authorities/committees/nsdec/data-ethics/self-assessment-2/>\)](https://uksa.statisticsauthority.gov.uk/authorities/committees/nsdec/data-ethics/self-assessment-2/)

## Product

---

**Code of conduct for data-driven health and care technology**  
(<https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology>)

---

**NHSX: A Buyer's Checklist for AI in Health and Care**  
([https://www.nhsx.nhs.uk/media/documents/A\\_Buyers\\_Checklist\\_for\\_AI\\_in\\_Health\\_and\\_Care.pdf](https://www.nhsx.nhs.uk/media/documents/A_Buyers_Checklist_for_AI_in_Health_and_Care.pdf))

## Product

---

[Intelligent security tools](https://www.ncsc.gov.uk/collectio) (<https://www.ncsc.gov.uk/collectio>

---

[The Dstl Biscuit Books](https://www.gov.uk/government/c) (<https://www.gov.uk/government/c>

## Product

---

---

[The Magneta Book](https://www.gov.uk/government/publications/the-magneta-book) (<https://www.gov.uk/government/publications/the-magneta-book>)

## Product

---

[Analytical Quality Assurance Book \(AQuA\)](https://www.aqua-book-guidance-on-producing-quality-analysis-for-gove) (<https://www.aqua-book-guidance-on-producing-quality-analysis-for-gove>)

---

[Service Standard - Service Manual](https://www.gov.uk/government/publications/service-standard-service-manual) (<https://www.gov.uk/government/publications/service-standard-service-manual>)

---

## Multilateral guidance

Product	Creator	Description
---------	---------	-------------

<a href="https://www.oecd.org-going-digital/ai/principles/">OECD principles on AI (https://www.oecd.org-going-digital/ai/principles/)</a>	OECD	The five value based OECD principles on / promote artificial intelligence (AI) that is innovative and trustworthy and that respects human rights democratic values. They were adopted in March 2019 by OECD member countries (including the
<a href="https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf">G20 AI principles (https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf)</a>	G20	The G20 AI principles draw from and agree with the OECD principles and recommendations

Published 21 July 2020



All content is available under the [Open Government Licence v3.0](#), except where otherwise stated

[© Crown copyright](#)

[Home](#)

Guidance

# Understanding artificial intelligence ethics and safety

Understand how to use artificial intelligence ethically and safely

---

From: [Department for Science, Innovation and Technology \(/government/organisations/department-for-science-innovation-and-technology\)](#), [Office for Artificial Intelligence \(/government/organisations/office-for-artificial-intelligence\)](#) and [Centre for Data Ethics and Innovation \(/government/organisations/centre-for-data-ethics-and-innovation\)](#)

Published 10 June 2019

## Contents

- [Who this guidance is for](#)
- [Understanding what AI ethics is](#)
- [Varying your governance for projects using AI](#)
- [Establish ethical building blocks for your AI project](#)
- [Start with a framework of ethical values](#)
- [Establish a set of actionable principles](#)

## Related content

[Managing your artificial intelligence project \(/guidance/managing-your-artificial-intelligence-project\)](#)

[A guide to using artificial intelligence in the public](#)

- Related guides

This guidance is part of a wider collection about [using artificial intelligence \(AI\) in the public sector](#) (<https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector>).

The Office for Artificial Intelligence (OAI) and the Government Digital Service (GDS) have produced the following chapter's guidance in partnership with The Alan Turing Institute's [public policy programme](#) (<https://www.turing.ac.uk/research/research-programmes/public-policy>). This chapter is a summary of [The Alan Turing Institute's detailed guidance](#) ([https://www.turing.ac.uk/sites/default/files/2019-06/understanding\\_artificial\\_intelligence\\_ethics\\_and\\_safety.pdf](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf)), and readers should refer to the full guidance when implementing these recommendations.

AI has the potential to make a substantial impact for individuals, communities, and society. To make sure the impact of your AI project is positive and does not unintentionally harm those affected by it, you and your team should make considerations of AI ethics and safety a high priority.

This section introduces AI ethics and provides a high-level overview of the ethical building blocks needed for the responsible delivery of an AI project.

The following guidance is designed to complement and supplement the [Data Ethics Framework](#) (<https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework>). The Data Ethics Framework is a tool that should be used in any project.

## sector

[\(/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector\)](#)

## Using natural language processing to structure market research

[\(/government/case-studies/using-natural-language-processing-to-structure-market-research\)](#)

## How DFID used satellite images to estimate populations

[\(/government/case-studies/how-dfid-used-satellite-images-to-estimate-populations\)](#)

## Using data from electricity meters to predict energy consumption

[\(/government/case-studies/using-data-from-electricity-meters-to-predict-energy-consumption\)](#)

## Collection

## A guide to using artificial intelligence in the public sector

[\(/government/collections/a-guide-to-using-artificial-](#)

# Who this guidance is for

intelligence-in-the-public-sector)

This guidance is for everyone involved in the design, production, and deployment of an AI project such as:

- data scientists
- data engineers
- domain experts
- delivery managers
- departmental leads

Ethical considerations will arise at every stage of your AI project. You should use the expertise and active cooperation of all your team members to address them.

## Understanding what AI ethics is

AI ethics is a set of values, principles, and techniques that employ widely accepted standards to guide moral conduct in the development and use of AI systems.

The field of AI ethics emerged from the need to address the individual and societal harms AI systems might cause. These harms rarely arise as a result of a deliberate choice - most AI developers do not want to build biased or discriminatory applications or applications which invade users' privacy.

The main ways AI systems can cause involuntary harm are:

- misuse - systems are used for purposes other than those for which they were designed and intended
- questionable design - creators have not thoroughly considered technical issues related to algorithmic bias and safety risks
- unintended negative consequences - creators have not thoroughly considered the potential negative impacts their systems may have on the individuals and communities they affect

The field of AI ethics mitigates these harms by providing project teams with the values, principles, and techniques needed to produce ethical, fair, and safe AI applications.

## Varying your governance for projects using AI

The guidance summarised in this chapter and presented at length in [The Alan Turing Institute's further guidance on AI ethics and safety](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf) ([https://www.turing.ac.uk/sites/default/files/2019-06/understanding\\_artificial\\_intelligence\\_ethics\\_and\\_safety.pdf](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf)) is as comprehensive as possible. However, not all issues discussed will apply equally to each project using AI.

An AI model which filters out spam emails, for example, will present fewer ethical challenges than one which identifies vulnerable children. You and your team should formulate governance procedures and protocols for each project using AI, following a careful evaluation of social and ethical impacts.

## Establish ethical building blocks for your AI project

You should establish ethical building blocks for the responsible delivery of your AI project. This involves building a culture of responsible innovation as well as a governance architecture to bring the values and principles of ethical, fair, and safe AI to life.

### Building a culture of responsible innovation

To build and maintain a culture of responsibility you and your team should prioritise 4 goals as you design, develop, and deploy your AI project. In particular, you should make sure your AI project is:

- ethically permissible - consider the impacts it may have on the wellbeing of affected stakeholders and communities

- fair and non-discriminatory - consider its potential to have discriminatory effects on individuals and social groups, mitigate biases which may influence your model's outcome, and be aware of fairness issues throughout the design and implementation lifecycle
- worthy of public trust - guarantee as much as possible the safety, accuracy, reliability, security, and robustness of its product
- justifiable - prioritise the transparency of how you design and implement your model, and the justification and interpretability of its decisions and behaviours

Prioritising these goals will help build a culture of responsible innovation. To make sure they are fully incorporated into your project you should establish a governance architecture consisting of a:

- framework of ethical values
- set of actionable principles
- process based governance framework

## Start with a framework of ethical values

You should understand the framework of ethical values which support, underwrite, and motivate the responsible design and use of AI. The Alan Turing Institute calls these 'the SUM Values':

- respect the dignity of individuals
- connect with each other sincerely, openly, and inclusively
- care for the wellbeing of all
- protect the priorities of social values, justice, and public interest

These values:

- provide you with an accessible framework to enable you and your team members to explore and discuss the ethical aspects of AI

- establish well-defined criteria which allow you and your team to evaluate the ethical permissibility of your AI project

You can read further guidance on SUM Values in [The Alan Turing Institute's comprehensive guidance on AI ethics and safety](#) ([https://www.turing.ac.uk/sites/default/files/2019-06/understanding\\_artificial\\_intelligence\\_ethics\\_and\\_safety.pdf](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf)).

## Establish a set of actionable principles

While the SUM Values can help you consider the ethical permissibility of your AI project, they are not specifically catered to the particularities of designing, developing, and implementing an AI system.

AI systems increasingly perform tasks previously done by humans. For example, AI systems can screen CVs as part of a recruitment process. However, unlike human recruiters, you cannot hold an AI system directly responsible or accountable for denying applicants a job.

This lack of accountability of the AI system itself creates a need for a set of actionable principles tailored to the design and use of AI systems. The Alan Turing Institute calls these the 'FAST Track Principles':

- fairness
- accountability
- sustainability
- transparency

Carefully reviewing the FAST Track Principles helps you:

- ensure your project is fair and prevent bias or discrimination
- safeguard public trust in your project's capacity to deliver safe and reliable AI

## Fairness

If your AI system processes social or demographic data, you should design it to meet a minimum level of discriminatory non-harm. To do this you should:

- use only fair and equitable datasets (data fairness)
- include reasonable features, processes, and analytical structures in your model architecture (design fairness)
- prevent the system from having any discriminatory impact (outcome fairness)
- implement the system in an unbiased way (implementation fairness)

## Accountability

You should design your AI system to be fully answerable and auditable. To do this you should:

- establish a continuous chain of responsibility for all roles involved in the design and implementation lifecycle of the project
- implement activity monitoring to allow for oversight and review throughout the entire project

## Sustainability

The technical sustainability of these systems ultimately depends on their safety, including their accuracy, reliability, security, and robustness.

You should make sure designers and users remain aware of:

- the transformative effects AI systems can have on individuals and society
- your AI system's real-world impact

## Transparency

Designers and implementers of AI systems should be able to:

- explain to affected stakeholders how and why a model performed the way it did in a specific context
- justify the ethical permissibility, the discriminatory non-harm, and the public trustworthiness of its outcome and of the processes behind its design and use

To assess these criteria in depth, [you should consult The Alan Turing Institute's guidance on AI ethics and safety](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf) ([https://www.turing.ac.uk/sites/default/files/2019-06/understanding\\_artificial\\_intelligence\\_ethics\\_and\\_safety.pdf](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf)).

## Build a process-based governance framework

The final method to make sure you use AI ethically, fairly, and safely is building a process-based governance framework. The Alan Turing Institute calls it a 'PBG Framework'. Its primary purpose is to integrate the SUM Values and the FAST Track Principles across the implementation of AI within a service.

Building a good PBG Framework for your AI project will provide your team with an overview of:

- the relevant team members and roles involved in each governance action
- the relevant stages of the workflow in which intervention and targeted consideration are necessary to meet governance goals
- explicit timeframes for any evaluations, follow-up actions, re-assessments, and continuous monitoring
- clear and well-defined protocols for logging activity and for implementing mechanisms to support end-to-end auditability

You may find it useful to [consider further guidance on allocating responsibility and governance for AI projects](https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution#allocating-responsibility-and-governance-for-ai-projects) (<https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution#allocating-responsibility-and-governance-for-ai-projects>).

## Related guides

- [Understanding artificial intelligence](https://www.gov.uk/government/publications/understanding-artificial-intelligence/a-guide-to-using-artificial-intelligence-in-the-public-sector)  
(<https://www.gov.uk/government/publications/understanding-artificial-intelligence/a-guide-to-using-artificial-intelligence-in-the-public-sector>)
- [Assessing if artificial intelligence is the right solution](https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution) (<https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution>)
- [Planning and preparing for artificial intelligence Implementation](https://www.gov.uk/guidance/planning-and-preparing-for-artificial-intelligence-implementation) (<https://www.gov.uk/guidance/planning-and-preparing-for-artificial-intelligence-implementation>)
- [Managing your AI project](https://www.gov.uk/guidance/managing-your-artificial-intelligence-project)  
(<https://www.gov.uk/guidance/managing-your-artificial-intelligence-project>)
- [Examples of real-world artificial intelligence use](https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector#examples-of-artificial-intelligence-use)  
(<https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector#examples-of-artificial-intelligence-use>)
- National Cyber Security Centre [guidance for assessing intelligent tools for cyber security](https://www.ncsc.gov.uk/collection/intelligent-security-tools)  
(<https://www.ncsc.gov.uk/collection/intelligent-security-tools>)
- The [Data Ethics Framework](https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework)  
(<https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework>)
- The [Technology Code of Practice](https://www.gov.uk/government/publications/technology-code-of-practice/technology-code-of-practice)  
(<https://www.gov.uk/government/publications/technology-code-of-practice/technology-code-of-practice>)

Published 10 June 2019

[↑ Contents](#)



All content is available under the [Open Government Licence v3.0](#), except where otherwise stated

© Crown copyright



[Home](#) > [Government](#) > [Government efficiency, transparency and accountability](#)

## Guidance

# Ethics, Transparency and Accountability Framework for Automated Decision-Making

Guidance for public sector organisations on how to use automated or algorithmic decision-making systems in a safe, sustainable and ethical way.

---

From: [Department for Science, Innovation and Technology \(/government/organisations/department-for-science-innovation-and-technology\)](#), [Centre for Data Ethics and Innovation \(/government/organisations/centre-for-data-ethics-and-innovation\)](#), [Cabinet Office \(/government/organisations/cabinet-office\)](#) and [Office for Artificial Intelligence \(/government/organisations/office-for-artificial-intelligence\)](#)

Published 13 May 2021

Last updated 29 November 2023 —

## Documents

## Details

The ‘Ethics, Transparency and Accountability Framework for Automated Decision-Making’ is a 7 point framework to help government departments use automated or algorithmic decision-making systems safely, sustainably and ethically. It is aimed at civil servants, particularly:

- senior owners of all major processes and services subject to automation consideration
- process and service risk owners
- senior leaders
- executive leaders
- operational staff
- those in digital, data and technology roles
- policy makers
- ministers when considering an algorithm or automated system

The ‘Ethics, Transparency and Accountability Risk Potential Assessment Form’ document will help teams assess the possible risk of an automated or algorithmic decision.

---

Published 13 May 2021

Last updated 29 November 2023 [+ show all updates](#)

---

### Explore the topic

[\*\*Government efficiency, transparency and accountability \(/government/government-efficiency-transparency-and-accountability\)\*\*](#)

---



All content is available under the [Open Government Licence v3.0](#), except where otherwise stated

© Crown copyright



[Home](#) > [Government](#) > [Public services](#)

Guidance

# Guidelines for AI procurement

A summary of best practice addressing specific challenges of acquiring Artificial Intelligence technologies in government.

---

From: [Department for Science, Innovation and Technology \(/government/organisations/department-for-science-innovation-and-technology\)](#), [Office for Artificial Intelligence \(/government/organisations/office-for-artificial-intelligence\)](#), [Department for Digital, Culture, Media & Sport \(/government/organisations/department-for-digital-culture-media-sport\)](#) and [Department for Business, Energy & Industrial Strategy \(/government/organisations/department-for-business-energy-and-industrial-strategy\)](#)

Published 8 June 2020

## Documents

## Details

Artificial Intelligence is a technology that has the potential to greatly improve our public services by reducing costs, enhancing quality, and freeing up valuable time of frontline staff.

Recognising this, the UK Government published the Data Ethics Framework and A Guide to using AI in the Public Sector to enable public bodies to adopt AI systems in a way that works for everyone in society.

These new procurement guidelines will help inform and empower buyers in the public sector, helping them to evaluate suppliers, then confidently and responsibly procure AI technologies for the benefit of citizens.

Published 8 June 2020

---

## Explore the topic

[Public services \(/government/public-services\)](#)

[Science and innovation \(/business-and-industry/science-and-innovation\)](#)



All content is available under the [Open Government Licence v3.0](#), except where otherwise stated

[© Crown copyright](#)

[Home](#)

Guidance

# Planning and preparing for artificial intelligence implementation

Guidance to help you plan and prepare for implementing artificial intelligence (AI).

---

From: [Department for Science, Innovation and Technology \(/government/organisations/department-for-science-innovation-and-technology\)](#), [Office for Artificial Intelligence \(/government/organisations/office-for-artificial-intelligence\)](#) and [Centre for Data Ethics and Innovation \(/government/organisations/centre-for-data-ethics-and-innovation\)](#)

Published 10 June 2019

## Contents

- [Planning your project](#)
- [Start your discovery phase](#)
- [Moving to the alpha phase](#)
- [Moving to the beta phase](#)

[↑ Contents](#)

## Related content

[Managing your artificial intelligence project \(/guidance/managing-your-artificial-intelligence-\)](#)

This guidance is part of a wider collection about [using artificial intelligence \(AI\) in the public sector](https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector) (<https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector>).

Once you have assessed whether AI can help your team meet your users' needs, this guidance will explore the steps you should take to plan and prepare before implementing AI. As with all technology projects and programmes, [you should follow the Technology Code of Practice](https://www.gov.uk/government/publications/technology-code-of-practice/technology-code-of-practice) (<https://www.gov.uk/government/publications/technology-code-of-practice/technology-code-of-practice>).

This guidance is for anyone responsible for:

- deciding how a project runs
- building teams and planning implementation

## Planning your project

As with all projects, you need to make sure you're [hypothesis-led](https://machinelearningmastery.com/what-is-a-hypothesis-in-machine-learning/) (<https://machinelearningmastery.com/what-is-a-hypothesis-in-machine-learning/>) and can constantly iterate to best help your users and their needs.

You should integrate your AI development with your wider project phases.

1. [Discovery](https://www.gov.uk/service-manual/agile-delivery/how-the-discovery-phase-works) (<https://www.gov.uk/service-manual/agile-delivery/how-the-discovery-phase-works>) - consider your current data state, decide whether to build, buy or collaborate, allocate responsibility for AI, assess your existing data, build your AI team, get your data ready for AI, and plan your AI modelling phase.
2. [Alpha](https://www.gov.uk/service-manual/agile-delivery/how-the-alpha-phase-works) (<https://www.gov.uk/service-manual/agile-delivery/how-the-alpha-phase-works>) - build and evaluate your machine learning model.

[↑ Contents](#)

[Natural language processing for Land Registry documentation in Sweden](#)  
 (/government/case-studies/natural-language-processing-for-land-registry-documentation-in-sweden)

[A guide to using artificial intelligence in the public sector](#)  
 (/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector)

[Understanding artificial intelligence ethics and safety](#)  
 (/guidance/understanding-artificial-intelligence-ethics-and-safety)

[Using natural language processing to structure market research](#)  
 (/government/case-studies/using-natural-language-processing-to-structure-market-research)

---

Collection

[A guide to using artificial intelligence in the public sector](#)  
 (/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector)

maintain your model.

intelligence-in-the-public-sector)

You should consider AI ethics and safety (<https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety>) throughout all phases.

Significant time is needed to understand the feasibility of using your data in a new way. This means the discovery phase tends to be longer and more expensive than for services without AI.

Your data scientists may be familiar with a lifecycle called CRISP-DM (<https://www.sv-europe.com/crisp-dm-methodology/>) and may wish to integrate parts of it into your project.

## Start your discovery phase

Discovery can help you understand the problem that needs to be solved.

## Assess your user needs and data sources

You should:

- thoroughly understand the problem and the needs of different users
- assess whether AI is the right tool to address the user needs
- understand the processes and how the AI model will connect with the wider service
- consider the location and condition of the data you will use

## Assess your existing data

To prepare for your AI project, you should assess your existing data. Training an AI system on error-strewn data can result in poor results due to:

- the dataset not containing clear patterns for the

[↑ Contents](#)

- the dataset containing clear but accidental patterns, resulting in the model learning biases

You can use a combination of accuracy, completeness, uniqueness, timeliness, validity, relevancy, representativeness, sufficiency or consistency to see if the data is high enough quality for an AI system to make predictions from.

When assessing your AI data, it's useful to collaborate with someone who has deep knowledge of your data, such as a [data scientist](#)

(<https://www.gov.uk/government/collections/digital-data-and-technology-profession-capability-framework#data:-data-scientist>). They will be familiar with the best practice for measuring, cleaning and maintaining good data standards for ongoing projects. [Make your data proportionate to user needs](#)

(<https://www.gov.uk/guidance/3-use-data-that-is-proportionate-to-the-user-need>) and [understand the limitations of the data](#) (<https://www.gov.uk/guidance/4-understand-the-limitations-of-the-data>) to help you assess your data readiness.

Questions for you to consider with data scientists are:

- do you have enough data for the model to learn from?
- do you understand the onward effects of using data in this way?
- is the data accurate and complete and how frequently is the data updated?
- is the data representative of the users the model's results will impact?
- was the data gathered using suitable, reliable, and impartial sources of measurement?
- is the data secure and do you have permission to use it?
- what modelling approaches could be suitable for the data available?
- do you have access to the data and how quickly can you access it?

[↑ Contents](#)

- what format is the data in and does it require significant preparation to be ready for modelling?
- is your data structured - for example can you store it in a table, or unstructured such as emails or webpages?
- are there any constraints on the data - for example does it contain sensitive information such as home addresses?
- can you link key variables within and between datasets?

If you're unsure about your use of data, consult the [Data Ethics Framework guidance](#) (<https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework>) to check your project is a safe application and deployment of AI.

## Build your team for AI implementation

As with other projects, your [team should be multidisciplinary](#) (<https://www.gov.uk/service-manual/the-team>), with a diverse combination of roles and skills to reduce bias and make sure your results are as accurate as possible. When working with AI you may need specialist roles such as a:

- [data architect](#) (<https://www.gov.uk/government/collections/digital-data-and-technology-profession-capability-framework#technical:-data-architect>) to set the vision for the organisation's use of data, through data design, to meet business needs
- [data scientist](#) (<https://www.gov.uk/government/collections/digital-data-and-technology-profession-capability-framework#data:-data-scientist>) to identify complex business problems while leveraging data value - often having at least 2 data scientists working on a project allows them to better collaborate and validate AI experiments
- [data engineer](#) (<https://www.gov.uk/government/collections/digital-data-and-technology-profession-capability-framework#data:-data-engineer>) to build and maintain the infrastructure required to support the data architecture and data science work

[↑ Contents](#)

products and services into systems and business processes

- ethicist to provide ethical judgements and assessments on the AI model's inputs
- domain expert who knows the environment where you will be deploying the AI model results - for example if the AI model will be investigating social care, collaborate with a social worker

You may not need all of these roles from the very beginning, but this may change as the work progresses. You may want to break up your discovery into smaller phases so you can evaluate what you are learning.

It can be useful for your team to have:

- experience of solving an AI problem similar to the one you're solving
- commercial experience of AI - understanding of machine learning techniques and algorithms, including production deployments at scale
- an understanding of cloud architecture, security, scalable deployment and open source tools and technologies
- hands-on experience of major cloud platforms
- experience with containers and container orchestrations - for example Docker and Kubernetes
- experience in or strong understanding of the fundamentals of computer science and statistics
- experience in software development - for example Python, R or Scala
- experience building large scale backend systems
- hands-on experience with a cluster-computing framework - for example Hadoop or Spark
- hands-on experience with data stores - for example SQL and No-SQL
- technical understanding of streaming data architectures

[↑ Contents](#)

## Managing infrastructure and suppliers

When preparing for AI implementation, you should identify how you can best [integrate AI with your existing technology and services](#) (<https://www.gov.uk/guidance/managing-legacy-technology>).

It's useful to consider how you'll manage:

- data collection pipelines to support reliable model performance and a clean input for modelling, such as batch upload or continuous upload
- storing your data in databases and how the type of database you choose will change depending on the complexity of the project and the different data sources required
- data mining and data analysis of the results
- any platforms your team will use to collate the technology used across the AI project to help speed up AI deployment

When choosing your AI tools, you should bring in specialists, such as data scientists or technical architects to assess what tools you currently have to support AI.

[Use Cloud First](#) (<https://www.gov.uk/guidance/use-cloud-first>) when setting up your infrastructure.

## Consider the benefits of AI platforms

A data science platform is a type of software tool which helps teams connect all of the technology they require across their project workflow, speeding up AI deployment and increasing the transparency and oversight over AI models.

When deciding on whether to use a data science platform, it's useful to consider how the platform can:

- provide access to flexible computation which allow

[↑ Contents](#)

- help your team build workflows for accessing and preparing datasets and allow for easy maintenance of the data
- provide common environments for sharing data and code so the team can work collaboratively
- let your teams clearly share their output through dashboards and applications
- provide a reproducible environment for your teams to work from
- help control and monitor project-specific or sensitive permissions

## Prepare your data for AI

After you've assessed your current data quality, you should prepare your data to make sure it is secure and unbiased. You may find it useful to [create a data factsheet \(\[https://www.turing.ac.uk/sites/default/files/2019-06/understanding\\\_artificial\\\_intelligence\\\_ethics\\\_and\\\_safety.pdf\]\(https://www.turing.ac.uk/sites/default/files/2019-06/understanding\_artificial\_intelligence\_ethics\_and\_safety.pdf\)\)](https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf) during discovery to keep a record of your data quality.

## Ensuring diversity in your data

In the same way you should have [diversity in your team \(<https://www.gov.uk/guidance/5-use-robust-practices-and-work-within-your-skillset>\)](https://www.gov.uk/guidance/5-use-robust-practices-and-work-within-your-skillset), your data should also be diverse and reflective of the population you are trying to model. This will reduce conscious or unconscious bias. Alongside this, a lack of diverse input could mean certain groups are disadvantaged, as the AI model may not cater for a diverse set of needs. You should read the Data Ethics Framework guidance to [understand the limitations of your data \(<https://www.gov.uk/guidance/4-understand-the-limitations-of-the-data>\)](https://www.gov.uk/guidance/4-understand-the-limitations-of-the-data) and how to recognise any bias present.

You should also:

- evaluate the accuracy of your data, how it was collected, and consider alternative sources
- consider if any particular groups might be at an advantage or disadvantage in the context in which

[↑ Contents](#)

- consider the social context of where, when and how the system is being deployed

## Keeping your data secure

Make sure you design your system to keep data secure. To help keep data safe:

- follow the [National Cyber Security Centre's \(NCSC\) guidance on using data with AI](https://www.ncsc.gov.uk/collection/intelligent-security-tools/dealing-with-data) (<https://www.ncsc.gov.uk/collection/intelligent-security-tools/dealing-with-data>)
- make sure your system is compliant with [GDPR](https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/) (<https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/>) and [DPA 2018](https://www.gov.uk/government/collections/data-protection-act-2018) (<https://www.gov.uk/government/collections/data-protection-act-2018>)

As with any other software, you should design and build modular, loosely coupled systems which can be easily iterated and adapted.

Writing and training algorithms can take a lot of time and computational power. In addition to ongoing cost, you'll need to think about the network and memory resources your team will need to train your model.

## Using historic data

Most of the data in government available to train our models is within legacy systems which might contain bias and might have poor controls around it. For legacy systems to be compatible with AI technology, you will often need to invest a lot of work to [bring your legacy systems up to modern standards](https://www.gov.uk/guidance/managing-legacy-technology) (<https://www.gov.uk/guidance/managing-legacy-technology>).

You'll also need to carefully consider the ethical and legal implications of working with historic data and whether you need to seek permission to use this information.

[↑ Contents](#)

## When you complete your data preparation phase you should have:

- a dataset ready for modelling in a technical environment
- a set of features (measurable properties) generated from the raw data set
- a data quality assessment using a combination of accuracy, bias, completeness, uniqueness, timeliness/currency, validity or consistency

## Researching the end to end service

During the discovery phase, you should explore the needs of the users of the end to end service. Like other digital services, you'll use this phase to determine whether there's a viable service you could build that would solve user needs, and that it's cost-effective to pursue the problem.

You'll be able to check guidance on how to [know when your discovery is finished](#) (<https://www.gov.uk/service-manual/agile-delivery/how-the-discovery-phase-works#how-you-know-discovery-is-finished>) before moving on to alpha.

## Moving to the alpha phase

### Plan and prototype your AI model build and service

If you have decided to build your AI model in-house, you should follow these steps.

1. Split the data.
2. Create a baseline model.
3. Build a prototype of the model and service.
4. Test the model and service.
5. Evaluate the model.
6. Assess and refine performance.

[↑ Contents](#)

## Split the data

Your team will need to train the models they build on data. Your team should split your data into a:

- training set to train algorithms during the modelling phase
- validation set for assessing the performance of your models
- test set for a final check on the performance of your best model

## Create a baseline model

Your team should build a simple baseline version model before they build any more complex models. This provides a benchmark that your team can later compare more complex models against, and will help your team identify problems in your data.

## Build a prototype of the model and service

Once you have a baseline model, your team can start prototyping more complex models. This is a highly iterative process, requiring substantial amounts of data, and will see your team probably build a number of AI models before deciding on the most effective and [appropriate algorithm](#) (<https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>) for your problem.

Keeping your team's first AI model simple and setting up the right end-to-end infrastructure will help smooth the transition from alpha to beta. You can action this by focusing on the infrastructure requirements for your AI pipelines at the same time as your team is developing your model. Your simple model will provide you with baseline metrics and information on the model's behaviour that you can use to test more complex models.

Throughout the build, you should [make sure your AI model security complies](#) (<https://www.ncsc.gov.uk/section/information-for/public->

[↑ Contents](#)

## Test the model and service

Your team will need to test your models throughout the process to mitigate against issues such as [overfitting or underfitting](#) (<https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/>) that could undermine your model's effectiveness once deployed.

Your team should only use the test set on your best model. Keep this data separate from your models until this final test. This test will provide you with the most accurate impression of how your model will perform once deployed.

## Evaluate the model

Your team will need to evaluate your model to assess how it is performing against unseen data. This will give you an indication of how your model will perform in the real world.

The best evaluation metric will depend on the problem you are trying to solve, and your chosen model. While you should select the evaluation metric with data scientists, you should also consider the ethical, economical and societal implications. These considerations make the fine tuning of AI systems relevant to both data scientists and delivery leads.

## Choose the final model

When choosing your final model, you will need to consider:

- what level of performance your problem needs
- how interpretable you need your model to be
- how frequently you need predictions or retraining
- the cost of maintaining the model

## Assess and refine performance

Once you select a final model, your team will need to assess its performance, and refine it to make sure it

[↑ Contents](#)

- how it performs compared to simpler models
- what level of performance you need before deploying the model
- what level of performance you can justify to the public, your stakeholders, and regulators
- what level of performance similar applications deliver in other organisations
- whether the model shows any signs of bias

If a model does not outperform human performance, it still may be useful. For example, a text classification algorithm might not be as accurate as a human when classifying documents, however they can perform at a far higher scale and speed than a human.

## Evaluate your Alpha phase

When you complete building your AI prototyping phase, you should have:

- a final model or set of predictive models and a summary of their performance and characteristics
- a decision on whether or not to progress to the beta phase
- a plan for your beta phase

## Moving to the beta phase

Moving from alpha to [beta \(<https://www.gov.uk/service-manual/agile-delivery/how-the-beta-phase-works>\)](https://www.gov.uk/service-manual/agile-delivery/how-the-beta-phase-works) involves integrating the model into the service's decision-making process and using live data for the model to make predictions on.

Using your model in your service has 3 stages.

1. Integrating your model - performance-test the model with live data and integrate it within the decision-making workflow. Integration can happen in a number of ways, from a local deployment to the creation of a custom application for staff or

[↑ Contents](#)

2. Evaluating your model - undertake continuous evaluation to make sure the model still meets business objectives and the model is performing at the level required. This will make sure the model performance is in line with the modelling phase and to help you identify when to retrain the model.
3. Helping users - make sure users feel confident in using, interpreting, and challenging any outputs or insights generated by the model.

You should continue to [collect user needs](#) (<https://www.gov.uk/service-manual/user-research/start-by-learning-user-needs#researching-users-and-their-needs>) so your team can use the model's outputs in the real world.

When moving from alpha to beta, there are some best-practice guidelines to smooth the transition.

## Iterate and deploy improved models

After creating a beta version, your team can use automated testing to create some high-level tests before moving to more thorough testing. Working in this way means you can launch new improvements without worrying about functionality once deployed.

## Maintain a cross-functional team

During alpha, you will have relied mostly on data scientists to assess the opportunity and your data state.

Moving to beta needs specialists with a strong knowledge of dev-ops, servers, networking, data stores, data management, data governance, containers, cloud infrastructure and security design.

This skillset is likely to be better suited to an engineer rather than a data scientist so maintaining a cross-functional team will help smooth the transition from alpha to beta.

[↑ Contents](#)

## When you complete your beta phase, you should have:

- AI running on top of your data, learning and improving its performance, and informing decisions
- a monitoring framework to evaluate the model's performance and rapidly identify incidents
- launched a [private beta](https://www.gov.uk/service-manual/agile-delivery/how-the-beta-phase-works) (<https://www.gov.uk/service-manual/agile-delivery/how-the-beta-phase-works>) followed by a public end-to-end beta prototype which users can use in full
- found a way to measure your service's success using new data you've got during the beta phase
- evidence that your service meets government [accessibility requirements](#) (<https://www.gov.uk/service-manual/helping-people-to-use-your-service/making-your-service-accessible-an-introduction>)
- tested the way you've designed [assisted digital support](#) (<https://www.gov.uk/service-manual/helping-people-to-use-your-service/assisted-digital-support-introduction>) for your service

## Related guides

- [Understanding artificial intelligence](#) (<https://www.gov.uk/government/publications/understanding-artificial-intelligence/a-guide-to-using-artificial-intelligence-in-the-public-sector>)
- [Assessing if artificial intelligence is the right solution](#) (<https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution>)
- [Managing your AI project](#) (<https://www.gov.uk/guidance/managing-your-artificial-intelligence-project>)
- [Understanding artificial intelligence ethics and safety](#) (<https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety>)
- [Examples of real-world artificial intelligence use](#) (<https://www.gov.uk/government/collections/a-guide-to-real-world-examples-of-artificial-intelligence>)

[↑ Contents](#)

- National Cyber Security Centre [guidance for assessing intelligent tools for cyber security](https://www.ncsc.gov.uk/collection/intelligent-security-tools) (<https://www.ncsc.gov.uk/collection/intelligent-security-tools>)
- The [Data Ethics Framework](https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework) (<https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework>)
- The [Technology Code of Practice](https://www.gov.uk/government/publications/technology-code-of-practice/technology-code-of-practice) (<https://www.gov.uk/government/publications/technology-code-of-practice/technology-code-of-practice>)

Published 10 June 2019

---



All content is available under the [Open Government Licence v3.0](#), except where otherwise stated

[© Crown copyright](#)

[↑ Contents](#)



[Home](#) > [Business and industry](#) > [Science and innovation](#) > [Artificial intelligence](#)

## Collection

# A guide to using artificial intelligence in the public sector

Guidance on building and using artificial intelligence in the public sector.

---

From: [Department for Science, Innovation and Technology \(/government/organisations/department-for-science-innovation-and-technology\)](#), [Office for Artificial Intelligence \(/government/organisations/office-for-artificial-intelligence\)](#) and [Centre for Data Ethics and Innovation \(/government/organisations/centre-for-data-ethics-and-innovation\)](#)

Published 10 June 2019

Last updated 18 October 2019 —

## Contents

- [Assess, plan and manage artificial intelligence](#)
- [Using artificial intelligence ethically and safely](#)
- [Examples of artificial intelligence use](#)

---

## Related content

[Understanding artificial intelligence ethics and safety \(/guidance/understanding-](#)

The Government Digital Service (GDS) and the Office for Artificial Intelligence (OAI) have published joint guidance on how to build and use artificial intelligence (AI) in the public sector.

This guidance covers how:

- to assess if using AI will help you meet user needs
- the public sector can best use AI
- to implement AI ethically, fairly and safely

OAI, GDS, and The Alan Turing Institute (ATI) have partnered to produce guidance on how to use AI ethically and safely.

Email [ai-guide@digital.cabinet-office.gov.uk](mailto:ai-guide@digital.cabinet-office.gov.uk) if you:

- want to talk about using AI in the public sector
- have any feedback on the AI guidance
- would like to share an AI case study with us

## Ministerial Foreword

Every day, artificial intelligence (AI) is changing how we experience the world. We already use AI to find the fastest route home, alert us of suspicious activity in our bank accounts and filter out spam emails.

The UK government recognises the importance of this technology's development to both business and the public sector. Indeed, Artificial Intelligence and Data was named as one of the four 'Grand Challenges' in the Industrial Strategy White Paper, which are global trends that will transform our future and contribute to the government's long-term plan to boost productivity in the UK.

PwC estimates that AI could contribute \$15.7tr to the global economy by 2030. The UK is in the top three countries globally in the development of AI technologies and this strength puts us in a prime position to unlock this projected global growth. The same estimates indicate AI could increase our

[artificial-intelligence-ethics-and-safety\)](#)

[Managing your artificial intelligence project](#)  
(/guidance/managing-your-artificial-intelligence-project)

[Using natural language processing to structure market research](#)  
(/government/case-studies/using-natural-language-processing-to-structure-market-research)

[Natural language processing for Land Registry documentation in Sweden](#)  
(/government/case-studies/natural-language-processing-for-land-registry-documentation-in-sweden)

[Using data from electricity meters to predict energy consumption](#)  
(/government/case-studies/using-data-from-electricity-meters-to-predict-energy-consumption)

productivity by 14.3% and grow our GDP up to 10.3% by 2030.

There are huge opportunities for government to capitalise on this exciting new technology to improve lives. We can deliver more for less, and give a better experience as we do so.

For citizens, the application of AI technologies will result in a more personalised and efficient experience. For people working in the public sector it means a reduction in the hours they spend on basic tasks, which will give them more time to spend on innovative ways to improve services.

When government and citizens benefit, so does the economy. This year, the UK government ranked second globally in terms of AI readiness, and as the country most prepared within Western Europe to realise the benefits of AI in delivering public services. Putting this readiness into practice and procuring innovative solutions from the UK's thriving tech sector will, in turn, benefit our economy and grow new and innovative markets across sectors.

We want the public sector to understand AI and embrace the opportunities here. As part of this work, a review into using AI in the public sector, led by the Government Digital Service and the Office for Artificial Intelligence (a joint DCMS / BEIS unit), was undertaken between November 2018 and April 2019. Its purpose was to show us where AI could have the most impact and where investment could yield the greatest benefit.

The findings of the review, published as part of the Government Technology Innovation Strategy, revealed that leaders across the public sector could benefit from better understanding the technology, the opportunities it presents and the limitations of its use.

It also found that delivery teams needed more specific guidance on the different considerations for projects with AI components.

This guide was produced to meet this need, drawing on best practice from the commercial sector and public sector.

We also need to practice what we preach and make sure the public sector is leading from the front in the safe and ethical deployment of AI and other emerging technologies. To reflect this, we have worked with the Alan Turing Institute, the UK's national institute for artificial intelligence, to produce additional guidance on AI ethics and safety in a public sector context.

Maximising the benefits of AI is a priority for government, and this guide is an important step forward towards reaching that goal. We encourage you to put this guidance into practice.

Minister of State for Digital and the Creative Industries, Margot James MP

Minister of State for Universities, Science, Research and Innovation, Chris Skidmore MP

Parliamentary Secretary and Minister for Implementation, Oliver Dowden CBE MP

## Assess, plan and manage artificial intelligence

Guidance from GDS and OAI on how to assess, plan and manage artificial intelligence.

---

[\*\*Understanding artificial intelligence\*\*](#)  
[\(/government/publications/understanding-artificial-intelligence\)](#)

10 June 2019    Guidance

---

[\*\*Assessing if artificial intelligence is the right solution\*\*](#)  
[\(/guidance/assessing-if-artificial-intelligence-is-the-right-solution\)](#)

10 June 2019    Guidance

## [Planning and preparing for artificial intelligence implementation \(/guidance/planning-and-preparing-for-artificial-intelligence-implementation\)](#)

10 June 2019    Guidance

---

## [Managing your artificial intelligence project \(/guidance/managing-your-artificial-intelligence-project\)](#)

10 June 2019    Guidance

## **Using artificial intelligence ethically and safely**

Guidance produced in partnership with The Alan Turing Institute on how to use AI ethically and safely.

---

## [Understanding artificial intelligence ethics and safety \(/guidance/understanding-artificial-intelligence-ethics-and-safety\)](#)

10 June 2019    Guidance

## **Examples of artificial intelligence use**

A collection of examples of how artificial intelligence is being used by the public sector and elsewhere.

---

### [How DFID used satellite images to estimate populations \(/government/case-studies/how-dfid-used-satellite-images-to-estimate-populations\)](#)

10 June 2019    Case study

---

### [How the Department for Transport used AI to improve MOT testing \(/government/case-studies/how-the-department-for-transport-used-ai-to-improve-mot-testing\)](#)

10 June 2019    Case study

## [\*\*How GDS used machine learning to make GOV.UK more accessible \(/government/case-studies/how-gds-used-machine-learning-to-make-govuk-more-accessible\)\*\*](#)

10 June 2019 Case study

---

## [\*\*How a signalling company used AI to help trains run on time \(/government/case-studies/how-a-signalling-company-used-ai-to-help-trains-run-on-time\)\*\*](#)

10 June 2019 Case study

---

## [\*\*Natural language processing for Land Registry documentation in Sweden \(/government/case-studies/natural-language-processing-for-land-registry-documentation-in-sweden\)\*\*](#)

10 June 2019 Case study

---

## [\*\*Using data from electricity meters to predict energy consumption \(/government/case-studies/using-data-from-electricity-meters-to-predict-energy-consumption\)\*\*](#)

10 June 2019 Case study

---

## [\*\*Using natural language processing to structure market research \(/government/case-studies/using-natural-language-processing-to-structure-market-research\)\*\*](#)

10 June 2019 Case study

---

## [\*\*How the Ministry of Justice used AI to compare prison reports \(/government/case-studies/how-the-ministry-of-justice-used-ai-to-compare-prison-reports-2\)\*\*](#)

26 June 2019 Case study

---

## [\*\*How a UK-based bank used AI to increase operational efficiency \(/government/case-studies/how-a-uk-based-bank-used-ai-to-increase-operational-efficiency\)\*\*](#)

18 October 2019 Case study

Published 10 June 2019

Last updated 18 October 2019 [+ show all updates](#)

## Explore the topic

[\*\*Artificial intelligence \(/business-and-industry/artificial-intelligence\)\*\*](#)



All content is available under the [Open Government Licence v3.0](#), except where otherwise stated

[© Crown copyright](#)



[Home](#)

Guidance

# Managing your artificial intelligence project

Understand how to manage a project which uses artificial intelligence.

---

From: [Department for Science, Innovation and Technology \(/government/organisations/department-for-science-innovation-and-technology\)](#), [Office for Artificial Intelligence \(/government/organisations/office-for-artificial-intelligence\)](#) and [Centre for Data Ethics and Innovation \(/government/organisations/centre-for-data-ethics-and-innovation\)](#)

Published 10 June 2019

## Contents

- [Governance when running your AI project](#)
- [Managing risk in your AI project](#)
- [Related guides](#)

This guidance is part of a wider collection about [using artificial intelligence \(AI\) in the public sector](#)

---

## Related content

[Understanding artificial intelligence ethics and safety \(/guidance/understanding-artificial-intelligence-ethics-and-safety\)](#)

[https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector\)](https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector).

Once you have planned and prepared for your AI implementation, you will need to make sure you effectively manage risk and governance.

This guidance is for people responsible for:

- setting governance
- managing risk

## Governance when running your AI project

The Alan Turing Institute (ATI) has written guidance on [how to use AI ethically and safely](#) (<https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety>).

## Safety

Governance in safety is important to make sure the model shows no signs of bias or discrimination. You can consider whether:

- the algorithm is performing in line with [safety and ethical considerations](#) (<https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety>)
- the model is explainable
- there is an agreed definition of fairness implemented in the model
- the data use aligns with the [Data Ethics Framework](#) (<https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework>)
- the algorithm's use of data complies with privacy and data processing legislation

## Purpose

[A guide to using artificial intelligence in the public sector](#)

[\(/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector\)](#)

[Using natural language processing to structure market research](#)  
[\(/government/case-studies/using-natural-language-processing-to-structure-market-research\)](#)

[Using data from electricity meters to predict energy consumption](#)  
[\(/government/case-studies/using-data-from-electricity-meters-to-predict-energy-consumption\)](#)

[How DFID used satellite images to estimate populations](#)  
[\(/government/case-studies/how-dfid-used-satellite-images-to-estimate-populations\)](#)

---

Collection

[A guide to using artificial intelligence in the public sector](#)  
[\(/government/collections/](#)

Governance in purpose makes sure the model is achieving its purpose/business objectives. You can consider whether:

[a-guide-to-using-artificial-intelligence-in-the-public-sector\)](#)

- the model solves the problem identified
- how and when you will evaluate the model
- the user experience aligns with existing government guidance

## Accountability

Governance in accountability provides a clear accountability framework for the model. You can consider:

- whether there is a clear and accountable owner of the model
- who will maintain the model
- who has the ability to change and modify the code

## Testing and monitoring

Governance in testing and monitoring makes sure a robust testing framework is in place. You can consider:

- how you will monitor the model's performance
- who will monitor the model's performance
- how often you will assess the model

## Public narrative

Governance in public narrative protects against reputational risks arising from the application of the model. You can consider whether:

- the project fits with the government organisation's use of AI
- the model fits with the government organisation's policy on data use
- the project fits with how citizens/users expect their data to be used

## Quality assurance

Governance in quality assurance makes sure the code has been reviewed and validated. You can consider whether:

- the team has validated the code
- the code is [open source](https://www.gov.uk/service-manual/technology/making-source-code-open-and-reusable) (<https://www.gov.uk/service-manual/technology/making-source-code-open-and-reusable>)

## Managing risk in your AI project

Risk	How to mitigate
Project shows signs of bias or discrimination	Make sure your model is fair, explainable, and you have a process for monitoring unexpected or biased outputs
Data use is not compliant with legislation, guidance or the government organisation's public narrative	Consult guidance on <a href="https://www.gov.uk/guidance/planning-and-preparing-for-artificial-intelligence-implementation#prepare-your-data-for-ai">preparing your data for AI</a> ( <a href="https://www.gov.uk/guidance/planning-and-preparing-for-artificial-intelligence-implementation#prepare-your-data-for-ai">https://www.gov.uk/guidance/planning-and-preparing-for-artificial-intelligence-implementation#prepare-your-data-for-ai</a> )
Security protocols are not in place to make sure you maintain confidentiality and uphold data integrity	Build a data catalogue to define the security protocols required

Risk	How to mitigate
You cannot access data or it is of poor quality	Map the datasets you will use at an early stage both within and outside your government organisation. It's then useful to assess the data against criteria for a combination of accuracy, completeness, uniqueness, relevancy, sufficiency, timeliness, representativeness, validity or consistency
You cannot integrate the model	Include engineers early in the building of the AI model to make sure any code developed is production-ready
There is no accountability framework for the model	Establish a clear responsibility record to define who has <u>accountability for the different areas of the AI model</u> ( <a href="https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution#allocating-responsibility-and-governance-for-AI-projects">https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution#allocating-responsibility-and-governance-for-AI-projects</a> )

## Related guides

- Understanding artificial intelligence (<https://www.gov.uk/government/publications/understanding-artificial-intelligence/a-guide-to-using-artificial-intelligence-in-the-public-sector>)
- Assessing if artificial intelligence is the right solution (<https://www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution>)
- Planning and preparing for artificial intelligence Implementation (<https://www.gov.uk/guidance/planning-and-preparing-for-artificial-intelligence-implementation>)
- Understanding artificial intelligence ethics and safety (<https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety>)
- Examples of real-world artificial intelligence use (<https://www.gov.uk/government/collections/a-guide-to-real-world-ai-examples>)

## [using-artificial-intelligence-in-the-public-sector#examples-of-artificial-intelligence-use\)](#)

- National Cyber Security Centre [guidance for assessing intelligent tools for cyber security](https://www.ncsc.gov.uk/collection/intelligent-security-tools) (<https://www.ncsc.gov.uk/collection/intelligent-security-tools>)
- The [Data Ethics Framework](https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework) (<https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework>)
- The [Technology Code of Practice](https://www.gov.uk/government/publications/technology-code-of-practice/technology-code-of-practice) (<https://www.gov.uk/government/publications/technology-code-of-practice/technology-code-of-practice>)

Published 10 June 2019

[↑ Contents](#)



All content is available under the [Open Government Licence v3.0](#), except where otherwise stated

[© Crown copyright](#)

## **RM6200 Artificial Intelligence Dynamic Purchasing System - Data Ethics Letter of Understanding**

Purpose: to ensure high standards of ethical conduct are upheld when adopting technologically assisted decision making in the public sector, in accordance with the principles and recommendations in the Committee on Standards and Public Life's report [Artificial Intelligence and Public Standards](#).

It is important that suppliers who bid for work under the RM6200 Artificial Intelligence (AI) Dynamic Purchasing System (DPS) are committed not only to delivering the technical elements of the procurement but also delivering ethically where a buyer has stated that there is an ethical dimension to their tender.

The Office for Artificial Intelligence (AI), Government Digital Service (GDS) and Alan Turing Institute published [ethical principles for data-driven technology](#) in the jointly issued [A Guide to Using Artificial Intelligence in the Public Sector](#), in June 2019.

The Department for Digital, Culture, Media and Sport published a collection page in July 2020, with main [data ethics and AI guidance](#). Public servants working with data and AI will use this collection of guidance when buying technology, products or services under the RM6200 Artificial Intelligence DPS. It is important that suppliers are aware of the standards and frameworks that will affect the buying decisions of Buyer organisations and will adhere to these as appropriate.

Suppliers may be asked to provide evidence of how the government's [Data Ethics Framework](#) principles have been followed during the development and implementation of the technology, product or service, at the award of an Order Contract.

The following list of requirements for Artificial Intelligence suppliers is an example and not exhaustive, and may be developed during the DPS Contract Period and by the Buyer organisation.

### Transparency and explainable AI

- The Supplier should describe the capabilities in the business to ensure the outputs of the AI technology are explainable, and that this explanation is widely available and understandable to a non-expert audience.

### Ethical considerations relation to data limitations, fairness and bias

- The Supplier should identify data limitations and implement strategies to address these data limitations.
- The Supplier should be able to describe the approach to eliminate (or minimise) bias, ethical issues, or other safety risks as a result of using the service.

- The Supplier should be able to describe how they have ensured that the data used to power the AI solution is sufficient in quantity, accuracy and relevance to the data available, and what measures have been taken to mitigate bias in the model.
- The Supplier should be able to demonstrate how they consider the skills, qualifications and diversity of the team developing and deploying AI systems.

#### Consent and Control

- The Supplier should adopt legally sound and ethical consent for processing and capturing data throughout the full lifecycle of the solution and be able to describe the level of human decision-making at critical points.

#### Privacy and cybersecurity

- The Supplier should be able to describe their privacy and cybersecurity approach for the proposed solution, in particular how the data will be protected.

The Supplier shall cooperate in good faith with CCS to develop efficiency tracking performance measures for Data Ethics Performance Indicators in accordance with RM6200 Artificial Intelligence DPS Schedule 4 (DPS Management) clause 4 (How the Supplier's Performance will be measured), if required to do so by CCS.

Suppliers appointed to the RM6200 Artificial Intelligence DPS will continue to meet government standards, guidelines and regulations as they develop in this industry.

In signing this letter of understanding you acknowledge that where a Buyer has stated that there is an ethical dimension to their procurement, you will only bid for work where you are willing to deliver both ethical as well as technical dimensions of a tender.

Signed by the Authorised Representative of .....  
*[insert company name]*

Signature .....

Name (please print) .....

Position .....

Date .....



[Home](#) > [Government](#)

Guidance

# The Sourcing Playbook

Key policies and guidance for making sourcing decisions for the delivery of public services

---

From: [Cabinet Office](#)

[\(/government/organisations/cabinet-office\)](#) and

[Government Commercial Function](#)

[\(/government/organisations/government-commercial-function\)](#)

Published 20 May 2021

Last updated 23 April 2024 —

---

## Documents

### Details

The Sourcing Playbook is the third annual update to the Outsourcing Playbook, capturing best practice from across government within 11 key policies that all central departments are expected to follow. The new Consultancy Playbook provides specific guidance on sourcing consultancy services.

---

### Related content

[The Green Book: appraisal and evaluation in central government](#)  
[\(/government/publications/the-green-book-appraisal-and-evaluation-in-central-government\)](#)

[Government Commercial Function:](#)

[Access 'Should Cost Modelling: Tools and Templates'](https://www.gov.uk/government/publications/should-cost-modelling-tools-and-templates)  
(<https://www.gov.uk/government/publications/should-cost-modelling-tools-and-templates>)

[Standards](#)  
([/government/collections/government-commercial-function](#))

---

Published 20 May 2021  
Last updated 23 April 2024 [+ show all updates](#)

[The Consultancy Playbook](#)  
([/government/publications/the-consultancy-playbook](#))

[Government Functional Standard GovS 008: Commercial and Commercial Continuous Improvement Assessment Framework](#)  
([/government/publications/commercial-operating-standards-for-government](#))

---

Collection

[Financing homebuilding and regeneration](#)  
([/government/collections/financing-homebuilding-and-regeneration](#))

---

## Explore the topic

[Government \(/government/all\)](#)



All content is available under the [Open Government Licence v3.0](#), except where otherwise stated

[© Crown copyright](#)

# The Alan Turing Institute

---

## Understanding artificial intelligence ethics and safety

A guide for the responsible  
design and implementation of AI  
systems in the public sector

Dr David Leslie  
Public Policy Programme



# The Alan Turing Institute

The Public Policy Programme at The Alan Turing Institute was set up in May 2018 with the aim of developing research, tools, and techniques that help governments innovate with data-intensive technologies and improve the quality of people's lives. We work alongside policy makers to explore how data science and artificial intelligence can inform public policy and improve the provision of public services. We believe that governments can reap the benefits of these technologies only if they make considerations of ethics and safety a first priority.

This document provides end-to-end guidance on how to apply principles of AI ethics and safety to the design and implementation of algorithmic systems in the public sector. We will shortly release a workbook to bring the recommendations made in this guide to life. The workbook will contain case studies highlighting how the guidance contained here can be applied to concrete AI projects. It will also contain exercises and practical tools to help strengthen the process-based governance of your AI project.

Please note, that this guide is a living document that will evolve and improve with input from users, affected stakeholders, and interested parties. We need your participation. Please share feedback with us at [policy@turing.ac.uk](mailto:policy@turing.ac.uk)

This work was supported exclusively by the Turing's Public Policy Programme. All research undertaken by the Turing's Public Policy Programme is supported entirely by public funds.

<https://www.turing.ac.uk/research/research-programmes/public-policy>

---

This work is licensed under the terms of the Creative Commons Attribution License 4.0 which permits unrestricted use, provided the original author and source are credited. The license is available at:  
<https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>

Cite this work as:

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute*.  
<https://doi.org/10.5281/zenodo.3240529>

# Table of Contents:

## What is AI ethics?

Intended audience and existing government guidance

AI ethics

## Why AI ethics?

An ethical platform for the responsible delivery of an AI project

Preliminary considerations about the ethical platform

Three building-blocks of a responsible AI project delivery ecosystem

## The SUM Values

### The FAST Track Principles

Fairness

Data fairness

Design fairness

Outcome fairness

Implementation fairness

Putting the principle of discriminatory non-harm into action

Accountability

Accountability deserves consideration both before and after model completion

Sustainability

Stakeholder Impact Assessment

Safety

Accuracy, reliability, security, and robustness

Risks posed to accuracy and reliability

Risks posed to security and robustness

Transparency

Defining transparent AI

Three critical tasks for designing and implementing transparent AI

Mapping AI transparency

## Process transparency: Establishing a Process-Based Governance Framework

## Outcome transparency: Explaining outcomes, clarifying content, implementing responsibly

Defining interpretable AI

Technical aspects of choosing, designing, and using an interpretable AI system

Guidelines for designing and delivering a sufficiently interpretable AI system

Guideline 1: Look first to context, potential impact, and domain-specific need

Guideline 2: Draw on standard interpretable techniques when possible

Guideline 3: Considerations for 'black box' AI systems

Guideline 4: Think about interpretability in terms of capacities for understanding

## Securing responsible delivery through human-centred implementation protocols and practices

Step 1: Consider aspects of application type and domain context to define roles

Step 2: Define delivery relations and map delivery processes

Step 3: Build an ethical implementation platform

## Conclusion

## Bibliography

## What is AI ethics?

### Intended audience and existing government guidance

The following guidance is designed to outline values, principles, and guidelines to assist department and delivery leads in ensuring that they develop and deploy AI ethically, safely, and responsibly. It is designed to complement and supplement the Data Ethics Framework. The [Data Ethics Framework](#) is a practical tool that should be used in any project initiation phase.

### AI ethics

A remarkable time of human promise has been ushered in by the convergence of the ever-expanding availability of big data, the soaring speed and stretch of cloud computing platforms, and the advancement of increasingly sophisticated machine learning algorithms.

This brave new digitally interconnected world is delivering rapid gains in the power of AI to better society. Innovations in AI are already dramatically improving the provision of essential social goods and services from healthcare, education, and transportation to food supply, energy, and environmental management. These bounties are, in fact, likely just the start. Because AI and machine learning systems organically improve with the enlargement of access to data and the growth of computing power, they will only become more effective and useful as the information age continues to develop apace. It may not be long before AI technologies become gatekeepers for the advancement of vital public interests and sustainable human development.

This prospect that progress in AI will help humanity to confront some of its most urgent challenges is exciting, but legitimate worries still abound. As with any new and rapidly evolving technology, a steep learning curve means that mistakes and miscalculations will be made and that both unanticipated and harmful impacts will inevitably occur. AI is no exception.

In order to manage these impacts responsibly and to direct the development of AI systems toward optimal public benefit, you will have to make considerations of **AI ethics and safety a first priority**.

This will involve integrating considerations of the social and ethical implications of the design and use of AI systems into **every stage** of the delivery of your AI project. It will also involve a **collaborative effort** between the data scientists, product managers, data engineers, domain experts, and delivery managers on your team to align the development of artificial intelligence technologies with ethical values and principles that safeguard and promote the wellbeing of the communities that these technologies affect.

By including a primer on AI ethics with the Guide, we are providing you with the conceptual resources and practical tools that will enable you to steward the responsible design and implementation of AI projects.

*AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies.*

These values, principles, and techniques are intended both to motivate morally acceptable practices and to prescribe the basic duties and obligations necessary to produce ethical, fair, and safe AI applications.

## Why AI ethics?

The field of AI ethics has largely emerged as a response to the range of individual and societal harms that the misuse, abuse, poor design, or negative unintended consequences of AI systems may cause. As a way to orient you to the importance of building a robust culture of AI ethics, here is a table that represents some of the most consequential forms that these potential harms may take:

### Potential Harms Caused by AI Systems

#### Bias and Discrimination

Because they gain their insights from the existing structures and dynamics of the societies they analyse, data-driven technologies can reproduce, reinforce, and amplify the patterns of marginalisation, inequality, and discrimination that exist in these societies.

Likewise, because many of the features, metrics, and analytic structures of the models that enable data mining are chosen by their designers, these technologies can potentially replicate their designers' preconceptions and biases.

Finally, the data samples used to train and test algorithmic systems can often be insufficiently representative of the populations from which they are drawing inferences. This creates real possibilities of biased and discriminatory outcomes, because the data being fed into the systems is flawed from the start.

#### Denial of Individual Autonomy, Recourse, and Rights

When citizens are subject to decisions, predictions, or classifications produced by AI systems, situations may arise where such individuals are unable to hold directly accountable the parties responsible for these outcomes.

AI systems automate cognitive functions that were previously attributable exclusively to accountable human agents. This can complicate the designation of responsibility in algorithmically generated outcomes, because the complex and distributed character of the design, production, and implementation processes of AI systems may make it difficult to pinpoint accountable parties.

In cases of injury or negative consequence, such an accountability gap may harm the autonomy and violate the rights of the affected individuals.

#### Non-transparent, Unexplainable, or Unjustifiable Outcomes

Many machine learning models generate their results by operating on high dimensional correlations that are beyond the interpretive capabilities of human scale reasoning. In these cases, the rationale of algorithmically produced outcomes that directly affect decision subjects remains opaque to those subjects. While in some use cases, this lack of explainability may be acceptable, in some applications, where the processed data could

harbour traces of discrimination, bias, inequity, or unfairness, the opaqueness of the model may be deeply problematic.

### Invasions of Privacy

Threats to privacy are posed by AI systems both as a result of their design and development processes, and as a result of their deployment. As AI projects are anchored in the structuring and processing of data, the development of AI technologies will frequently involve the utilisation of personal data. This data is sometimes captured and extracted without gaining the proper consent of the data subject or is handled in a way that reveals (or places under risk the revelation of) personal information.

On the deployment end, AI systems that target, profile, or nudge data subjects without their knowledge or consent could in some circumstances be interpreted as infringing upon their ability to lead a private life in which they are able to intentionally manage the transformative effects of the technologies that influence and shape their development. This sort of privacy invasion can consequently harm a person's more basic right to pursue their goals and life plans free from unchosen influence.

### Isolation and Disintegration of Social Connection

While the capacity of AI systems to curate individual experiences and to personalise digital services holds the promise of vastly improving consumer life and service delivery, this benefit also comes with potential risks. Excessive automation, for example, might reduce the need for human-to-human interaction, while algorithmically enabled hyper-personalisation, by limiting our exposure to worldviews different from ours, might polarise social relationships. Well-ordered and cohesive societies are built on relations of trust, empathy, and mutual understanding. As AI technologies become more prevalent, it is important that these relations be preserved.

### Unreliable, Unsafe, or Poor-Quality Outcomes

Irresponsible data management, negligent design and production processes, and questionable deployment practices can, each in their own ways, lead to the implementation and distribution of AI systems that produce unreliable, unsafe, or poor-quality outcomes. These outcomes can do direct damage to the wellbeing of individual persons and the public welfare. They can also undermine public trust in the responsible use of societally beneficial AI technologies, and they can create harmful inefficiencies by virtue of the dedication of limited public resources to inefficient or even detrimental AI technologies.

## An ethical platform for the responsible delivery of an AI project

Building a project delivery environment, which enables the ethical design and deployment of AI systems, requires a multidisciplinary team effort. It demands the active cooperation of all team members both in maintaining a **deeply ingrained culture of responsibility** and in executing a **governance architecture that adopts ethically sound practices at every point in the innovation and implementation lifecycle**.

This task of uniting an in-built culture of responsible innovation with a governance architecture that brings the values and principles of ethical, fair, and safe AI to life, will require that you and your team accomplish several goals:

- You will have to ensure that your AI project is ***ethically permissible*** by considering the impacts it may have on the wellbeing of affected stakeholders and communities.
- You will have to ensure that your AI project is ***fair and non-discriminatory*** by accounting for its potential to have discriminatory effects on individuals and social groups, by mitigating biases that may influence your model's outputs, and by being aware of the issues surrounding fairness that come into play at every phase of the design and implementation pipeline.
- You will have to ensure that your AI project is ***worthy of public trust*** by guaranteeing to the extent possible the safety, accuracy, reliability, security, and robustness of its product.
- You will have to ensure that your AI project is ***justifiable*** by prioritising both the transparency of the process by which your model is designed and implemented, and the transparency and interpretability of its decisions and behaviours.

We call this governance architecture an ***ethical platform*** for two important reasons. First, it is intended to provide you with a solid, process-based footing of values, principles, and protocols—*an ethical platform to stand on*—so that you and your team are better able to design and implement AI systems ethically, equitably, and safely. Secondly, it is intended to help you facilitate a culture of responsible AI innovation—to *help you provide an ethical platform to stand for*—so that your project team can be united in a collaborative spirit to develop AI technologies for the public good.

### Preliminary considerations about the ethical platform

Our aim for the remainder of this document is to provide you with guidance that is as comprehensive as possible in its presentation of the values, principles, and governance mechanisms necessary to serve the purpose of responsible innovation. Keep in mind, however, that not all issues discussed in this document will apply equally to each project. Clearly, a machine learning algorithm trained to detect spam emails will present fewer ethical challenges compared to one trained to detect cancer in blood samples. Similarly, image recognition systems used for sorting and routing mail raise fewer ethical dilemmas compared to the facial recognition technologies used in law enforcement.

Low-stakes AI applications that are not safety-critical, do not directly impact the lives of people, and do not process potentially sensitive social and demographic data will need less proactive ethical stewardship than high-stakes projects. You and your project team will need to evaluate the scope and possible impacts of your project on affected individuals and communities, and you will have to apply reasonable assessments of the risks posed to individual wellbeing and public welfare in order to formulate proportional governance procedures and protocols.

Be that as it may, you should also keep in mind that all AI projects have social and ethical impacts on stakeholders and communities even if just by diverting or redistributing limited intellectual, material, and economic resources away from other concerns and possibilities for socially beneficial innovation. Ethical considerations and principles-based policy formation should therefore play a salient role in every prospective AI project.

## Three building-blocks of a responsible AI project delivery ecosystem

Setting up an ethical platform for responsible AI project delivery involves not only *building from the cultural ground up*; it involves providing your team with the means to accomplish the goals of establishing the ethical permissibility, fairness, trustworthiness, and justifiability of your project. It will take three building-blocks to make such an ethical platform possible:

1. At the most basic level, it necessitates that you gain a working knowledge of a framework of **ethical values** that *Support, Underwrite, and Motivate* a responsible data design and use ecosystem. These will be called **SUM Values**, and they will be composed of four key notions: *Respect, Connect, Care, and Protect*. The objectives of these SUM Values are (1) to provide you with an accessible framework to start thinking about the moral scope of the societal and ethical impacts of your project and (2) to establish well-defined criteria to evaluate its ethical permissibility.
2. At a second and more concrete level, an ethical platform for responsible AI project delivery requires a set of **actionable principles** that facilitate an orientation to the responsible design and use of AI systems. These will be called **FAST Track Principles**, and they will be composed of four key notions: *Fairness, Accountability, Sustainability, and Transparency*. The objectives of these FAST Track Principles are to provide you with the moral and practical tools (1) to make sure that your project is bias-mitigating, non-discriminatory, and fair, and (2) to safeguard public trust in your project's capacity to deliver safe and reliable AI innovation.
3. At a third and most concrete level, an ethical platform for responsible AI project delivery requires a **process-based governance framework (PBG Framework)** that **operationalises the SUM Values and the FAST Track Principles** across the entire AI project delivery workflow. The objective of this PBG Framework is to set up transparent processes of design and implementation that safeguard and enable the justifiability of both your AI project and its product.

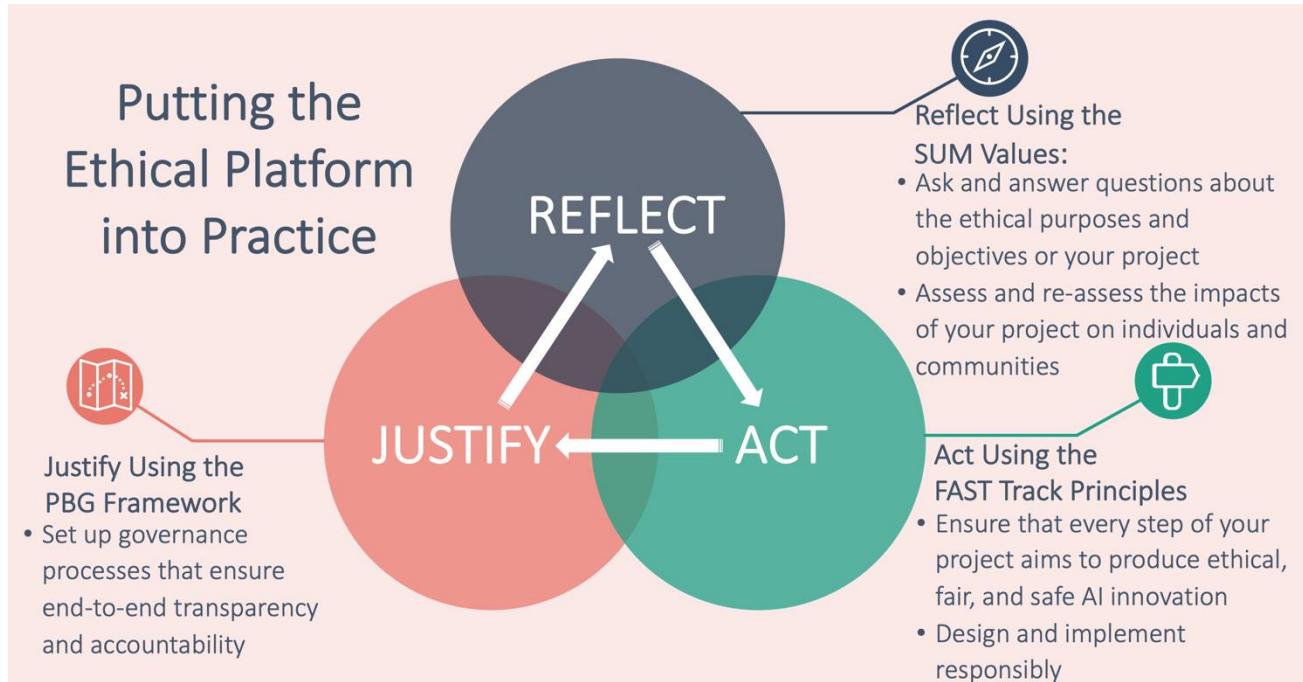
Here is a summary visualisation of these three building blocks of the platform:

### Ethical Platform for the Responsible Delivery of an AI Project



## *How to use this guide*

This guide is intended to assist you in stewarding practices of responsible AI innovation. This entails that the ethical platform be put into practice at every step of the design and implementation workflow. Turning the SUM Values, the FAST Track Principles, and the PBG Framework into practice will require that you and your team continuously **reflect, act, and justify**:



## The SUM Values

### Background

The challenge of creating a culture of responsible innovation begins with the task of building an **accessible moral vocabulary** that will allow team members to explore and discuss the ethical stakes of the AI projects that they are involved in or are considering taking on.

In the field of AI ethics, this moral vocabulary draws primarily on two traditions of moral thinking: (1) **bioethics** and (2) **human rights discourse**. **Bioethics** is the study of the ethical impacts of biomedicine and the applied life sciences. **Human rights discourse** draws inspiration from the UN Declaration of Human Rights. It is anchored in a set of universal principles that build upon the idea that all humans have an equal moral status as bearers of intrinsic human dignity.

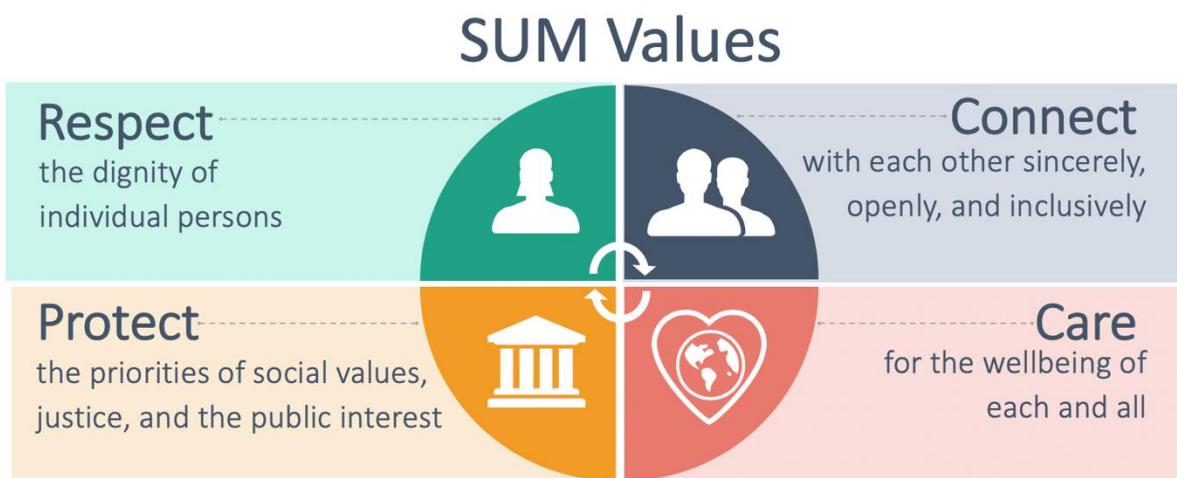
Whereas bioethics largely stresses the normative values that underlie the safeguarding of **individuals** in instances where technological practices affect their interests and wellbeing, human rights discourse mainly focuses on the set of **social, political, and legal entitlements** that are due to all human beings under a universal framework of juridical protection and the rule of law.

The main principles of bioethics include respecting the autonomy of the individual, protecting people from harm, looking after the well-being of others, and treating all individuals equitably and justly. The main tenets of human rights include the entitlement to equal freedom and dignity under the law, the protection of civil, political, and social rights, the universal recognition of personhood, and the right to free and unencumbered participation in the life of the community.

#### *The SUM Values: Respect, Connect, Care, and Protect*

While the SUM Values incorporate conceptual elements from both bioethics and human rights discourse, they do so with an eye to applying the most critical of these elements to the specific social and ethical problems raised by the potential misuse, abuse, poor design, or harmful unintended consequences of AI systems.

They are also meant to be utilised as guiding values throughout the innovation lifecycle: from the preliminary steps of project evaluation, planning, and problem formulation, through processes of design, development, and testing, to the stages of implementation and reassessment. The SUM Values can be visualised as follows:



#### Key Concept: Normativity/Normative

In the context of practical ethics, the word '**normativity**' means that a given concept, value, or belief puts a moral demand on one's practices, i.e. that such a concept, value, or belief indicates what one '**should**' or '**ought to**' do in circumstances where that concept, value, or belief applies. For example, if I hold the moral belief that helping people in need is a good thing, then, when confronted with a sick person in the street who requires help, I should help them. My belief puts a normative demand on me to act in accordance with what it is indicating that I ought to do, namely to come to the needy person's aid.

In order to focus in on a more detailed exploration of each of the values' meanings, their contents will be presented individually. Formulating it as a question: What are each of these values charging you to do?

→ **RESPECT the dignity of individual persons:**

- Ensure their abilities to make free and informed decisions about their own lives
- Safeguard their autonomy, their power to express themselves, and their right to be heard
- Secure their capacities to make well-considered and independent contributions to the life of the community
- Support their abilities to flourish, to fully develop themselves, and to pursue their passions and talents according to their own freely determined life plans

→ **CONNECT with each other sincerely, openly, and inclusively:**

- Safeguard the integrity of interpersonal dialogue, meaningful human connection, and social cohesion
- Prioritise diversity, participation, and inclusion at all points in the design, development, and deployment processes of AI innovation.
- Encourage all voices to be heard and all opinions to be weighed seriously and sincerely throughout the production and use lifecycle
- Use the advancement and proliferation of AI technologies to strengthen the developmentally essential relationship between interacting human beings.
- Utilise AI innovations *pro-socially* so as to enable bonds of interpersonal solidarity to form and individuals to be socialised and recognised by each other
- Use AI technologies to foster this capacity to connect so as to reinforce the edifice of trust, empathy, reciprocal responsibility, and mutual understanding upon which all ethically well-founded social orders rest

→ **CARE for the wellbeing of each and all:**

- Design and deploy AI systems to foster and to cultivate the welfare of all stakeholders whose interests are affected by their use
- Do no harm with these technologies and minimise the risks of their misuse or abuse

- Prioritise the safety and the mental and physical integrity of people when scanning horizons of technological possibility and when conceiving of and deploying AI applications

→ PROTECT the priorities of social values, justice, and the public interest:

- Treat all individuals equally and protect social equity
- Use digital technologies as an essential support for the protection of fair and equal treatment under the law
- Prioritise social welfare, public interest, and the consideration of the social and ethical impacts of innovation in determining the legitimacy and desirability of AI technologies
- Use AI to empower and to advance the interests and well-being of as many individuals as possible
- Think big-picture about the wider impacts of the AI technologies you are conceiving and developing. Think about the ramifications of their effects and externalities for others around the globe, for future generations, and for the biosphere as a whole

As a general rule, these SUM Values should orient you in deliberating about the **ethical permissibility** of a prospective AI project. They should also provide you with a framework of concepts to consider the **ethical impacts of an AI system across the design, use, and monitoring lifecycle**.

Taking these SUM Values as a starting point of conversation, you should also encourage discussion within your team of how to weigh the values against one another and how to consider trade-offs should use case specific circumstances arise when the values come into tension with each other.

## The FAST Track Principles:

### Background

While the SUM Values are intended to provide you with some general normative guideposts and moral motivations for thinking through the social and ethical aspects of AI project delivery, they are not specifically catered to the actual processes involved in developing and deploying AI systems.

To make clear what is needed for this next step toward a more actionable orientation to the responsible design and use of AI technologies, it would be helpful to briefly touch upon what has necessitated the emergence of AI ethics in the first place.

Marvin Minsky, the great cognitive scientist and AI pioneer, defined AI as follows: ‘Artificial Intelligence is the science of ***making computers do things that require intelligence*** when done by humans.’ This standard definition should key us in to the principal motivation that has driven the development of the field of the applied ethics of artificial intelligence:

When humans do ‘things that require intelligence,’ we hold them responsible for the accuracy, reliability, and soundness of their judgements. Moreover, we demand of them that their actions and decisions be supported by good reasons, and we hold them accountable for the fairness, equity, and reasonableness of how they treat others.

What creates the need for principles tailored to the design and use of AI systems is that their emergence and expanding power ‘to do things that require intelligence’ has heralded a shift of a wide array of cognitive functions to algorithmic processes that themselves can be held neither directly responsible nor immediately accountable for the consequences of their behaviour.

As inert and program-based machinery, AI systems are not morally accountable agents. This has created an ethical breach in the sphere of the applied science of AI that the growing number of frameworks for AI ethics are currently trying to fill. Targeted principles such as fairness, accountability, sustainability, and transparency are meant to ‘fill the gap’ between the new ‘smart agency’ of machines and their fundamental lack of moral responsibility.

### The FAST Track Principles: Fairness, Accountability, Sustainability, and Transparency

By becoming well-acquainted with the FAST Track Principles, *all members* of your project delivery team will be better able to support a responsible environment for data innovation.

Issues of fairness, accountability, sustainability, and transparency operate at every juncture and at every level of the AI project delivery workflow and demand the cooperative attention and deliberative involvement of those with technical expertise, domain knowledge, project/product management skill, and policy competence. Ethical AI innovation is a team effort from start to finish.

To introduce you to the scope of the FAST Track Principles, here is a summary visualisation of them:

## FAST Track Principles



You should keep in mind, initially, that while fairness, accountability, sustainability, and transparency are grouped together in the FAST acronym, they do not necessarily relate to each other on the same plane or as equivalents. The principles of accountability and transparency are ***end-to-end governing principles***. Accountability entails that humans are answerable for the parts they play across the entire AI design and implementation workflow. It also demands that the results of this work are traceable from start to finish. The principle of transparency entails that design and implementation processes are justifiable through and through. It demands as well that an algorithmically influenced outcome is interpretable and made understandable to affected parties.

The governing roles of accountability and transparency are very different from the more dependent roles of fairness and sustainability. These latter two are *qualities* of algorithmic systems for which their designers and implementers are ***held accountable*** through the ***transparency both of the outcomes of their practices and of the practices themselves***. According to the principle of fairness, designers and implementers are held accountable for being equitable and for not harming anyone through bias or discrimination. According to the principle of sustainability, designers and implementers are held accountable for producing AI innovation that is safe and ethical in its outcomes and wider impacts.

Whereas the principles of transparency and accountability thus provide the procedural mechanisms and means through which AI systems can be justified and by which their producer and implementers can be held responsible, fairness and sustainability are the crucial aspects of the design, implementation, and outcomes of these systems which establish the normative criteria for such governing constraints. These four principles are therefore all deeply interrelated, but they are not equal.

There is one more important thing to keep in mind before we delve into the details of the FAST Track principles. Transparency, accountability, and fairness are *also data protection principles*, and where algorithmic processing involves personal data, complying with them is not simply a matter of ethics or good practice, but a legal requirement, which is enshrined in the General Data Protection Regulation (GDPR) and the Data Protection Act of 2018 (DPA 2018). For more detailed information about the specific meanings of transparency, accountability, and fairness as data protection principles in the context of the GDPR and the DPA 2018, please refer to the [Guide to Data Protection](#) produced by the Information Commissioner's Office.

## Fairness

When thinking about fairness in the design and deployment of AI systems, it is important to always keep in mind that these technologies, no matter how neutral they may seem, are designed and produced by human beings, who are bound by the limitations of their contexts and biases.

Human error, prejudice, and misjudgement can enter into the innovation lifecycle and create biases at any point in the project delivery process from the preliminary stages of data extraction, collection, and pre-processing to the critical phases of problem formulation, model building, and implementation.

Additionally, data-driven technologies achieve accuracy and efficacy by building inferences from datasets that record complex social and historical patterns, which themselves may contain culturally crystallised forms of bias and discrimination. There is no silver bullet when it comes to remediating the dangers of discrimination and unfairness in AI systems. The problem of fairness and bias mitigation in algorithmic design and use therefore has no simple or strictly technical solution.

That said, best practices of fairness-aware design and implementation (both at the level of non-technical self-assessment and at the level of technical controls and means of evaluation) hold great promise in terms of securing just, morally acceptable, and beneficial outcomes that treat affected stakeholders fairly and equitably.

While there are different ways to characterise or define fairness in the design and use of AI systems, you should consider the **principle of discriminatory non-harm** as a minimum required threshold of fairness. This principle directs us to do no harm to others through the biased or discriminatory outcomes that may result from practices of AI innovation:

**Principle of Discriminatory Non-Harm:** The designers and users of AI systems, which process social or demographic data pertaining to features of human subjects, societal patterns, or cultural formations, should prioritise the mitigation of bias and the exclusion of discriminatory influences on the outputs and implementations of their models. Prioritising discriminatory non-harm implies that the designers and users of AI systems ensure that the decisions and behaviours of their models do not generate discriminatory or inequitable impacts on affected individuals and communities. This entails that these designers and users ensure that the AI systems they are developing and deploying:

1. Are trained and tested on properly representative, relevant, accurate, and generalisable datasets (**Data Fairness**)
2. Have model architectures that do not include target variables, features, processes, or analytical structures (correlations, interactions, and inferences) which are unreasonable, morally objectionable, or unjustifiable (**Design Fairness**)
3. Do not have discriminatory or inequitable impacts on the lives of the people they affect (**Outcome Fairness**)
4. Are deployed by users sufficiently trained to implement them responsibly and without bias (**Implementation Fairness**)

### *Data fairness*

Responsible data acquisition, handling, and management is a necessary component of algorithmic fairness. If the results of your AI project are generated by biased, compromised, or skewed datasets, affected stakeholders will not adequately be protected from discriminatory harm. Your project team should keep in mind the following key elements of data fairness:

- **Representativeness:** Depending on the context, either underrepresentation or overrepresentation of disadvantaged or legally protected groups in the data sample may lead to the systematic disadvantaging of vulnerable stakeholders in the outcomes of the trained model. To avoid such kinds of sampling bias, domain expertise will be crucial to assess the fit between the data collected or procured and the underlying population to be modelled. Technical team members should, if possible, offer means of remediation to correct for representational flaws in the sampling.
- **Fit-for-Purpose and Sufficiency:** An important question to consider in the data collection and procurement process is: Will the amount of data collected be sufficient for the intended purpose of the project? The quantity of data collected or procured has a significant impact on the accuracy and reasonableness of the outputs of a trained model. A data sample not large enough to represent with sufficient richness the significant or qualifying attributes of the members of a population to be classified may lead to unfair outcomes. Insufficient datasets may not equitably reflect the qualities that should rationally be weighed in producing a justified outcome that is consistent with the desired purpose of the AI system. Members of the project team with technical and policy competences should collaborate to determine if the data quantity is, in this respect, sufficient and fit-for-purpose.
- **Source Integrity and Measurement Accuracy:** Effective bias mitigation begins at the very commencement of data extraction and collection processes. Both the sources and instruments of measurement may introduce discriminatory factors into a dataset. When incorporated as inputs in the training data, biased prior human decisions and judgments—such as prejudiced scoring, ranking, interview-data or evaluation—will become the ‘ground truth’ of the model and replicate the bias in the outputs of the system. In order to secure discriminatory non-harm, you must do your best to make sure your data sample has optimal source integrity. This involves securing or confirming that the data gathering processes involved suitable, reliable, and impartial sources of measurement and sound methods of collection.
- **Timeliness and Recency:** If your datasets include outdated data then changes in the underlying data distribution may adversely affect the generalisability of your trained model. Provided these distributional drifts reflect changing social relationship or group dynamics, this loss of accuracy with regard to the actual characteristics of the underlying population may introduce bias into your AI system. In preventing discriminatory outcomes, you should scrutinise the timeliness and recency of all elements of the data that constitute your datasets.
- **Relevance, Appropriateness and Domain Knowledge:** The understanding and utilisation of the most appropriate sources and types of data are crucial for building a robust and unbiased AI system. Solid domain knowledge of the underlying population distribution and of the predictive or classificatory goal of the project is instrumental for choosing optimally relevant measurement inputs that contribute to the reasonable determination of the defined solution. You should make sure that domain experts collaborate closely with your technical team to assist in the determination of the optimally appropriate categories and sources of measurement.

To ensure the uptake of best practices for responsible data acquisition, handling, and management across your AI project delivery workflow, you should initiate the creation of a **Dataset Factsheet** at the alpha stage of your project. This factsheet should be maintained diligently throughout the design and implementation lifecycle in order to secure optimal data quality, deliberate bias-mitigation aware practices, and optimal auditability. It should include a **comprehensive record of data provenance, procurement, pre-processing, lineage, storage, and security** as well as qualitative input from team members about determinations made with regard to data representativeness, data sufficiency, source integrity, data timeliness, data relevance, training/testing/validating splits, and unforeseen data issues encountered across the workflow.

### *Design Fairness*

Because human beings have a hand in all stages of the construction of AI systems, fairness-aware design must take precautions across the AI project workflow to prevent bias from having a discriminatory influence:

- **Problem Formulation:** At the initial stage of problem formulation and outcome definition, technical and non-technical members of your team should work together to translate project goals into measurable targets. This will involve the use of both domain knowledge and technical understanding to define what is being optimised in a formalisable way and to translate the project's objective into a target variable or measurable proxy, which operates as a statistically actionable rendering of the defined outcome.

At each of these points, choices must be made about the design of the algorithmic system that may introduce structural biases which ultimately lead to discriminatory harm. Special care must be taken here to identify affected stakeholders and to consider how vulnerable groups might be negatively impacted by the specification of outcome variables and proxies. Attention must also be paid to the question of whether these specifications are reasonable and justifiable given the general purpose of the project and the potential impacts that the outcomes of the system's use will have on the individuals and communities involved.

These challenges of fairness aware design at the problem formulation stage show the need for making diversity and inclusive participation a priority from the start of the AI project lifecycle. This involves both the collaboration of the entire team and the attainment of stakeholder input about the acceptability of the project plan. This also entails collaborative deliberation across the project team and beyond about the ethical impacts of the design choices made.

- **Data Pre-Processing:** Human judgment enters into the process of algorithmic system construction at the stage of labelling, annotating, and organising the training data to be utilised in building the model. Choices made about how to classify and structure raw inputs must be taken in a fairness aware manner with due consideration given to the sensitive social contexts that may introduce bias into such acts of classification. Similar fairness aware processes should be put in place to review automated or outsourced classifications. Likewise, efforts should be made to attach solid contextual information and ample metadata to the datasets, so that downstream analyses of data processing have access to properties of concern in bias mitigation.

- **Feature Determination and Model-Building:** The constructive task of selecting the attributes or features that will serve as input variables for your model involves human decisions being made about what sorts of information may or may not be relevant or rationally required to yield an accurate *and* unbiased classification or prediction. Moreover, the feature engineering tasks of aggregating, extracting, or decomposing attributes from datasets may introduce human appraisals that have biasing effects. For this reason, discrimination awareness should play a large role at this stage of the AI model-building workflow as should domain knowledge and policy expertise. Your team should proceed in the modelling stage aware that choices made about grouping or separating and including or excluding features as well as more general judgements about the comprehensiveness or coarseness of the total set of features may have significant consequences for vulnerable or protected groups.

The process of tuning hyperparameters and setting metrics at the modelling, testing, and evaluation stages also involves human choices that may have discriminatory effects in the trained model. Your technical team should proceed with an attentiveness to bias risk, and continual iterations of peer review and project team consultation should be encouraged to ensure that choices made in adjusting the dials and metrics of the model are in line with bias mitigation and discriminatory non-harm.

- **Evaluating Analytical Structures:** Design fairness also demands close assessment of the existence in the trained model of lurking or hidden proxies for discriminatory features that may act as significant factors in its output. Including such hidden proxies in the structure of the model may lead to implicit ‘redlining’ (the unfair treatment of a sensitive group on the basis of an unprotected attribute or interaction of attributes that ‘stands in’ for a protected or sensitive one).

Designers must additionally scrutinise the moral justifiability of the significant correlations and inferences that are determined by the model’s learning mechanisms themselves. In cases of the processing of social or demographic data related to human features, where the complexity and high dimensionality of machine learning models preclude the confirmation of the discriminatory non-harm of these inferences (for reason of their uninterpretability by human assessors), these models should be avoided. In AI systems that process and draw analytics from data arising from human relationships, societal patterns, and complex socioeconomic and cultural formations, designers must prioritise a degree of interpretability that is sufficient to ensure that the inferences produced by these systems are non-discriminatory. In cases where this is not possible, a different, more transparent and explainable model or portfolio of models should be chosen.

Analytical structures must also be confirmed to be *procedurally fair*. Any rule or procedure employed in an AI system should be consistently and uniformly applied to every decision subject whose information is being processed by that system. Your team should be able to certify that when a rule or procedure has been used to render an outcome for any given individual, the same rule or procedure will be applied to any other individual in the same way regardless of that other subject’s similarities with or differences from the first.

Implementers, in this respect, should be able to show that any algorithmic output is replicable when the same rules and procedures are applied to the same inputs. Such a uniformity of the application of rules and procedures secures the equal procedural treatment of decision subjects and precludes any rule-changes in the algorithmic processing targeted at a specific person that may disadvantage that individual vis-à-vis any other.

### *Outcome fairness*

As part of this minimum safeguarding of discriminatory non-harm, forethought and well-informed consideration must be put into *how you are going to define and measure the fairness of the impacts and outcomes of the AI system you are developing*.

There is a great diversity of beliefs in the area of **outcome fairness** as to how to properly classify what makes the consequences of an algorithmically supported decision equitable, fair, and allocatively just. Different approaches—detailed below—stress different principles: some focus on demographic parity, some on individual fairness, others on error rates equitably distributed across subpopulations.

Your determination of outcome fairness should heavily depend both on the **specific use case for which the fairness of outcome is being considered** and the **technical feasibility of incorporating your chosen criteria into the construction of the AI system**. (Note that different fairness-aware methods involve different types of technical interventions at the pre-processing, modelling, or post-processing stages of production). Again, this means that determining your fairness definition should be a **cooperative and multidisciplinary effort across the project team**.

You will find below a summary table of some of the main definitions of outcome fairness that have been integrated by researchers into formal models as well as a list of current articles and technical resources, which should be consulted to orient your team to the relevant knowledge base. (Note that this is a rapidly developing field, so your technical team should keep updated about further advances.) The first four fairness types fall under the category of group fairness and allow for comparative criteria of non-discrimination to be considered in model construction and evaluation. The final two fairness types focus instead on cases of individual fairness, where context-specific issues of effective bias are considered and assessed at the level of the individual agent.

Take note, though, that these technical approaches have limited scope in terms of the bigger picture issues of algorithmic fairness that we have already stressed. Many of the formal approaches work only in use cases that have *distributive or allocative consequences*. In order to carry out group comparisons, these approaches require access to data about sensitive/protected attributes (that may often be unavailable or unreliable) as well as accurate demographic information about the underlying population distribution. Furthermore, there are unavoidable trade-offs and inconsistencies between these technical definitions that must be weighed in determining which of them are best fit for your use case. Consult those on your project team with the technical expertise to consider the use case appropriateness of a desired formal approach.

Some Formalisable Definitions of Outcome Fairness	
Type of Fairness	Definition
<b>Demographic/ Statistical Parity</b>	An outcome is fair if each group in the selected set receives benefit in equal or similar proportions, i.e. if there is no correlation between a sensitive or protected attribute and the allocative result. This approach is intended to prevent <i>disparate impact</i> , which occurs when the outcome of an algorithmic process disproportionately harms members of disadvantaged or protected groups.
<b>True Positive Rate Parity</b>	An outcome is fair if the ‘true positive’ rates of an algorithmic prediction or classification are equal across groups. This approach is intended to align the goals of bias mitigation and accuracy by ensuring that the accuracy of the model is equivalent between relevant population subgroups. This method is also referred to as ‘equal opportunity’ fairness because it aims to secure equalised odds of an advantageous outcome for qualified individuals in a given population regardless of the protected or disadvantaged groups of which they are members.
<b>False Positive Rate Parity</b>	An outcome is fair if it does not disparately mistreat people belonging to a given social group by misclassifying them at a higher rate than the members of a second social group, for this would place the members of the first group at an unfair disadvantage. This approach is motivated by the position that sensitive groups and advantaged groups should have similar error rates in outcomes of algorithmic decisions.
<b>Positive Predictive Value Parity</b>	An outcome is fair if the rates of positive predictive value (the fraction of correctly predicted positive cases out of all predicted positive cases) are equal across sensitive and advantaged groups. Outcome fairness is defined here in terms of a parity of precision, where the probability of members from different groups actually having the quality they are predicted to have is the same across groups.
<b>Individual Fairness</b>	An outcome is fair if it treats individuals with similar relevant qualifications similarly. This approach relies on the establishment of a similarity metric that shows the degree to which pairs of individuals are alike with regard to a specific task.
<b>Counterfactual Fairness</b>	An outcome is fair if an automated decision made about an individual belonging to a sensitive group would have been the same were that individual a member of a different group in a closest possible alternative (or counterfactual) world. Like the individual fairness approach, this method of defining fairness focuses on the specific circumstances of an affected decision subject, but, by using the tools of contrastive explanation, it moves beyond individual fairness insofar as it brings out the causal influences behind the algorithmic output. It also presents the possibility of offering the subject of an automated decision knowledge of what factors, if changed, could have influenced a different outcome. This could provide them with actionable recourse to change an unfavourable decision.

## Selected References and Technical Resources

- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM. (Statistical Parity and Individual Fairness)
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, 325–333. (Demographic Parity)
- Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323. (Equality of Opportunity)
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2):153–163. (Balancing Error Rates)
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM. (Test for Disparate Impact)
- Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. International World Wide Web Conferences Steering Committee. (Disparate Mistreatment)
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, 1-7. Fairware '18. (Summary and Comparison)
- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4066–4076. (Counterfactual Fairness)
- Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10-19. (Extension of Counterfactual Fairness)

### Technical Resources for Exploring Fairness Tools:

- <https://dsapp.uchicago.edu/projects/aequitas/> (University of Chicago's open source bias audit toolkit for machine learning developers)
- <http://www.fairness-measures.org/> and [https://github.com/megantosh/fairness\\_measures\\_code/](https://github.com/megantosh/fairness_measures_code/) (Datasets and software for detecting algorithmic discrimination from TU Berlin and Eurecat)
- <https://github.com/columbia/fairtest> (Fairtest unwarranted association discovery platform from Columbia University)
- <http://aif360.mybluemix.net/#> (IBM's Fairness 360 open source toolkit)

### Fairness Position Statement:

Once you and your project team have thoroughly considered the use case appropriateness as well as technical feasibility of the formal models of fairness most relevant for your system and have incorporated the model into your application, you should prepare a **Fairness Position Statement (FPS)** in which the fairness criteria being employed in the AI system is made explicit and explained in plain and non-technical language. This FPS should then be made publicly available for review by all affected stakeholders.

### *Implementation fairness*

When your project team is approaching the beta stage, you should begin to build out your plan for implementation training and support. This plan should include adequate preparation for the responsible and unbiased deployment of the AI system by its on-the-ground users. Automated

decision-support systems present novel risks of bias and misapplication at the point of delivery, so special attention should be paid to preventing harmful or discriminatory outcomes at this critical juncture of the AI project lifecycle.

In order to design an optimal regime of implementer training and support, you should pay special attention to the unique pitfalls of bias-in-use to which the deployment of AI technologies give rise. These can be loosely classified as decision-automation bias (more commonly just ‘automation bias’) and automation-distrust bias:

- **Decision-Automation Bias/The Technological Halo Effect:** Users of automated decision-support systems may tend to become hampered in their critical judgment, rational agency, and situational awareness as a result of their faith in the perceived objectivity, neutrality, certainty, or superiority of the AI system.

This may lead to **over-reliance** or **errors of omission**, where implementers lose the capacity to identify and respond to the faults, errors, or deficiencies, which might arise over the course of the use of an automated system, because they become complacent and overly deferent to its directions and cues. Decision-automation bias may also lead to **over-compliance** or **errors of commission** where implementers defer to the perceived infallibility of the system and thereby become unable to detect problems emerging from its use for reason of a failure to hold the results against available information.

Both over-reliance and over-compliance may lead to what is known as out-of-loop syndrome where the degradation of the role of human reason and the deskilling of critical thinking hampers the user’s ability to complete the tasks that have been automated. This condition may bring about a loss of the ability to respond to system failure and may lead both to safety hazards and to dangers of discriminatory harm.

To combat risks of decision-automation bias, you should operationalise strong regimes of accountability at the site of user deployment to steer human decision-agents to act on the basis of good reasons, solid inferences, and critical judgment.

- **Automation-Distrust Bias:** At the other extreme, users of an automated decision-support system may tend to disregard its salient contributions to evidence-based reasoning either as a result of their distrust or skepticism about AI technologies in general or as a result of their over-prioritisation of the importance of prudence, common sense, and human expertise. An aversion to the non-human and amoral character of automated systems may also influence decision subjects’ hesitation to consult these technologies in high impact contexts such as healthcare, transportation, and law.

In order to secure and safeguard fair implementation of AI systems by users well-trained to utilise the algorithmic outputs as tools for making evidence-based judgements, you should consider the following measures:

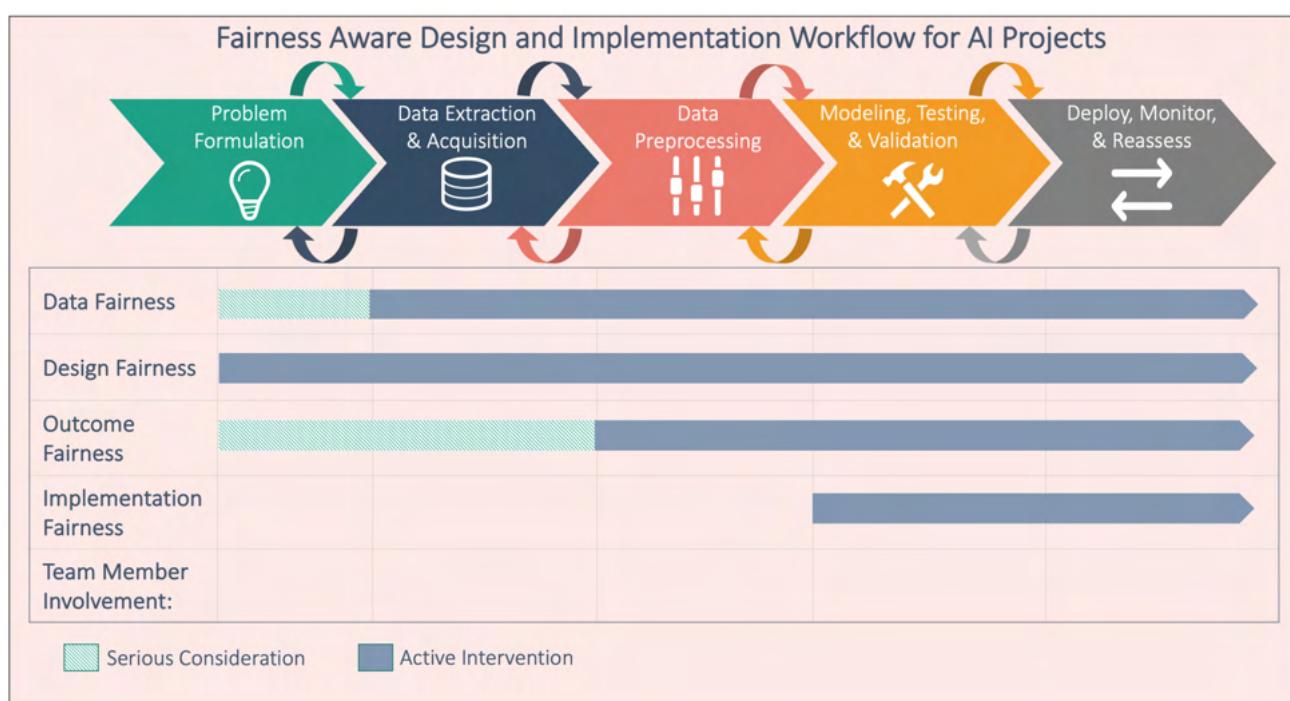
- Training of implementers should include the conveyance of basic knowledge about the statistical and probabilistic character of machine learning and about the limitations of AI and automated decision-support technologies. This training should avoid any anthropomorphic

(or human-like) portrayals of AI systems and should encourage users to view the benefits and risks of deploying these systems in terms of their role in assisting human judgment rather than replacing it.

- Forethought should be given in the design of the user-system interface about human factors and about possibilities for implementation biases. The systems should be *designed into* processes that encourage active user judgment and situational awareness. The interface between the user and the system should be designed to make clear and accessible to the user the system's rationale, compliance to fairness standards, and confidence level. Ideally this should happen in a 'runtime' manner.
- Training of implementers should include a pre-emptive exploration of the cognitive and judgmental biases that may occur across deployment contexts. This should be done in a use case based manner that highlights the particular misjudgements that may occur when people weigh statistical evidence. Examples of the latter may include overconfidence in prediction based on the historical consistency of data, illusions that any clustering of data points necessarily indicates significant insights, and discounting of societal patterns that exist beyond the statistical results.

#### *Putting the principle of discriminatory non-harm into action*

When you are considering how to put the principle of discriminatory non-harm into action, you should come together with all the managers on the project team to map out team member involvement at each stage of the AI project pipeline from alpha through beta. Considering fairness aware design and implementation from a workflow perspective will allow you, as a team, to concretise and make explicit end-to-end paths of accountability in a clear and peer-reviewable manner. This is essential for establishing a robust accountability framework. Here is a schematic representation of the fairness aware workflow. You will have to complete the final row.



Considering fairness aware design and implementation from such a workflow perspective will also assist you in pinpointing risks of bias or downstream discrimination and streamlining possible solutions in a proactive, pre-emptive, and anticipatory way. At each stage of the AI project pipeline (i.e. at each column of the above table), you and the relevant members of your team should carry out a collaborative self-assessment with regard to the applicable dimension of fairness. This is a three-step process:

#### Discriminatory Non-Harm Self-Assessment

Step 1: Identify the fairness and bias mitigation dimensions that apply to the specific stage under consideration (for example, at the data pre-processing stage, dimensions of data fairness, design fairness, and outcome fairness may be at issue).

Step 2: Scrutinise how your particular AI project might pose risks or have unintended vulnerabilities in each of these areas.

Step 3: Take action to correct any existing problems that have been identified, strengthen areas of weakness that have possible discriminatory consequences, and take proactive bias-prevention measures in areas that have been identified to pose potential future risks.

## Accountability

When considering the role of accountability in the AI project delivery lifecycle, it is important first to make sure that you are taking a ‘best practices’ approach to data processing that is aligned with [Principle 6 of the Data Ethics Framework](#). Beyond following this general guidance, however, you should pay special attention to the new and unique challenges posed to public sector accountability by the design and implementation of AI systems.

Responsible AI project delivery requires that two related challenges to public sector accountability be confronted directly:

1. **Accountability gap:** As mentioned above, automated decisions are not self-justifiable. Whereas human agents can be called to account for their judgements and decisions in instances where those judgments and decisions affect the interests of others, the statistical models and underlying hardware that compose AI systems are not responsible in the same morally relevant sense. This creates an accountability gap that must be addressed so that clear and imputable sources of human answerability can be attached to decisions assisted or produced by an AI system.
2. **Complexity of AI production processes:** Establishing human answerability is not a simple matter when it comes to the design and deployment of AI systems. This is due to the complexity and multi-agent character of the development and use of these systems. Typically, AI project delivery workflows include department and delivery leads, technical

experts, data procurement and preparation personnel, policy and domain experts, implementers, and others. Due to this production complexity, it may become difficult to answer the question of who among these parties involved in the production of AI systems should bear responsibility if these systems' uses have negative consequences and impacts.

Meeting the special requirements of accountability, which are born out of these two challenges, call for a sufficiently fine-grained concept of what would make an AI project properly accountable. This concept can be broken down into two subcomponents of accountability: **answerability** and **auditability**:

- **Answerability:** The principle of accountability demands that the onus of justifying algorithmically supported decisions be placed on the shoulders of the human creators and users of those AI systems. This means that it is essential to establish a continuous chain of human responsibility across the whole AI project delivery workflow. Making sure that accountability is effective from end to end necessitates that no gaps be permitted in the answerability of responsible human authorities from first steps of the design of an AI system to its algorithmically steered outcomes.

Answerability also demands that explanations and justifications of both the content of algorithmically supported decisions and the processes behind their production be offered by competent human authorities in plain, understandable, and coherent language. These explanations and justifications should be based upon sincere, consistent, sound, and impartial reasons that are accessible to non-technical hearers.

- **Auditability:** Whereas the notion of answerability responds to the question of *who is accountable* for an automation supported outcome, the notion of auditability answers the question of *how the designers and implementers of AI systems are to be held accountable*. This aspect of accountability has to do with **demonstrating** both the **responsibility of design and use practices** and the **justifiability of outcomes**.

Your project team must ensure that every step of the process of designing and implementing your AI project is accessible for audit, oversight, and review. Successful audit requires builders and implementers of algorithmic systems to keep records and to make accessible information that enables monitoring of the soundness and diligence of the innovation processes that produced the AI system.

Auditability also requires that your project team keep records and make accessible information that enables monitoring of data provenance and analysis from the stages of collection, pre-processing, and modelling to training, testing, and deploying. This is the purpose of the previously mentioned Dataset Factsheet.

Moreover, it requires your team to enable peers and overseers to probe and to critically review the dynamic operation of the system in order to ensure that the procedures and operations which are producing the model's behaviour are safe, ethical, and fair. Practically transparent algorithmic models must be **built for auditability, reproducible**, and **equipped for end-to-end recording and monitoring** of their data processing.

The deliberate incorporation of both of these elements of accountability (answerability and auditability) into the AI project lifecycle may be called **Accountability-by-Design**:

**Accountability by Design:** All AI systems must be designed to facilitate end-to-end answerability and auditability. This requires both **responsible humans-in-the-loop** across the entire design and implementation chain as well as **activity monitoring protocols** that enable end-to-end oversight and review.

### *Accountability deserves consideration across the entire design and implementation workflow*

As a best practice, you should actively consider the different demands that accountability by design places on you before and after the roll out of your AI project. We will refer to the process of ensuring accountability during the design and development stages of your AI project as '**anticipatory accountability**.' This is because you are anticipating your AI project's accountability needs prior to it being completed. Following a similar logic, we will refer to the process of addressing accountability after the start of the deployment of your AI project as '**remedial accountability**.' This is because after the initial implementation of your system, you are remedying any of the issues that may be raised by its effects and potential externalities. These two subtypes of accountability are sometimes referred to as *ex-ante* (or before-the-event) accountability and *ex-post* (after-the-event) accountability respectively.

- **Anticipatory Accountability:** Treating accountability as an anticipatory principle entails that you take as of primary importance the decisions made and actions taken by your project delivery team prior to the outcome of an algorithmically supported decision process.

This kind of *ex ante* accountability should be prioritised over remedial accountability, which focuses instead on the corrective or justificatory measures that can be taken after that automation supported process had been completed.

By ensuring the AI project delivery processes are accountable prior to the actual application of the system in the world, you will bolster the soundness of design and implementation processes and thereby more effectively pre-empt possible harms to individual wellbeing and public welfare.

Likewise, by establishing strong regimes of anticipatory accountability and by making the design and delivery process as open and publicly accessible as possible, you will put affected stakeholders in a position to make better informed and more knowledgeable decisions about their involvement with these systems in advance of potentially harmful impacts. In doing so, you will also strengthen the public narrative and help to safeguard the project from reputational harm.

- **Remedial Accountability:** While remedial accountability should be seen, along these lines, as a necessary fallback rather than as a first resort for imputing responsibility in the design and deployment of AI systems, strong regimes of remedial accountability are no less important in

providing necessary justifications for the bearing these systems have on the lives of affected stakeholders.

**Putting in place comprehensive auditability regimes as part of your accountability framework and establishing transparent design and use practices, which are methodically logged throughout the AI project delivery lifecycle, are essential components for this sort of remedial accountability.**

One aspect of remedial accountability that you must pay close attention to is the need to provide **explanations** to affected stakeholders for algorithmically supported decisions. This aspect of accountable and transparent design and use practices will be called **explicability**, which literally means the ability to make explicit the meaning of the algorithmic model's result.

Offering explanations for the results of algorithmically supported decision-making involves furnishing decision subjects and other interested parties with an understandable account of the rationale behind the specific outcome of interest. It also involves furnishing the decision subject and other interested parties with an explanation of the ethical permissibility, the fairness, and the safety of the use of the AI system. These tasks of **content clarification** and **practical justification** will be explored in more detail below as part of the section on transparency.

## Sustainability

Designers and users of AI systems should remain aware that these technologies may have transformative and long-term effects on individuals and society. In order to ensure that the deployment of your AI system remains sustainable and supports the sustainability of the communities it will affect, you and your team should proceed with a continuous sensitivity to the real-world impacts that your system will have.

### *Stakeholder Impact Assessment*

You and your project team should come together to evaluate the social impact and sustainability of your AI project through a **Stakeholder Impact Assessment (SIA)**, whether the AI project is being used to deliver a public service or in a back-office administrative capacity. When we refer to 'stakeholders' we are referring primarily to affected individual persons, but the term may also extend to groups and organisations in the sense that individual members of these collectives may also be impacted as such by the design and deployment of AI systems. Due consideration to stakeholders should be given at both of these levels.

The purpose of carrying out an SIA is multidimensional. SIAs can serve several purposes, some of which include:

- (1) Help to build public confidence that the design and deployment of the AI system by the public sector agency has been done responsibly
- (2) Facilitate and strengthen your accountability framework
- (3) Bring to light unseen risks that threaten to affect individuals and the public good

- (4) Underwrite well-informed decision-making and transparent innovation practices
- (5) Demonstrate forethought and due diligence not only within your organisation but also to the wider public

Your team should convene to evaluate the social impact and sustainability of your AI project through the SIA at three critical points in the project delivery lifecycle:

- 1. Alpha Phase (Problem Formulation):** Carry out an initial Stakeholder Impact Assessment (SIA) to determine the ethical permissibility of the project. Refer to the SUM Values as a starting point for the considerations of the possible effects of your project on individual wellbeing and public welfare. In cases where you conclude that your AI project will have significant ethical and social impacts, you should open your initial SIA to the public so that their views can be properly considered. This will bolster the inclusion of a diversity of voices and opinions into the design and development process through the participation of a more representative range of stakeholders. You should also consider consulting with internal organisational stakeholders, whose input will likewise strengthen the openness, inclusivity, and diversity of your project.
- 2. From Alpha to Beta (Pre-Implementation):** Once your model has been trained, tested, and validated, you and your team should revisit your initial SIA to confirm that the AI system to be implemented is still in line with the evaluations and conclusions of your original assessment. This check-in should be logged on the pre-implementation section of the SIA with any applicable changes added and discussed. Before the launch of the system, this SIA should be made publicly available. At this point you must also set a timeframe for re-assessment once the system is in operation as well as a public consultation which predates and provides input for that re-assessment. Timeframes for these re-assessments should be decided by your team on a case-by-case basis but should be proportional to the scale of the potential impact of the system on the individuals and communities it will affect.
- 3. Beta Phase (Re-Assessment):** After your AI system has gone live, your team should intermittently revisit and re-evaluate your SIA. These check-ins should be logged on the re-assessment section of the SIA with any applicable changes added and discussed. Re-assessment should focus both on evaluating the existing SIA against real world impacts and on considering how to mitigate the unintended consequences that may have ensued in the wake of the deployment of the system. Further public consultation for input at the beta stage should be undertaken before the re-assessment so that stakeholder input can be included in re-assessment deliberations.

You should keep in mind that, in its specific focus on social and ethical sustainability, your Stakeholder Impact Assessment constitutes just one part of the governance platform for your AI project and should be a complement to your accountability framework and other auditing and activity-monitoring documentation.

Your SIA should be broken down into four sections of questions and responses. In the 1<sup>st</sup> section, there should be general questions about the possible big-picture social and ethical impacts of the use of the AI system you plan to build. In the 2<sup>nd</sup> section, your team should collaboratively formulate relevant sector-specific and use case-specific questions about the impact of the AI system on

affected stakeholders. The 3<sup>rd</sup> section should provide answers to the additional questions relevant to pre-implementation evaluation. The 4<sup>th</sup> section should provide the opportunity for members of your team to reassess the system in light of its real-world impacts, public input, and possible unintended consequences.

Here is a prototype of an SIA:

<b><u>Stakeholder Impact Assessment for (Project Name)</u></b>	
<p><b>1. Alpha Phase (Problem Formulation) General Questions</b></p> <p><b>Completed on this Date:</b></p>	<p><b>I. Identifying Affected Stakeholders</b></p> <p>Who are the stakeholders that this AI project is most likely to affect? What groups of these stakeholders are most vulnerable? How might the project negatively impact them?</p> <p><b>II. Goal-Setting and Objective-Mapping</b></p> <p>How are you defining the outcome (the target variable) that the system is optimising for? Is this a fair, reasonable, and widely acceptable definition?</p> <p>Does the target variable (or its measurable proxy) reflect a reasonable and justifiable translation of the project's objective into the statistical frame?</p> <p>Is this translation justifiable given the general purpose of the project and the potential impacts that the outcomes of its implementation will have on the communities involved?</p> <p><b>III. Possible Impacts on the Individual</b></p> <p>How might the implementation of your AI system impact the abilities of affected stakeholders to make free, independent, and well-informed decisions about their lives? How might it enhance or diminish their autonomy?</p> <p>How might it affect their capacities to flourish and to fully develop themselves?</p> <p>How might it do harm to their physical or mental integrity? Have risks to individual health and safety been adequately considered and addressed?</p> <p>How might it infringe on their privacy rights, both on the data processing end of designing the system and on the implementation end of deploying it?</p> <p><b>IV. Possible Impacts on Society and Interpersonal Relationships</b></p> <p>How might the implementation of your AI system adversely affect each stakeholder's fair and equal treatment under the law? Are there any aspects of the project that expose vulnerable communities to possible discriminatory harm?</p> <p>How might the use of your system affect the integrity of interpersonal dialogue, meaningful human connection, and social cohesion?</p>

	<p>Have the values of civic participation, inclusion, and diversity been adequately considered in articulating the purpose and setting the goals of the project? If not, how might these values be incorporated into your project design?</p> <p>Does the project aim to advance the interests and well-being of as many affected individuals as possible? Might any disparate socioeconomic impacts result from its deployment?</p> <p>Have you sufficiently considered the wider impacts of the system on future generations and on the planet as a whole?</p>
<p><b>2. Alpha Phase (Problem Formulation)</b> <b>Sector-Specific and Use Case-Specific Questions</b></p> <p><b>Completed on this Date:</b></p>	<p>In this section you should consider the sector-specific and use case-specific issues surrounding the social and ethical impacts of your AI project on affected stakeholders. Compile a list of the questions and concerns you anticipate. State how your team is attempting to address these questions and concerns.</p>
<p><b>3. From Alpha to Beta (Pre-Implementation)</b></p> <p><b>Completed on this Date:</b></p>	<p>After reviewing the results of your initial SIA, answer the following questions:</p> <p>Are the trained model's actual objective, design, and testing results still in line with the evaluations and conclusions contained in your original assessment? If not, how does your assessment now differ?</p> <p>Have any other areas of concern arisen with regard to possibly harmful social or ethical impacts as you have moved from the alpha to the beta phase?</p> <p>You must also set a reasonable timeframe for public consultation and beta phase re-assessment:</p> <p><b>Dates of Public Consultation on Beta-Phase Impacts:</b></p> <p><b>Date of Planned Beta Phase Re-Assessment:</b></p>
<p><b>4. Beta Phase (Re-Assessment)</b></p> <p><b>Completed on this Date:</b></p>	<p>Once you have reviewed the most recent version of your SIA and the results of the public consultation, answer the following questions:</p> <p>How does the content of the existing SIA compare with the real-world impacts of the AI system as measured by available evidence of performance, monitoring data, and input from implementers and the public?</p> <p>What steps can be taken to rectify any problems or issues that have emerged?</p> <p>Have any unintended harmful consequences ensued in the wake of the deployment of the system? If so, how might these negative impacts be mitigated and redressed?</p>

	<p>Have the maintenance processes for your AI model adequately taken into account the possibility of distributional shifts in the underlying population? Has the model been properly retuned and retrained to accommodate changes in the environment?</p> <p><b>Dates of Public Consultation on Beta-Phase Impacts:</b></p> <p><b>Date of Next Planned Beta Phase Re-Assessment:</b></p>
--	--

## Safety

Beyond safeguarding the sustainability of your AI project as it relates to its social impacts on individual wellbeing and public welfare, your project team must also confront the related challenge of **technical sustainability or safety**. A technically sustainable AI system is **safe, accurate, reliable, secure, and robust**. Securing these goals, however, is a difficult and unremitting task.

Because AI systems operate in a world filled with uncertainty, volatility, and flux, the challenge of building safe and reliable AI can be especially daunting. This job, however, must be met head-on. Only by making the goal of producing safe and reliable AI technologies central to your project, will you be able to mitigate risks of your system failing at scale when faced with real-world unknowns and unforeseen events. The issue of **AI safety** is of paramount importance, because these potential failures may both produce harmful outcomes and undermine public trust.

In order to safeguard that your AI system functions safely, you must prioritise the technical objectives of **accuracy, reliability, security, and robustness**. This requires that your technical team put careful forethought into how to construct a system that **accurately and dependably operates in accordance with its designers' expectations even when confronted with unexpected changes, anomalies, and perturbations**. Building an AI system that meets these safety goals also requires rigorous testing, validation, and re-assessment as well as the integration of adequate mechanisms of oversight and control into its real-world operation.

### *Accuracy, reliability, security, and robustness*

It is important that you gain a strong working knowledge of each of the safety relevant operational objectives (**accuracy, reliability, security, and robustness**):

- **Accuracy and Performance Metrics:** In machine learning, the accuracy of a model is the proportion of examples for which it generates a correct output. This performance measure is also sometimes characterised conversely as an **error rate** or the fraction of cases for which the model produces an incorrect output. Keep in mind that, in some instances, the choice of an acceptable error rate or accuracy level can be adjusted in accordance with the use case specific needs of the application. In other instances, it may be largely set by a domain established benchmark.

As a performance metric, accuracy should be a central component to establishing and nuancing your team's approach to safe AI. That said, specifying a reasonable performance level for your system may also often require you to refine or exchange your measure of accuracy. For instance, if certain errors are more significant or costly than others, a metric for total cost can be integrated into your model so that the cost of one class of errors can be weighed against that of another. Likewise, if the precision and sensitivity of the system in detecting uncommon events is a priority (say, in instances of the medical diagnosis of rare cases of a disease), you can use the technique of precision and recall. This method of addressing imbalanced classification would allow you to weigh the proportion of the system's correct detections—both of frequent and of rare outcomes—against the proportion of actual detections of the rare event (i.e. the ratio of the true detections of the rare outcome to the sum of the true detections of that outcome and the missed detections or false negatives for that outcome).

In general, measuring accuracy in the face of uncertainty is a challenge that must be given significant thought. The confidence level of your AI system will depend heavily on problems inherent in attempts to model a chaotic and changing reality. Concerns about accuracy must cope with issues of unavoidable noise present in the data sample, architectural uncertainties generated by the possibility that a given model is missing relevant features of the underlying distribution, and inevitable changes in input data over time.

- **Reliability:** The objective of reliability is that an AI system behaves exactly as its designers intended and anticipated. A reliable system adheres to the specifications it was programmed to carry out. Reliability is therefore a measure of **consistency** and can establish confidence in the safety of a system based upon the dependability with which it operationally conforms to its intended functionality.
- **Security:** The goal of security encompasses the protection of several operational dimensions of an AI system when confronted with possible adversarial attack. A secure system is capable of maintaining the **integrity** of the information that constitutes it. This includes protecting its architecture from the unauthorised modification or damage of any of its component parts. A secure system also remains continuously **functional** and **accessible** to its authorised users and keeps **confidential** and **private information** secure even under hostile or adversarial conditions.
- **Robustness:** The objective of robustness can be thought of as the goal that an AI system functions reliably and accurately under harsh conditions. These conditions may include adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications). The measure of robustness is therefore the strength of a system's integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, and undesirable reinforcement learning behaviour.

*Risks posed to accuracy and reliability:*

**Concept Drift:** Once trained, most machine learning systems operate on static models of the world that have been built from historical data which have become fixed in the systems' parameters. This freezing of the model before it is released 'into the wild' makes its accuracy and reliability especially vulnerable to changes in the underlying distribution of data. When the historical data that have crystallised into the trained model's architecture cease to reflect the population concerned, the model's mapping function will no longer be able to accurately and reliably transform its inputs into its target output values. These systems can quickly become prone to error in unexpected and harmful ways.

There has been much valuable research done on methods of detecting and mitigating concept and distribution drift, and you should consult with your technical team to ensure that its members have familiarised themselves with this research and have sufficient knowledge of the available ways to confront the issue. In all cases, you should remain vigilant to the potentially rapid concept drifts that may occur in the complex, dynamic, and evolving environments in which your AI project will intervene. Remaining aware of these transformations in the data is crucial for safe AI, and your team should actively formulate an action plan to anticipate and to mitigate their impacts on the performance of your system.

**Brittleness:** Another possible challenge to the accuracy and reliability of machine learning systems arises from the inherent limitations of the systems themselves. Many of the high-performing machine learning models such as deep neural nets (DNN) rely on massive amounts of data and brute force repetition of training examples to tune the thousands, millions, or even billions of parameters, which collectively generate their outputs.

However, when they are actually running in an unpredictable world, these systems may have difficulty processing unfamiliar events and scenarios. They may make unexpected and serious mistakes, because they have neither the capacity to contextualise the problems they are programmed to solve nor the common-sense ability to determine the relevance of new 'unknowns'. Moreover, these mistakes may remain unexplainable given the high-dimensionality and computational complexity of their mathematical structures. This fragility or brittleness can have especially significant consequences in safety-critical applications like fully automated transportation and medical decision support systems where undetectable changes in inputs may lead to significant failures. While progress is being made in finding ways to make these models more robust, it is crucial to consider safety first when weighing up their viability.

### *Risks posed to security and robustness*

**Adversarial Attack:** Adversarial attacks on machine learning models maliciously modify input data—often in imperceptible ways—to induce them into misclassification or incorrect prediction. For instance, by undetectably altering a few pixels on a picture, an adversarial attacker can mislead a model into generating an incorrect output (like identifying a panda as a gibbon or a 'stop' sign as a 'speed limit' sign) with an extremely high confidence. While a good amount of attention has been paid to the risks that adversarial attacks pose in deep learning applications like computer vision, these kinds of perturbations are also effective across a vast range of machine learning techniques and uses such as spam filtering and malware detection.

These vulnerabilities of AI systems to adversarial examples have serious consequences for AI safety. The existence of cases where subtle but targeted perturbations cause models to be misled into gross miscalculation and incorrect decisions have potentially serious safety implication for the adoption of critical systems like applications in autonomous transportation, medical imaging, and security and surveillance. In response to concerns about the threats posed to a safe and trusted environment for AI technologies by adversarial attacks a field called **adversarial machine learning** has emerged over the past several years. Work in this area focuses on securing systems from disruptive perturbations at all points of vulnerability across the AI pipeline.

One of the major safety strategies that has arisen from this research is an approach called **model hardening**, which has advanced techniques that combat adversarial attacks by strengthening the architectural components of the systems. Model hardening techniques may include adversarial training, where training data is methodically enlarged to include adversarial examples. Other model hardening methods involve architectural modification, regularisation, and data pre-processing manipulation. A second notable safety strategy is **run-time detection**, where the system is augmented with a discovery apparatus that can identify and trace in real-time the existence of adversarial examples. You should consult with members of your technical team to ensure that the risks of adversarial attack have been taken into account and mitigated throughout the AI lifecycle. A valuable collection of resources to combat adversarial attack can be found at <https://github.com/IBM/adversarial-robustness-toolbox>.

**Data Poisoning:** A different but related type of adversarial attack is called data poisoning. This threat to safe and reliable AI involves a malicious compromise of data sources at the point of collection and pre-processing. Data poisoning occurs when an adversary modifies or manipulates part of the dataset upon which a model will be trained, validated, and tested. By altering a selected subset of training inputs, a poisoning attack can induce a trained AI system into curated misclassification, systemic malfunction, and poor performance. An especially concerning dimension of targeted data poisoning is that an adversary may introduce a ‘backdoor’ into the infected model whereby the trained system functions normally until it processes maliciously selected inputs that trigger error or failure.

In order to combat data poisoning, your technical team should become familiar with the state of the art in filtering and detecting poisoned data. However, such technical solutions are not enough. Data poisoning is possible because data collection and procurement often involves potentially unreliable or questionable sources. When data originates in uncontrollable environments like the internet, social media, or the Internet of Things, many opportunities present themselves to ill-intentioned attackers, who aim to manipulate training examples. Likewise, in third- party data curation processes (such as ‘crowdsourced’ labelling, annotation, and content identification), attackers may simply handcraft malicious inputs. Your project team should focus on the best practices of responsible data management, so that they are able to tend to data quality as an end-to-end priority.

- **Misdirected Reinforcement Learning Behaviour:** A different set of safety risks emerges from the approach to machine learning called reinforcement learning (RL). In the more widely

applied methods of supervised learning that have largely been the focus of this guide, a model transforms inputs into outputs according to a fixed mapping function that has resulted from its passively received training. In RL, by contrast, the learner system actively solves problems by engaging with its environment through trial and error. This exploration and ‘problem-solving’ behaviour is determined by the objective of maximising a reward function that is defined by its designers.

This flexibility in the model, however, comes at the price of potential safety risks. An RL system, which is operating in the real-world without sufficient controls, may determine a reward-optimising course of action that is optimal for achieving its desired objective but harmful to people. Because these models lack context-awareness, common sense, empathy, and understanding, they are unable to identify, on their own, scenarios that may have damaging consequences but that were not anticipated and constrained by their programmers. This is a difficult problem, because the unbounded complexity of the world makes anticipating all of its pitfalls and detrimental variables veritably impossible.

Existing strategies to mitigate such risks of misdirected reinforcement learning behaviour include:

- Running extensive simulations during the testing stage, so that appropriate measures of constraint can be programmed into the system
- Continuous inspection and monitoring of the system, so that its behaviour can be better predicted and understood
- Finding ways to make the system more interpretable so that its decisions can be better assessed
- Hard-wiring mechanisms into the system that enable human override and system shut-down

## End-to-End AI Safety

The safety risks you face in your AI project will depend, among other factors, on the sort of algorithm(s) and machine learning techniques you are using, the type of applications in which those techniques are going to be deployed, the provenance of your data, the way you are specifying your objective, and the problem domain in which that specification applies. As a best practice, regardless of this variability of techniques and circumstances, safety considerations of accuracy, reliability, security, and robustness should be in operation at every stage of your AI project lifecycle.

This should involve both **rigorous protocols of testing, validating, verifying, and monitoring the safety of the system** and the performance of **AI safety self-assessments** by relevant members of your team at each stage of the workflow. Such self-assessments should evaluate how your team’s design and implementation practices line up with the safety objectives of accuracy, reliability, security, and robustness. Your AI safety self-assessments should be logged across the workflow on a single document in a running fashion that allows review and re-assessment.

## Transparency

## Defining transparent AI

It is important to remember that ***transparency as a principle of AI ethics*** differs a bit in meaning from the everyday use of the term. The common dictionary understanding of transparency defines it as either (1) the quality an object has when one can see clearly through it or (2) the quality of a situation or process that can be clearly justified and explained because it is open to inspection and free from secrets.

Transparency as a principle of AI ethics encompasses *both* of these meanings:

On the one hand, transparent AI involves the interpretability of a given AI system, i.e. **the ability to know how and why a model performed the way it did in a specific context and therefore to understand the rationale behind its decision or behaviour**. This sort of transparency is often referred to by way of the metaphor of ‘opening the black box’ of AI. It involves ***content clarification and intelligibility*** or ***explicability***.

On the other hand, transparent AI involves **the justifiability both of the processes that go into its design and implementation and of its outcome**. It therefore involves the ***soundness of the justification of its use***. In this more normative meaning, transparent AI is ***practically justifiable*** in an unrestricted way if one can demonstrate that both the design and implementation processes that have gone into the particular decision or behaviour of a system and the decision or behaviour itself are ***ethically permissible, non-discriminatory/fair, and worthy of public trust/safety-securing***.

## *Three critical tasks for designing and implementing transparent AI*

This two-pronged definition of transparency as a principle of AI ethics asks that you to think about transparent AI both in terms of the *process* behind it (the design and implementation practices that lead to an algorithmically supported outcome) and in terms of its *product* (the content and justification of that outcome). Such a process/product distinction is crucial, because it clarifies the three tasks that your team will be responsible for in safeguarding the transparency of your AI project:

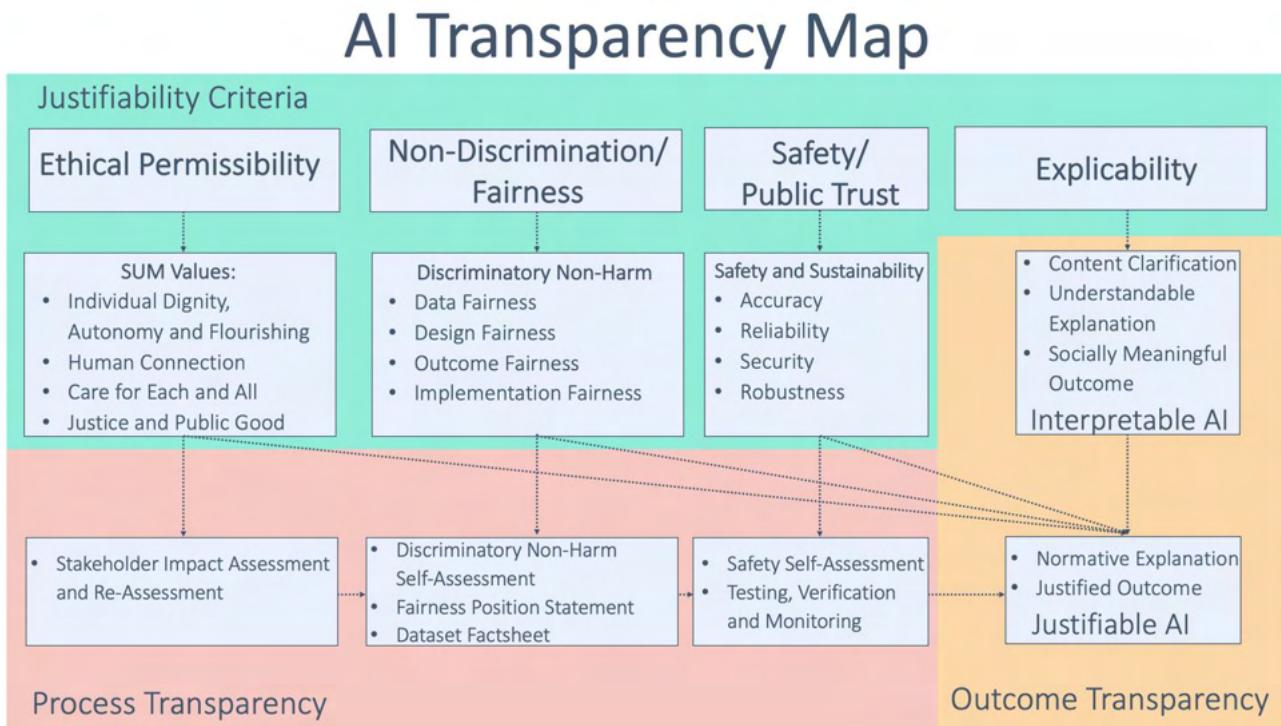
- **Process Transparency, Task 1: Justify Process.** In offering an explanation to affected stakeholders, you should be able to demonstrate that considerations of ethical permissibility, non-discrimination/fairness, and safety/public trustworthiness were operative end-to-end in the design and implementation processes that lead to an automated decision or behaviour. This task will be supported both by following the best practices outlined herein throughout the AI project lifecycle and by putting into place robust auditability measures through an accountability-by-design framework.
- **Outcome Transparency, Task 2: Clarify Content and Explain Outcome.** In offering an explanation to affected stakeholders, you should be able to show in plain language that is understandable to non-specialists how and why a model performed the way it did in a specific decision-making or behavioural context. You should therefore be able to clarify and communicate the rationale behind its decision or behaviour. This explanation should be *socially meaningful* in the sense that the terms and logic of the explanation should not simply

reproduce the formal characteristics or the technical meanings and rationale of the mathematical model but should rather be translated into the everyday language of human practices and therefore be understandable in terms of the societal factors and relationships that the decision or behaviour implicates.

- **Outcome Transparency, Task 3: Justify Outcome.** In offering an explanation to affected stakeholders, you should be able to demonstrate that a specific decision or behaviour of your system is ethically permissible, non-discriminatory/fair, and worthy of public trust/safety-securing. This outcome justification should take the content clarification/explicated outcome from task 2 as its starting point and weigh that explanation against the justifiability criteria adhered to throughout the design and use pipeline: ethical permissibility, non-discrimination/fairness, and safety/public trustworthiness. Undertaking an optimal approach to process transparency from the start should support and safeguard this demand for normative explanation and outcome justification.

### *Mapping AI transparency*

Before exploring each of the three tasks individually, it may be helpful to visualise the relationship between these connected components of transparent AI:



### Process Transparency: Establishing a Process-Based Governance Framework

The central importance of the end-to-end operability of good governance practices should guide your strategy to build out responsible AI project workflow processes. Three components are essential to creating a such a responsible workflow: (1) Maintaining strong regimes of professional and institutional transparency; (2) Having a clear and accessible Process-Based Governance

Framework (PBG Framework); (3) Establishing a well-defined auditability trail in your PBG Framework through robust activity logging protocols that are consolidated digitally in a process log.

1. **Professional and Institutional Transparency:** At every stage of the design and implementation of your AI project, team members should be held to rigorous standards of conduct that secure and maintain professionalism and institutional transparency. These standards should include the core values of **integrity, honesty, sincerity, neutrality, objectivity and impartiality**. All professionals involved in the research, development, production, and implementation of AI technologies are, first and foremost, acting as **fiduciaries of the public interest** and must, in keeping with these core values of the Civil Service, put the obligations to serve that interest above any other concerns.

Furthermore, from start to finish of the AI project lifecycle, the design and implementation process should be as transparent and as open to public scrutiny as possible with restrictions on accessibility to relevant information limited to the reasonable protection of justified public sector confidentiality and of analytics that may tip off bad actors to methods of gaming the system of service provision.

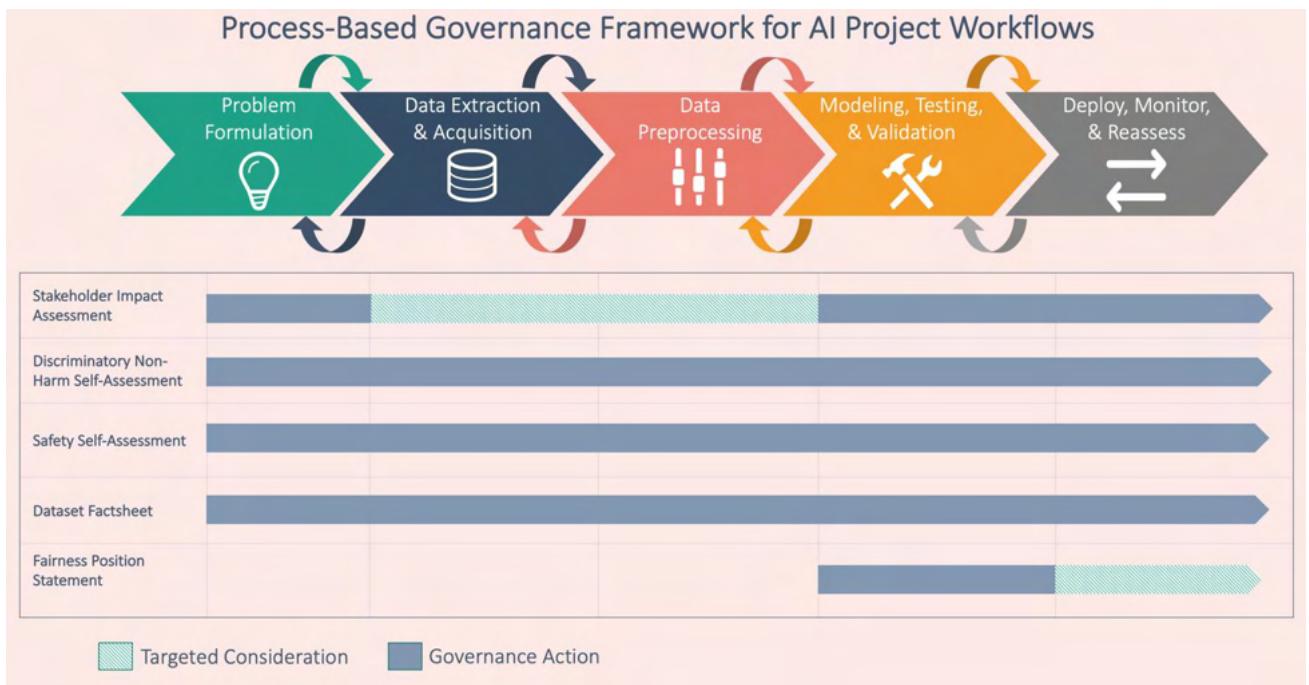
2. **Process-Based Governance Framework:** So far, this guide has presented some of the main steps that are necessary for establishing responsible innovation practices in your AI project. Perhaps the most vital of these measures is the effective operationalisation of the values and principles that underpin the development of ethical and safe AI. By organising all of your governance considerations and actions into a PBG Framework, you will be better able to accomplish this task.

The purpose of a PBG Framework is to provide a template for the integrations of the norms, values, and principles, which motivate and steer responsible innovation, with the actual processes that characterise the AI design and development pipeline. While the accompanying Guide has focused primarily on the Cross Industry Standard Process for Data Mining (CRISP-DM), keep in mind that such a structured integration of values and principles with innovation processes is just as applicable in other related workflow models like Knowledge Discovery in Databases (KDD) and Sample, Explore, Modify, Model, and Assess (SEMMA).

Your PBG Framework should give you a landscape view of the governance procedures and protocols that are organising the control structures of your project workflow. Constructing a good PBG Framework will provide you and your team with a big picture of:

- The relevant team members and roles involved in each governance action.
- The relevant stages of the workflow in which intervention and targeted consideration are necessary to meet governance goals
- Explicit timeframes for any necessary follow-up actions, re-assessments, and continual monitoring
- Clear and well-defined protocols for logging activity and for instituting mechanisms to assure end-to-end auditability

To help you get a summary picture of where the components of process transparency explored so far fit into a PBG Framework, here is a landscape view:



3. **Enabling Auditability with a Process Log:** With your controls in place and your governance framework organised, you will be better able to manage and consolidate the information necessary to assure end-to-end auditability. This information should include both the records and activity-monitoring results that are yielded by your PBG Framework and the model development data gathered across the modelling, training, testing, verifying, and implementation phases.

By centralising your information digitally in a process log, you are preparing the way for optimal process transparency. A process log will enable you to make available, in one place, information that may assist you in demonstrating to concerned parties and affected decision subjects both the responsibility of design and use practices and the justifiability of the outcomes of your system's processing behaviour.

Such a log will also allow you to differentially organise the accessibility and presentation of the information yielded by your project. Not only is this crucial to preserving and protecting data that legitimately should remain unavailable for public view, it will afford your team the capacity to cater the presentation of results to different tiers of stakeholders with different interests and levels of expertise. This ability to curate your explanations with the user-receiver in mind will be vital to achieving the goals of interpretable and justifiable AI.

## Outcome transparency: Explaining outcome and clarifying content

Beyond enabling process transparency through your PBG Framework, you must also put in place standards and protocols to ensure that clear and understandable explanations of the outcomes of your AI system's decisions, behaviours, and problem-solving tasks can:

1. Properly inform the evidence-based judgments of the implementers that they are designed to support;
2. Be offered to affected stakeholders and concerned parties in an accessible way.

This is a multifaceted undertaking that will demand careful forethought and participation across your entire project team.

There is no simple technological solution for how to effectively clarify and convey the rationale behind a model's output in a particular decision-making or behavioural context. Your team will have to use sound judgement and common sense in order to bring together the **technical aspects** of choosing, designing, using a sufficiently interpretable AI system and the **delivery aspects** of being able to clarify and communicate in plain, non-technical, and socially meaningful language how and why that system performed the way it did in a specific decision-making or behavioural context.

Having a good grasp of the rationale and criteria behind the decision-making and problem-solving behaviour of your system is essential for producing safe, fair, and ethical AI. If your AI model is not sufficiently interpretable—if you aren't able to draw from it humanly understandable explanations of the factors that played a significant role in determining its behaviours—then you may not be able to tell how and why things go wrong in your system when they do.

This is a crucial and unavoidable issue for reasons we have already explored. Ensuring the safety of high impact systems in transportation, medicine, infrastructure, and security requires human verification that these systems have properly learned the critical tasks they are charged to complete. It also requires confirmation that when confronted with unfamiliar circumstances, anomalies, and perturbations, these systems will not fail or make unintuitive errors. Moreover, ensuring that these systems operate without causing discriminatory harms requires effective ways to detect and to mitigate sources of bias and inequitable influence that may be buried deep within their feature spaces, inferences, and architectures. Without interpretability each one of these tasks necessary for delivering safe and morally justifiable AI will remain incomplete.

### **Defining Interpretable AI**

To gain a foothold in both the technical and delivery dimensions of AI interpretability, you will first need a solid working definition of what interpretable AI is. To this end, it may be useful to recall once again the definition of AI offered in the accompanying Guide: '*Artificial Intelligence is the science of making computers do things that require intelligence when done by humans.*'

This characterisation is important, because it brings out an essential feature of the explanatory demands of interpretable AI: to do things that require intelligence when done by humans means to do things that require *reasoning processes and cognitive functioning*. This cognitive dimension has a direct bearing on how you should think about offering suitable explanations about algorithmically generated outcomes:

**Explaining an algorithmic model's decision or behaviour should involve making explicit how the particular set of factors which determined that outcome can play the role of evidence in supporting**

the conclusion reached. It should involve making intelligible to affected individuals the rationale behind that decision or behaviour as if it had been produced by a reasoning, evidence-using, and inference-making person.

What makes this explanation-giving task so demanding when it comes to AI systems is that reasoning processes do not occur, for humans, at just one level. Rather, human-scale reasoning and interpreting includes:

1. Aspects of **logic** (applying the basic principles of validity that lie behind and give form to sound thinking): *This aspect aligns with the need for formal or logical explanations of AI systems.*
2. Aspects of **semantics** (gaining an understanding of how and why things work the way they do and what they mean): *This aspect aligns with the need for explanations of the technical rationale behind the outcomes AI systems.*
3. Aspects of the **social understanding of practices, beliefs, and intentions** (clarifying the content of interpersonal relations, societal norms, and individual objectives): *This aspect aligns with the need for the clarification of the socially meaningful content of the outcomes of AI systems.*
4. Aspects of **moral justification** (making sense of what should be considered right and wrong in our everyday activities and choices): *This aspect aligns with the justifiability of AI systems.*

There are good reasons why ***all four of these dimensions of human reasoning processes*** must factor in to explaining the decisions and behaviours of AI systems: First and most evidently, understanding the logic and technical innerworkings (i.e. semantic content) of these systems is a precondition for ensuring their safety and fairness. Secondly, because they are designed and used to achieve human objectives and to fulfil surrogate cognitive functions *in the everyday social world*, we need to make sense of these systems in terms of the consequential roles that their decisions and behaviours play in that human reality. The social context of these outcomes matters greatly. Finally, because they actually affect individuals and society in direct and morally consequential ways, we need to be able to understand and explain their outcomes not just in terms of their mathematical logic, technical rationale, and social context but also in terms of the justifiability of their impacts on people.

Delving more deeply into the technical and delivery aspects of interpretable AI will show how these four dimensions of human reasoning directly line up with the different levels of demand for explanations of the outcomes of AI systems. In particular, the logical and semantic dimensions will weigh heavily in technical considerations whereas the social and moral dimensions will be significant at the point of delivery.

Note here, though, that these different dimensions of human reasoning are not necessarily mutually exclusive but build off and depend upon each other in significant and cascading ways. Approaching explanations of interpretable AI should therefore be treated holistically and inclusively. Technical explanation of the logic and rationale of a given model, for instance, should be seen as a support for the context-based clarification of its socially meaningful content, just as that socially meaningful content should be viewed as forming the basis of explaining an outcome's moral justifiability. When

considering how to make the outcomes of decision-making and problem-solving AI systems maximally transparent to affected stakeholders, you should take this rounded view of human reasoning into account, because it will help you address more effectively the spectrum of concerns that these stakeholders may have.

### Technical aspects of choosing, designing, and using an interpretable AI system

Keep in mind that, while, on the face of it, the task of choosing between the numerous AI and machine learning algorithms may seem daunting, it need not be so. By sticking to the priority of outcome transparency, you and your team will be able to follow some straightforward and simple guidelines for selecting sufficiently interpretable but optimally performing algorithmic techniques.

Before exploring these guidelines, it is necessary to provide you with some background information to help you better understand what facets of explanation are actually involved in technically interpretable AI. A good grasp of what is actually needed from such an explanation will enable you to effectively target the interpretability needs of your AI project.

**Facets of explanation in technically interpretable AI:** A good starting point for understanding how the technical dimension of explanation works in interpretable AI systems is to remember that these systems are largely mathematical models that carry out step-by-step computations in transforming sets of statistically interacting or independent inputs into sets of target outputs. Machine learning is, at bottom, just applied statistics and probability theory fortified with several other mathematical techniques. As such, it is subject to same methodologically rigorous requirements of logical validation as other mathematical sciences.

Such a demand for rigour informs the facet of **formal and logical explanation of AI systems** that is sometimes called the **mathematical glass box**. This characterisation refers to the transparency of strictly formal explanation: No matter how complicated it is (even in the case of a deep neural net with a hundred million parameters), an algorithmic model is a closed system of effectively computable operations where rules and transformations are mechanically applied to inputs to determine outputs. In this restricted sense, all AI and machine learning models are fully intelligible and mathematically transparent if only **formally and logically** so.

This is an important characteristic of AI systems, because it makes it possible for supplemental and eminently interpretable computational approaches to model, approximate, and simplify even the most complex and high dimensional among them. In fact, such a possibility fuels some of the technical approaches to interpretable AI that will soon be explored.

This formal way of understanding the technical explanation of AI and machine learning systems, however, has immediate limitations. It can tell us that a model is mathematically intelligible because it operates according to a collection of fixed operations and parameters, but it cannot tell us much about how or why the components of the model transformed a specified group of inputs into their corresponding outputs. It cannot tell us anything about the *rationale behind the algorithmic generation of a given outcome*.

This second dimension of technical explanation has to do with the *semantic facet* of interpretable AI. A **semantic explanation** offers an interpretation of the functions of the individual parts of the

algorithmic system in the generation of its output. Whereas formal and logical explanation presents an account of the stepwise application of the procedures and rules that comprise the formal framework of the algorithmic system, semantic explanation helps us to understand the meaning of those procedures and rules in terms of their purpose in the input-output mapping operation of the system, i.e. what role they play in determining the outcome of the model's computation.

The difficulties surrounding the interpretability of algorithmic decisions and behaviours arise in this semantic dimension of technical explanation. It is easiest to illustrate this by starting from the simplest case.

When a machine learning model is very basic, the task of following the rationale of how it transforms a given set of inputs into a given set of outputs can be relatively unproblematic. For instance, in the simple linear regression,  $y = a + bx + \varepsilon$ , with a single predictor variable  $x$  and a response variable  $y$ , the predictive relationship of  $x$  to  $y$  is directly expressed in a regression coefficient  $b$ , representing the rate and direction at which  $y$  is predicted to change as  $x$  changes. This hypothetical model is completely interpretable from the technical perspective for the following reasons:

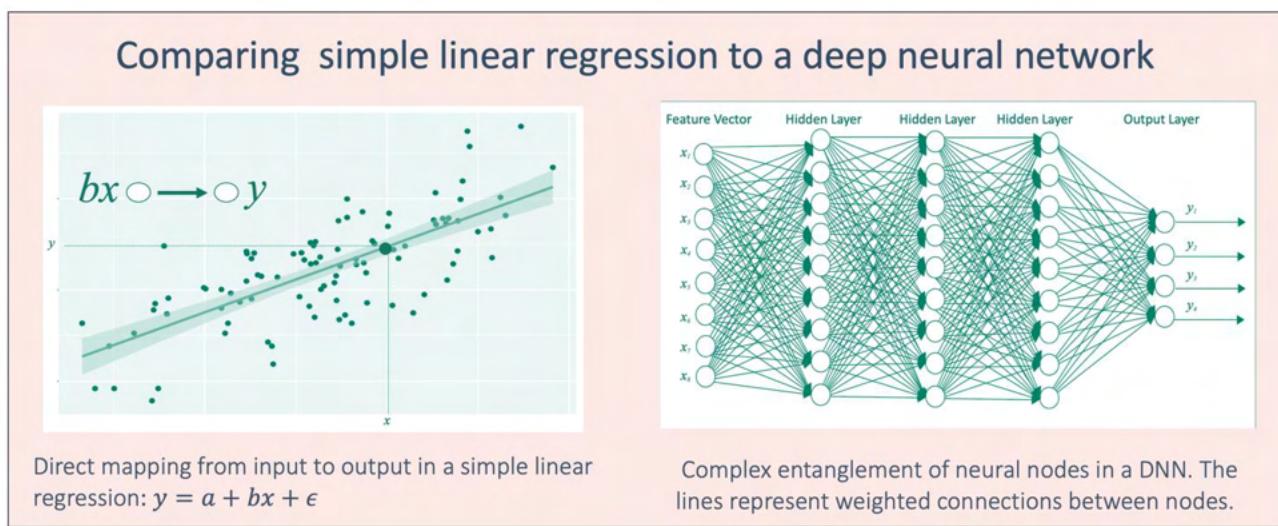
- **Linearity:** Any change in the value of the predictor variable is directly reflected in a change in the value of the response variable at a constant rate  $b$ . The interpretable prediction yielded by the model can therefore be directly inferred. This linearity dimension of predictive models has been an essential feature of the automated decision-making systems in many heavily regulated and high-impact sectors, because the predictions yielded have high inferential clarity and strength.
- **Monotonicity:** When the value of the predictor changes in a given direction, the value of the response variable changes consistently either in the same or opposite direction. The interpretable prediction yielded by the model can thus be directly inferred. This monotonicity dimension is also a highly desirable interpretability condition of predictive models in many heavily regulated sectors, because it incorporates reasonable expectations about the consistent application of sector specific selection constraints into automated decision-making systems. So, for example, if the selection criteria to gain employment at an agency or firm includes taking an exam, a reasonable expectation of outcomes would be that if candidate A scored better than candidate B, then candidate B, all other things being equal, would not be selected for employment when A is not. A monotonic predictive model that uses the exam score as the predictor variable and application success as the response variable would, in effect, guarantee this expectation is met by disallowing situations where A scores better than B but B gets selected and A does not.
- **Non-Complexity:** The number of features (dimensionality) and feature interactions is low enough and the mapping function is simple enough to enable a clear 'global' understanding of the function of each part of the model in relation to its outcome.

While, all three of these desirable interpretability characteristics of the imagined model allow for direct and intuitive reasoning about the relation of the predictor and response variables, the model itself is clearly too minimal to capture the density of relationships and interactions between attributes in complex real-world situations where some degree of noisiness is unavoidable and the task of apprehending the subtleties of underlying data distributions is tricky.

In fact, one of the great strides forward that has been enabled by the contemporary convergence of expanding computing power and big data availability with more advanced machine learning models has been exactly this capacity to better capture and model the intricate and complicated dynamics of real-world situations. Still, this incorporation of the complexity of scale into the models themselves has also meant significant challenges to the semantic dimension of the technical explanation of AI systems.

As machine learning systems have come to possess both ever greater access to big data and increasing computing power, their designers have correspondingly been able both to enlarge the feature spaces (the number of input variables) of these systems and to turn to gradually more complex mapping functions. In many cases, this has meant vast improvements in the predictive and classificatory performance of more accurate and expressive models, but this has also meant the growing prevalence of non-linearity, non-monotonicity, and high-dimensional complexity in an expanding array of so-called ‘black-box’ models.

Once high-dimensional feature spaces and complex functions are introduced into machine learning systems, the effects of changes in any given input become so entangled with the values and interactions of other inputs that understanding how individual components are transformed into outputs becomes extremely difficult. The complex and unintuitive curves of the decision functions of many of these models preclude linear and monotonic relations between their inputs and outputs. Likewise, the high-dimensionality of their optimisation techniques—frequently involving millions of parameters and complex correlations—ranges well beyond the limits of human-scale cognition and understanding. To illustrate the increasing complexity involved in comprehending input-output mappings, here is a visual representation that depicts the difference of between a linear regression function and a deep neural network:



These rising tides of computational complexity and algorithmic opacity consequently pose a key challenge for the responsible design and deployment of safe, fair, and ethical AI systems: how should the potential to advance the public interest through the implementation of high performing but increasingly uninterpretable machine learning models be weighed against the tangible risks posed by the lack of interpretability of such systems?

A careful answer to this question is, in fact, not so simple. While the trade-off between performance and interpretability may be real and important in *some domain-specific applications*, in others there exist increasingly sophisticated developments of standard interpretable techniques such as regression extensions, decision trees, and rule lists that may prove just as effective for use cases where the need for transparency is paramount. Furthermore, supplemental interpretability tools, which function to make ‘black box’ models more semantically and qualitatively explainable are rapidly advancing day by day.

These are all factors that you and your team should consider as you work together to decide on which models to use for your AI project. As a starting point for those considerations, let us now turn to some basic guidelines that may help you to steer that dialogue toward points of relevance and concern.

### Guidelines for designing and delivering a sufficiently interpretable AI system

You should use the table below to begin thinking about how to integrate interpretability into your AI project. While aspects of this topic can become extremely technical, it is important to make sure that dialogue about making your AI system interpretable remains multidisciplinary and inclusive. Moreover, it is crucial that key stakeholders be given adequate consideration when deciding upon the delivery mechanisms of your project. These should include policy or operational design leads, the technical personnel in charge of operating the trained models, the implementers of the models, and the decision subjects, who are affected by their outcomes.

Note that the first three guidelines focus on the big picture issues you will need to consider in order to incorporate interpretability needs into your project planning and workflow, whereas the last two guidelines shift focus to the user-centred requirements of designing and implementing a sufficiently interpretable AI system.

#### Guidelines for designing and delivering a sufficiently interpretable AI system

##### **Guideline 1: Look first to context, potential impact, and domain-specific need when determining the interpretability requirements of your project**

There are several related factors that should be taken into account as you formulate your project’s approach to interpretability:

- 1. Type of application:** Start by assessing both the kind of tool you are building and the environment in which it will apply. Clearly there is a big difference between a computer vision system that sorts handwritten employee feedback forms and one that sorts safety risks at a security checkpoint. Likewise, there is a big difference between a random forest model that triages applicants at a licencing agency and one that triages sick patients in an emergency department.

Understanding your AI system’s purpose and context of application will give you a better idea of the stakes involved in its use and hence also a good starting point to think about the scope of its interpretability needs. For instance, low-stakes AI models that are

not safety-critical, do not directly impact the lives of people, and do not process potentially sensitive social and demographic data will likely have a lower need for extensive resources to be dedicated to a comprehensive interpretability platform.

2. **Domain specificity:** By acquiring solid domain knowledge of the environment in which your AI system will operate, you will gain better insight into any potential sector-specific standards of explanation or benchmarks of justification which should inform your approach to interpretability. Through such knowledge, you may also obtain useful information about organisational and public expectations regarding the scope, content, and depth of explanations that have been previously offered in relevant use cases.
3. **Existing technology:** If one of the purposes of your AI project is to replace an existing algorithmic technology that may not offer the same sort of expressive power or performance level as the more advanced machine learning techniques that you are planning to deploy, you should carry out an assessment of the performance and interpretability levels of the existing technology. Acquiring this knowledge will provide you with an important reference point when you are considering possible trade-offs between performance and interpretability that may occur in your own prospective system. It will also allow you to weigh the costs and benefits of building a more complex system with higher interpretability-support needs in comparison to the costs and benefits of using a simpler model.

## Guideline 2: Draw on standard interpretable techniques when possible

In order to actively integrate the aim of sufficient interpretability into your AI project, your team should approach the model selection and development process with the goal of finding the right fit between **(1) domain-specific risks and needs, (2) available data resources and domain knowledge, and (3) task appropriate machine learning techniques**. Effectively assimilating these three aspects of your use case requires open-mindedness and practicality.

Often times, it may be the case that high-impact, safety-critical, or other potentially sensitive environments heighten demands for the thoroughgoing accountability and transparency of AI projects. In some of these instances, such demands may make choosing standard but sophisticated non-opaque techniques an overriding priority. These techniques may include **decisions trees, linear regression and its extensions like generalised additive models, decision/rule lists, case-based reasoning, or logistic regression**. In many cases, reaching for the ‘black box’ model first may not be appropriate and may even lead to inefficiencies in project development, because more interpretable models, which perform very well but do not require supplemental tools and techniques for facilitating interpretable outcomes, are also available.

Again, solid domain knowledge and context awareness are key components here. In use cases where data resources lend to well-structured, meaningful representations and domain expertise can be incorporated into model architectures, interpretable techniques may often be more desirable than opaque ones. Careful data pre-processing and iterative model development can, in these cases, hone the accuracy of such interpretable systems in ways that may make the advantages gained by the combination of their performance and transparency outweigh the benefits of more semantically intransparent approaches.

In other use cases, however, data processing needs may disqualify the deployment of these sorts of straightforward interpretable systems. For instance, when AI applications are sought for classifying images, recognising speech, or detecting anomalies in video footage, the most effective machine learning approaches will likely be opaque. The feature spaces of these kinds of AI systems grow exponentially to hundreds of thousands or even millions of dimensions. At this scale of complexity, conventional methods of interpretation no longer apply. Indeed, it is the unavoidability of hitting such an **interpretability wall** for certain important applications of supervised, unsupervised, and reinforcement learning that has given rise to an entire subfield of machine learning research which focuses on providing technical tools to facilitate interpretable and explainable AI.

When the use of ‘black box’ models best fits the purpose of your AI project, you should proceed diligently and follow the procedures recommended in Guideline 3. For clarity, let us define a ‘black box’ model as **any AI system whose innerworkings and rationale are opaque or inaccessible to human understanding**. These systems may include **neural networks** (including recurrent, convolutional, and deep neural nets), **ensemble methods** (an algorithmic technique such as the random forest method that strengthens an overall prediction by combining and aggregating the results of several or many different base models), and **support vector machines** (a classifier that uses a special type of mapping function to build a divider between two sets of features in a high dimensional feature space).

**Guideline 3: When considering the use of ‘black box’ AI systems, you should:**

1. Thoroughly weigh up impacts and risks;
2. Consider the options available for supplemental interpretability tools that will ensure a level of semantic explanation which is both *domain appropriate* and *consistent with the design and implementation of safe, fair, and ethical AI*;
3. Formulate an interpretability action plan, so that you and your team can put adequate forethought into how explanations of the outcomes of your system’s decisions, behaviours, or problem-solving tasks can be optimally provided to users, decision subjects, and other affected parties.

It may be helpful to explore each of these three suggested steps of assessing the viability of the responsible design and implementation of a ‘black box’ model in greater detail.

**(1) Thoroughly weigh up impacts and risks:** Your first step in evaluating the feasibility of using a complex AI system should be to focus on issues of ethics and safety. As a general policy, you and your team should utilise ‘black box’ models only:

- where their potential impacts and risks have been thoroughly considered in advance, and you and your team have determined that your use case and domain specific needs support the responsible design and implementations of these systems;

- where supplemental interpretability tools provide your system with a domain appropriate level of semantic explainability that is reasonably sufficient to mitigate its potential risks and that is therefore consistent with the design and implementation of safe, fair, and ethical AI.

**(2) Consider the options available for supplemental interpretability tools:** Next, you and your team should assess whether there are technical methods of explanation-support that **both** satisfy the specific interpretability needs of your use case as determined by the deliberations suggested in Guideline 1 **and** are appropriate for the algorithmic approach you intend to use. You should consult closely with your technical team at this stage of model selection. The exploratory processes of trial-and-error, which often guide this discovery phase in the innovation lifecycle, should be informed and constrained by a solid working knowledge of the technical art of the possible in the domain of available and useable interpretability approaches.

The task of lining up the model selection process with the demands of interpretable AI requires a few conceptual tools that will enable thoughtful evaluation of whether proposed supplemental interpretability approaches sufficiently meet your project's explanatory needs. First and most importantly, you should be prepared to ask the right questions when evaluating any given interpretability approach. This involves establishing with as much clarity as possible **how the explanatory results of that approach can contribute to the user's ability to offer solid, coherent, and reasonable accounts of the rationale behind any given algorithmically generated output**. Relevant questions to ask that can serve this end are:

- What sort of explanatory resources will the interpretability tool provide users and implementers in order (1) to enable them to exercise better-informed evidence-based judgments and (2) to assist them in offering plausible, sound, and reasonable accounts of the logic behind algorithmically generated output to affected individuals and concerned parties?
- Will the explanatory resources that the interpretability tool offers be useful for providing affected stakeholders with a sufficient understanding of a given outcome?
- How, if at all, might the explanatory resources offered by the tool be misleading or confusing?

You and your team should take these questions as a starting point for evaluating prospective interpretability tools. These tools should be assessed in terms of their capacities to render the reasoning behind the decisions and behaviours of the uninterpretable 'black box' systems sufficiently intelligible to users and affected stakeholders given use case and domain specific interpretability needs.

Keeping this in mind, there are two technical dimensions of supplemental interpretability approaches that should be systematically incorporated into evaluation processes at this stage of the innovation workflow.

The first involves the possible **explanatory strategies** you choose to pursue over the course of the design and implementation lifecycle. Such strategies will largely determine the paths to understanding you will be able to provide for its users and decision subjects. They will largely define *how you explain your model and its outcomes* and hence *what kinds of explanation you are able offer*.

The second involves the **coverage and scope** of the actual explanations themselves. The choices you make about explanatory coverage will determine the extent to which the kinds of explanations you are planning to pursue will address *single instances* of the model's outputs or range more broadly to cover the *underlying rationale of its behaviour in general and across instances*. Choices you make about explanatory coverage will largely govern the extent to which your AI system is locally and/or globally interpretable.

The very broad-brushed overview of these two dimensions that follows is just meant to orient you to some of the basic concepts in an expanding field of research, so that you are more prepared for working with your technical team to think through the strengths and weaknesses of various approaches. Note, additionally, that this is a rapidly developing area. Relevant members of your team should keep abreast of the latest developments in the field of interpretable AI or XAI (Explainable AI):

#### Two technical dimensions of supplemental interpretability approaches:

1. **Determining explanatory strategies:** To achieve the goal of securing a sufficiently interpretable AI system, you and your team will need to get clear on **how to explain** your model and its outcomes. The explanatory strategies you decide to pursue will shape the paths to understanding you are able to provide for the users of your model and for its decision subjects.

There are four such explanatory strategies to which you should pay special attention:

- a) **Internal explanation:** Pursuing the internal explanation of an opaque model involves making intelligible how the components and relationships within it function. There are two ways that such a goal of internal explanation can be interpreted. On the one hand, it can be seen as an endeavour to explain the operation of the model by considering it globally *as a comprehensible whole*. Here, the aspiration is to 'pry open the black box' by building an explanatory model that enables a full grasp of the opaque system's internal contents. The strengths and weaknesses of such an approach will be discussed in the next section on global interpretability.

On the other hand, the search for internal explanation can indicate the pursuit of a kind of **engineering insight**. In this sense, internal explanation can be seen as attempting to shed descriptive and inferential light on the parts and operation of the system as a whole in order to try to make it work better. Acquiring this sort of internal understanding of the more general relationships that the working parts of a trained model have with patterns of its responses can allow researchers to advance step-by-step in gaining a better data scientific grasp on

why it does what it does and how to improve it. Similarly, this type of internal explanation can be seen as attempting to shed light on an opaque model's operation by breaking it down into more understandable, analysable, and digestible parts (for instance, in the case of a DNN: into interpretable characteristics of its vectors, features, layers, parameters, etc.).

From a practical point of view, this kind of aspiration to *engineering insight* in the ends of data scientific advancement should inform the goals of your technical team throughout the model selection and design workflow.

Numerous methods exist to help provide informative representations of the innerworkings of various 'black box' systems. Gaining a clearer descriptive understanding of the internal composition of your system will contribute greatly to your project's ability to achieve a higher degree of outcome transparency and to its capacity to foster best practices in the pursuit of responsible data science in general.

- b) ***External or post-hoc explanation:*** External or post-hoc explanation attempts to capture essential attributes of the observable behaviour of a 'black box' system by subjecting it to a number of different techniques that reverse engineer explanatory insight. Some post-hoc approaches test the sensitivity of the outputs of an opaque model to perturbations in its inputs; others allow for the interactive probing of its behavioural characteristics; others, still, build proxy-based models that utilise simplified interpretable techniques to gain a better understanding of particular instances of its predictions and classifications.

This external or post-hoc approach has, at present, established itself in machine learning research as a go-to explanatory strategy and for good reason. It allows data scientists to pose mathematical questions to their opaque systems by testing them and by building supplemental models which enable greater insight through the inferences drawn from their experimental interventions. Such a post-hoc approach allows them, moreover, to seek out evidence for the reasoning behind a given opaque model's prediction or classification by utilising maximally interpretable techniques like linear regression, decision trees, rule lists, or case-based reasoning. Several examples of post-hoc explanation will be explored below in the section on local interpretability.

Take note initially though that, as some critics have rightly pointed out, because they are approximations or simplified supplemental models of the more complex originals, many post-hoc explanations can fail to accurately represent certain areas of the opaque model's feature space. This deterioration of accuracy in parts of the original model's domain can frequently produce misleading and uncertain results in the post-hoc explanations of concern.

- c) ***Supplemental explanatory infrastructure:*** A different kind of explanatory strategy involves actually incorporating secondary explanatory facilities into the system you are building. For instance, an image recognition system could have a primary component, like a convolutional neural net, that extracts features from

its inputs and classifies them while a secondary component, like a built-in recurrent neural net with an ‘attention-directing’ mechanism, translates the extracted features into a natural language representation that produces a sentence-long explanation of the result to the user. In other words, a system like this is designed to provide simple explanations of its own data processing results.

Research into integrating ‘attention-based’ interfaces like this in AI systems is continuing to advance toward making their implementations more sensitive to user needs, more explanation-forward, and more human-understandable. For instance, multimodal methods of combining visualisation tools and textual interface are being developed that may make the provision of explanations more interpretable for both implementers and decision subjects. Furthermore, the incorporation of domain knowledge and logic-based or convention-based structures into the architectures of complex models are increasingly allowing for better and more user-friendly representations and prototypes to be built into them. This is gradually enabling more sophisticated explanatory infrastructures to be integrated into opaque systems and makes it essential to think about building explanation-by-design into your AI projects.

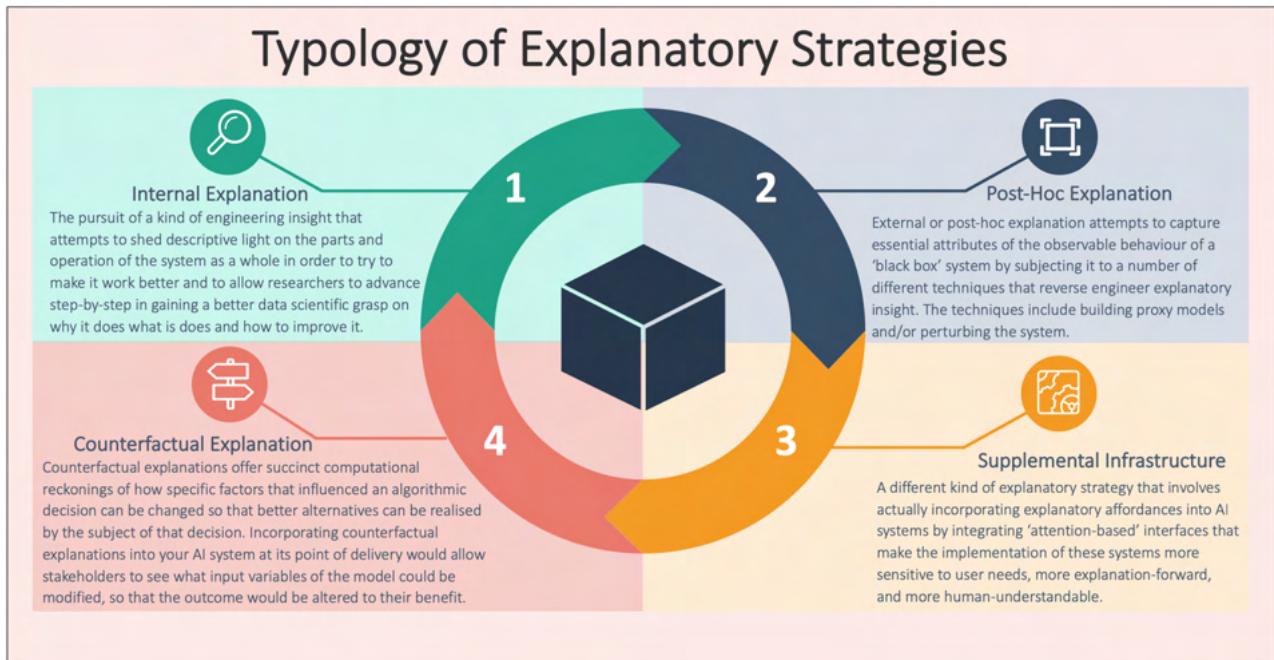
- d) ***Counterfactual explanation:*** While counterfactual explanation is a kind of post-hoc approach, it deserves special attention insofar as it moves beyond other post-hoc explanations to provide affected stakeholders with clear and precise options for actionable recourse and practical remedy.

Counterfactual explanations are contrastive explanations: They offer succinct computational reckonings of how specific factors that influenced an algorithmic decision can be changed so that better alternatives can be realised by the subject of that decision. Incorporating counterfactual explanations into your AI system at its point of delivery would allow stakeholders to see what input variables of the model can be modified, so that the outcome could be altered to their benefit. Additionally, from a responsible design perspective, incorporating counterfactual explanation into the development and testing phases of your system would allow your team to build a model that incorporates ***actionable variables***, i.e. input variables that will afford decision subjects with concise options for making practical changes that would improve their chances of obtaining the desired outcome. **Counterfactual explanatory strategies can be used as way to incorporate reasonableness and the encouragement of agency into the design and implementation of your AI project.**

All that said, it is important to recognise that, while counterfactual explanation does offer an innovative way to contrastively explore how feature importance may influence an outcome, it is not a complete solution to the problem of AI interpretability. In certain cases, for instance, the sheer number of potentially significant features that could be at play in counterfactual explanations of a given result can make a clear and direct explanation difficult to obtain and selected sets of explanations seem potentially arbitrary. Moreover, there are as

yet limitations on the types of datasets and functions to which these kinds of explanations are applicable. Finally, because this kind of explanation concedes the opacity of the algorithmic model outright, it is less able to address concerns about potentially harmful feature interactions and multivariate relationships that may be buried deep within the model's architecture.

Here is an at-a-glance view of a typology of these explanatory strategies:



2. **Coverage and Scope:** The main questions you will need to broach in the dimension of the coverage and scope of your supplemental interpretability approach are: To what extent does our interpretability approach cover the explanation of *singe predictions or classifications* of the model and to what extent does it cover the explanation of the *innerworkings and rationale of the model as a whole and across predictions*? To what extent does it cover both?

This distinction between single instance and total model explanation is often characterised as the difference between **local interpretability** and the **global interpretability**. Both types of explanation offer potentially helpful support for the provision of significant information about the rationale behind an algorithmic decision or behaviour, but both, in their own ways, also face difficulties.

**Local Interpretability:** A local semantic explanation aims to enable the interpretability of **individual cases**. The general idea behind attempts to explain a 'black box' system in terms of specific instances is that, regardless of how complex the architecture or decision function of that system may be, it is possible to gain interpretive insight into its innerworkings by focusing on single data points or neighbourhoods in its feature space. In other words, even if the high dimensionality and curviness of a model makes it opaque *as a whole*, there is an expectation that insight-generating

interpretable methods can be applied *locally* to smaller sections of the model, where changes in isolated or grouped variables are more manageable and understandable.

This general explanatory perspective has yielded several different interpretive strategies that have been successfully applied in significant areas of ‘black box’ machine learning. One family of such strategies has zeroed in on neural networks (DNNs, in particular) by identifying what features of an input vector’s data points make it representative of the target concept that a given model is trying to classify. So, for example, if we have a digital image of a dog that is converted into a vector of pixel values and then processed it through a dog-classifying deep neural net, this interpretive approach will endeavour to tell us why the system yielded a ‘dog-positive’ output by isolating the slices of this set of data points that are most relevant to its successful classification by the model.

This can be accomplished in several related ways. What is called **sensitivity analysis** identifies the most relevant features of an input vector by calculating local gradients to determine how a data point has to be moved to change the output label. Here, an output’s sensitivity to such changes in input values identifies the most relevant features. Another method to identify feature relevance that is downstream from sensitivity analysis is called **salience mapping**, where a strategy of moving backward through the layers of a neural net graph allows for the mapping of patterns of high activation in the nodes and ultimately generates interpretable groupings of salient input variables that can be visually represented in a heat or pixel attribution map.

A second local interpretive strategy also seeks to explain feature importance in a single prediction or classification by perturbing input variables. However, instead of using these nudges in the feature space to highlight areas of saliency, it uses them to prod the opaque model in the area around the relevant prediction, so that a supplemental interpretable model can be constructed which establishes the relative importance of features in the black box model’s output.

The most well-known example of this strategy is called **LIME (Local Interpretable Model-Agnostic Explanation)**. LIME works by fitting an interpretable model to a specific prediction or classification produced by the opaque system of concern. It does this by sampling data points at random around the target prediction or classification and then using them to build a local approximation of the decision boundary that can account for the features which figure prominently in the specific prediction or classification under scrutiny.

The way this works is relatively uncomplicated: LIME generates a simple linear regression model by weighting the values of the data points, which were produced by randomly perturbing the opaque model, according to their proximity to the original prediction or classification. The closest of these values to the instance being explained are weighted the heaviest, so that the supplemental model can produce an explanation of feature importance that is **locally faithful** to that instance. Note that the type of model that LIME uses most prominently is a sparse linear regression

function for reasons of semantic transparency that were discussed above. Other interpretable models such as decision trees can likewise be employed.

While LIME does indeed appear to be a step in the right direction for the future of interpretable AI, a host of issues that present challenges to the approach remains unresolved. For instance, the crucial aspect of how to properly define the proximity measure for the ‘neighbourhood’ or ‘local region’ where the explanation applies remains unclear, and small changes in the scale of the chosen measure can lead to greatly diverging explanations. Likewise, the explanation produced by the supplemental linear model can quickly become unreliable even with small and virtually unnoticeable perturbations of the system it is attempting to approximate. This challenges the basic assumption that there is always some simplified linear model that successfully approximates the underlying model reasonably well near any given data point.

LIME’s creators have largely acknowledged these shortcomings and have recently offered a new explanatory approach that they call ‘**anchors**’. These ‘high precision rules’ incorporate into their formal structures ‘reasonable patterns’ that are operating within the underlying model (such as the implicit linguistic conventions that are at work in a sentiment prediction model), so that they can establish suitable and faithful boundaries of their explanatory coverage of its predictions or classifications.

A related and equally significant local interpretive strategy is called **SHAP (Shapley Additive exPlanations)**. SHAP uses concepts from game theory to define a ‘Shapley value’ for a feature of concern that provides a measurement of its influence on the underlying model’s prediction. Broadly, this value is calculated for a feature by averaging its marginal contribution to *every possible prediction* for the instance under consideration.

This might seem impossible, but the strategy is straightforward. SHAP calculates the marginal contribution of the relevant feature for all possible combinations of inputs in the feature space of the instance. So, if the opaque model that it is explaining has 15 features, SHAP would calculate the marginal contribution of the feature under consideration 32,768 times (i.e. one calculation for each combination of all possible combinations of features:  $2^{15}$ , or  $2^k$  when  $k = 15$ ).

This method then allows SHAP to estimate the Shapley values for all input features in the set to produce the complete distribution of the prediction for the instance. In our example, this would entail 491,520 calculations. While such a procedure is computationally burdensome and becomes intractable beyond a certain threshold, this means that *locally*, that is, for the calculation of the specific instance, SHAP can axiomatically guarantee the consistency and accuracy of its reckoning of the marginal effect of the feature. (Note that the SHAP platform does offer methods of approximation to avoid this excessive computational expense.)

Despite this calculational robustness, SHAP also faces some of the same kinds of difficulties that LIME does. The way SHAP calculates marginal contributions is by

constructing two instances: the first instance includes the feature being measured while the second leaves it out. After calculating the prediction for each of these instances by plugging their values into the underlying model, the result of the second is subtracted from that of the first to determine the marginal contribution of the feature. This procedure is then repeated for all possible combinations of features so that the weighted average of all of the marginal contributions of the feature of concern can be computed.

The contestable part of this process comes with how SHAP defines the *absence* of variables under consideration. To leave out a feature—whether it's the one being directly measured or one of the others not included in the combination under consideration—SHAP replaces it with a *stand-in feature value* drawn from a selected donor sample (that is itself drawn from the existing dataset). This method of sampling values assumes feature independence (i.e. that values sampled are not correlated in ways that might significantly affect the output for a particular calculation). As a consequence, the interaction effects engendered by and between stand-in variables are necessarily unaccounted for when conditional contributions are approximated. The result is the introduction of uncertainty into the explanation that is produced because the complexity of multivariate interactions in the underlying model may not be sufficiently captured by the simplicity of this supplemental interpretability technique. This drawback in sampling (as well as a certain degree of arbitrariness in domain definition) can cause SHAP to become unreliable even with minimal perturbations of the model it is approximating.

Despite these limitations in the existing tools of local interpretability, it is important that you think ‘local-first’ when considering the issue of the coverage and scope of the explanatory approaches you plan to incorporate into your project. Being able to provide explanations of specific predictions and classifications is of paramount importance both to securing optimal outcome transparency and also to ensuring that your AI system will be implemented responsibly and reasonably.

**Global interpretability:** The motivation behind the creation of local interpretability tools like LIME or SHAP (as well as many others not mentioned here) has derived, at least in part, from a need to find a way of avoiding the kind of difficult *double bind* faced by the alternative approach to the coverage and scope of interpretable AI: global interpretability.

On the prevailing view, providing a global explanation of a ‘black box’ model entails offering an alternative interpretable model that captures the innerworkings and logic of a ‘black box’ model *in sum* and across predictions or classifications. The difficulty faced by global interpretability arises in the seemingly unavoidable trade-off between the need for the global explanatory model to be sufficiently simple so that it is understandable by humans and the need for that model to be sufficiently complex so that it can capture the intricacies of how the mapping function of a ‘black box’ model works as a whole.

While this is clearly a real problem that appears to be theoretically inevitable, it is important to keep in mind that, *from a practical standpoint*, a serviceable notion of global interpretability need not be limited to such a conceptual puzzle. There are at least two less ambitious but more constructive ways to view global interpretability as a potentially meaningful contributor to the responsible design and implementation of interpretable AI.

First, many useful attempts have already been made at building explanatory models that employ interpretable methods (like decision trees, rule lists, and case-based classification) to globally approximate neural nets, tree ensembles, and support vector machines. These results have enabled a deeper understanding of the way human interpretable logics and conventions (like if-then rules and representationally generated prototypes) can be measured against or mapped onto high dimensional computational structures and even allow for some degree of targeted comprehensibility of the logic of their parts.

This capacity to ‘peek into the black box’ is of great practical importance in domains where trust, user-confidence, and public acceptance are critical for the realisation optimal outcomes. Moreover, this ability to move back and forth between interpretable architectures and high-dimensional processing structures can enable knowledge discovery as well as insights into the kinds of dataset-level and population-level patterns, which are crucial for well-informed macroscale decision-making in areas ranging from public health and economics to the science of climate change.

Being able to uncover global effects and relationships between complex model behaviour and data distributions at the demographic and ecological level may prove vital for establishing valuable and practically useful knowledge about unobservable but significant biophysical and social configurations. Hence, although these models have not solved the understandability-complexity puzzle as such, they have opened up new pathways for innovative thinking in the applied data sciences that may be of immense public benefit in the future.

Secondly, as mentioned above, under the auspices of the aspiration to **engineering insight**, a *descriptive and analytical kind of global interpretability* can be seen as a driving force of data scientific advancement. When seen through a practitioner-centred lens, this sort of global interpretability allows data scientists to take a wide-angled and discovery-oriented view of a ‘black box’ model’s relationship to patterns that arise across the range of its predictions. Figuring out how an opaque system works and how to make it work better by more fully understanding these patterns is a continuous priority of good research. So too is understanding the relevance of features and of their complex interactions through dataset level measurement and analysis. These dimensions of incorporating the explanatory aspirations of global interpretability into best practices of research and innovation should be encouraged in your AI project.

(3) **Formulate an interpretability action plan:** The final step you will need to take to ensure a responsible approach to using ‘black box’ models is to formulate an interpretability action plan so that you and your team can put adequate forethought into how explanations of the outcomes of your system’s decisions, behaviours, or problem-solving tasks can be optimally provided to users, decision subjects, and other affected parties.

This action plan should include the following:

- A **clear articulation of the explanatory strategies** your team intends to use and a detailed plan that indicates the stages in the project workflow when the design and development of these strategies will need to take place.
- A succinct formulation of your **explanation delivery strategy**, which addresses the special provisions for clear, simple, and user-centred explication that are called for when supplemental interpretability tools for ‘black box’ models are utilised. See more about delivery and implementation in Guideline 5.
- A **detailed timeframe for evaluating your team’s progress** in executing its interpretability action plan and a **role responsibility list**, which maps in detail the various task-specific responsibilities that will need to be fulfilled to execute the plan.

#### Guideline 4: Think about interpretability in terms of the capacities of human understanding

When you begin to deliberate about the specific scope and content of your interpretability platform, it is important to reflect on what it is that you are exactly aiming to do in making your model sufficiently interpretable. A good initial step to take in this process is to think about what makes even the simplest explanations **clear and understandable**. In other words, you should begin by thinking about interpretability in terms of the capacities and limitations of human cognition.

From this perspective, it becomes apparent that even the most straightforward model like a linear regression function or a decision tree can become uninterpretable when its dimensionality presses beyond the cognitive limits of a thinking human. Recall our example of the simple linear regression:  $y = a + bx + \epsilon$ . In this instance, only one feature  $x$  relates to the response variable  $y$ , so understanding the predictive relationship is easy. The model is parsimonious.

However, if we started to add more features as covariates, even though the model would remain linear and hence intuitively predictable, being able to understand the relationship between the response variable and all the predictors and their coefficients (feature weights) would quickly become difficult. So, say we added ten thousand features and trained the model:  $y = a + b_0x_0 + b_1x_1 + \dots + b_{10000}x_{10000} + \epsilon$ . Understanding *how* this model’s prediction comes about—what role each of the individual parts play in producing the prediction—would become difficult because of a certain cognitive limit in the quantity of entities that human thinking can handle at any given time. This model would lose a significant degree of interpretability.

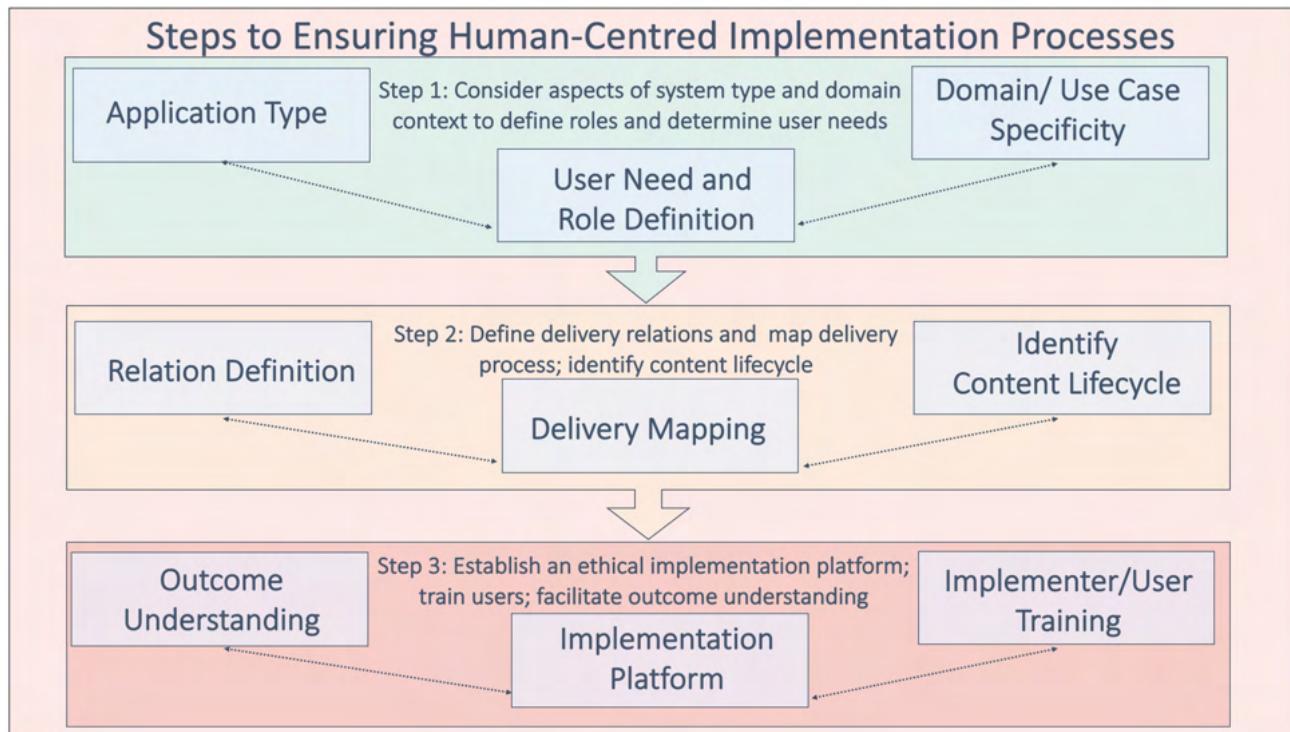
Seeing interpretability as a continuum of comprehensibility that is dependent on the capacities and limits of the individual human interpreter should key you in to what is needed in order to deliver an interpretable AI system. Such limits to consider should include not only cognitive

boundaries but also varying levels of access to relevant vocabularies of explanation; an explanation about the results of a trained model that uses a support vector machine to divide a 26-dimensional feature space with a planar separator, for instance, may be easy to understand for a technical operator or auditor but entirely inaccessible to a non-specialist. Offering good explanations should take expertise level into account. **Your interpretability platform should be cognitively equitable.**

## Securing responsible delivery through human-centred implementation protocols and practices

The demand for sensitivity to human factors should inform your approach to devising delivery and implementation processes from start to finish. To provide clear and effective explanations about the content and rationale of algorithmic outputs, you will have to begin by building ***from the human ground up***. You will have to pay close attention to the circumstances, needs, competences, and capacities of the people whom your AI project aims to assist and serve.

This means that ***context will be critical***. By understanding your use case well and by drawing upon solid domain knowledge, you will be better able to **define roles and relationships**. You will better be able to **train the users and implementers of your system**. And, you will be better able to **establish an effectual implementation platform, to clarify content, and to facilitate understanding of outcomes** for users and affected stakeholders alike. Here is a diagram of what securing human-centred implementation protocols and practices might look like:



Let us consider these steps in turn by building a checklist of essential actions that should be taken to help ensure the human-centred implementation of your AI project. Because the specifics of your approach will depend so heavily on the context and potential impacts of your project, we'll assume a

generic case and construct the checklist around a hypothetical algorithmic decision-making system that will be used for predictive risk assessment.

### Step 1: Consider aspects of application type and domain context to define roles and determine user needs

- (1) Assess which members of the communities you are serving will be most affected by the implementation of your AI system. Who are the most vulnerable among them? How will their socioeconomic, cultural, and education backgrounds affect their capacities to interpret and understand the explanations you intend to provide? How can you fine-tune your explanatory strategy to accommodate their requirements and provide them with clear and non-technical details about the rationale behind the algorithmically supported result?

When thinking about providing explanations to affected stakeholders, you should start with the needs of the most disadvantaged first. Only in this way, will you be able to establish an acceptable baseline for the equitable delivery of interpretable AI.

- (2) After reviewing [Guideline 1](#) above, make a list of and define all the roles that will potentially be involved at the delivery stage of your AI project. As you go through each role, specify levels of technical expertise and domain knowledge as well as possible goals and objectives for each role. For instance, in our predictive risk assessment case:

- **Decision Subject (DS)-**
  - **Role:** Subject of the predictive analytics.
  - **Possible Goals and Objectives:** To receive a fair, unbiased, and reasonable determination, which makes sense; to discover which factors might be changed to receive a different outcome.
  - **Technical and Domain Knowledge:** Most likely low to average technical expertise and average domain knowledge.
- **Advocate for the DS-**
  - **Role:** Support for the DS (for example, legal counsel or care worker) and concerned party to the automated decision.
  - **Possible Goals and Objectives:** To make sure the best interests of the DS are safeguarded throughout the process; to help make clear to the DS what is going on and how and why decisions are being made.
  - **Technical and Domain Knowledge:** Most likely average technical expertise and high level of domain knowledge.
- **Implementer-**
  - **Role:** User of the AI system as decision support.
  - **Possible Goals and Objectives:** To make an objective and fair decision that is sufficiently responsive to the particular circumstances of the DS and that is anchored in solid reasoning and evidence-based judgement.
  - **Technical and Domain Knowledge:** Most likely average technical expertise and high level of domain knowledge.
- **System Operator/Technician-**
  - **Role:** Provider of support and maintenance for the AI system and its use.

- **Possible Goals and Objectives:** To make sure the machine learning system is performing well and running in accordance with its intended design; to handle the technical dimension of information processing for the DS's particular case; to answer technical questions about the system and its results as they arise.
  - **Technical and Domain Knowledge:** Most likely high level of technical expertise and average domain knowledge.
- **Delivery Manager-**
  - **Role:** Member of the implementation team who oversees its operation and responds to problems as they arise.
  - **Possible Goals and Objectives:** To ensure that the quality of the automation-supported assessment process is high and that the needs of the decision subject are being served as intended by the project; to oversee the overall quality of the relationships within the implementation team and between the members of that team and the communities they serve.
  - **Technical and Domain Knowledge:** Most likely average technical expertise and good to high level domain knowledge

## Step 2: Define delivery relations and map delivery processes

- (1) Assess the possible relationships between the defined roles that will have significant bearing on your project's implementation and formulate a descriptive account of this relationship with an eye to the part it will play in the delivery process. For the predictive risk assessment example:
  - **Decision Subject/Advocate to Implementer:** This is the primary relationship of the implementation process. It should be information-driven and dialogue-driven with the implementer's exercise of unbiased judgment and the DS's comprehension of the outcome treated as the highest priorities. Implementers should be prepared to answer questions and to offer evidence-based clarifications and justifications for their determinations. The achievement of well-informed mutual understanding is a central aim.
  - **Implementer to System Operator:** This is the most critical operational relationship within the implementation team. Communication levels should be kept high from case to case, and the shared goal of the two parties should be to optimise the quality of the decisions by optimising the use of the algorithmic decision-support system in ways that are accessible both to the user and to the DS. The conversations between implementers and system operators should be problem-driven and should avoid, as much as possible, focus on the specialised vocabularies of each party's domain of expertise.
  - **Delivery Manager to Operator to Implementer:** The quality of this cross-disciplinary relationship within the implementation team will have direct bearing on the overall quality of the delivery of the algorithmically supported decisions. Safeguarding the latter will require that open and easily accessible lines of communication be maintained between delivery managers, operators, and implementers, so that unforeseen implementation problems can be tackled from multiple angles and in ways that anticipate and stem future difficulties. Additionally, different use cases may present different explanatory challenges that are best addressed by multidisciplinary

team input. Good communications within the implementation team will be essential to enable that such challenges are addressed in a timely and efficient manner.

- (2) Start building a map of the delivery process. This should involve incorporating your understanding of the needs, roles, and relationships of relevant actors involved in the implementation of your AI system into the wider objective of providing clear, informative, and understandable explanations of algorithmically supported decisions.

It is vital to recognise, at this implementation-planning stage of your project, that the principal goal of the delivery process is two-fold: *to translate statistically expressed results into humanly significant reasons and to translate algorithmic outputs into socially meaningful outcomes*.

These overlapping objectives should have a direct bearing on the way you build a map for your project's delivery process, because they organise the duties of implementation into two task-specific components:

1. A **technical component**, which involves determining the most effective way to convey and communicate to users and decision subjects the statistical results of your model's information processing so that the factors that figured into the logic and rationale of those results can be translated into understandable reasons that can be subjected to rational evaluation and critical assessment; and
2. A **social component**, which involves clarifying the socially meaningful content of the outcome of a given algorithmically assisted decision by translating that model's technical machinery—its input and output variables, parameters, and functional rationale—back into the everyday language of the humanly relevant categories and relationships that informed the formulation of its purpose, objective, and intended elements of design in the first place. Only through this re-translation will the effects of that model's output on the real human life it impacts be understandable in terms of the specific social and individual context of that life and be conveyable as such.

These two components of the delivery process will be fleshed out in turn.

**Technical component of responsible implementation:** As a general rule, we use the results of statistical analysis to guide our actions, because, when done properly, this kind of analysis offers a solid basis of empirically derived evidence that helps us to exercise sound and well-supported judgment about the matters it informs.

Having a good understanding of the factors that are at work in producing the result of a particular statistical analysis (such as in an algorithmic decision-support system) means that we are able to grasp these factors (for instance, input features that weigh heavily in determining a given algorithmically generated decision) as reasons that may warrant the rational acceptability of that result. After all, seen from the perspective of the interpretability of such an analysis, these factors are, in fact, nothing other than *reasons that are operating to support its conclusions*.

Clearly understood, these factors that lie behind the logic of the result or decision are not ‘causes’ of it. Rather, they form the evidentiary basis of its rational soundness and of the goodness of the inferences that support it. Whether or not we ultimately agree with the decision or the result of the analysis, the reasons that work together to comprise its conclusions make ***claims to validity*** and can *as such* be called before a tribunal of ***rational criticism***. These reasons, in other words, must bear the burden of continuous assessment, evaluation, and contestation.

This is an element especially crucial for the responsible implementation of AI systems: **Because they serve surrogate cognitive functions in society, their decisions and results are in no way immune from these demands for rational justification and thus must be delivered to be optimally responsive to such demands.**

The results of algorithmic decision support systems, in this sense, serve as stand-ins for acts of speech and representation and therefore bear the justificatory burdens of those cognitive functions. They must establish the validity of their conclusions and operate under the constraint of being surrogates of the dialogical goal to convince through good reasons.

This charge to be responsive to the demands of rational justification should be essential to the way you map out your delivery strategy. **When you devise how best to relay and explain the statistical results of your AI systems, you need to start from the role they play in supporting evidence-based reasoning.**

This, however, is no easy job. Interpreting the results of data scientific analysis is, more often than not, a highly technical activity and can depart widely from the conventional, everyday styles of reasoning that are familiar to most. Moreover, the various performance metrics deployed in AI systems can be confusing and, at times, seem to be at cross-purposes with each other, depending upon the metrics chosen. There is also an unavoidable dimension of uncertainty that must be accounted for and expressed in confidence intervals and error bars which may only bring further confusion to users and decision subjects.

Be that as it may, by taking a **deliberate and human-centred approach** to the delivery process, you should be able to find the most effective way to convey your model’s statistical results to users and decision subjects in non-technical and socially meaningful language that enables them to understand and evaluate the rational justifiability of those results. A good point of departure for this is to divide your map-building task into the *means of content delivery* and the *substance of the content to be delivered*.

***Means of content delivery:*** When you start mapping out serviceable ways of presenting and communicating your model’s results, you should consider **the users’ and decision subjects’ perspectives to be of primary importance**. Here are a few guiding questions to ask as you sketch out this dimension of your delivery process as well as some provisional answers to them:

- How can the delivery process of explaining the AI system’s results aid and augment the user’s and decision subject’s *mental models* (their ways of organising and filtering information), so that they can get a clear picture of the technical meaning of the

**assessment or explanation? What is the best way to frame the statistical inferences and meanings so that they can be effectively integrated into each user's own cognitive space of concepts and beliefs?**

While answering these questions will largely depend both on your use case and on the type of AI application you are building, it is just as important that you start responding to them by concentrating on the differing needs and capabilities of your explainees. To do this properly, you should first seek input from domain experts, users, and affected stakeholders, so that you can suitably scan the horizons of existing needs and capabilities. Likewise, you should take a human-centred approach to exploring the types of explanation delivery methods that would best be suited for each of your target groups. Much valuable research has been done on this in the field of human-computer interaction and in the study of human factors. This work should be consulted when mapping delivery means.

Once you have gathered enough background information, you should begin to plan out how you are going to **line up your means of delivery with the varying levels of technical literacy, expertise, and cognitive need possessed by the relevant stakeholder groups, who will be involved in the implementation of your project**. Such a **multi-tiered approach** minimally requires that individual attention be paid to the explanatory needs and capacities of implementers, system operators, and decision subjects and their advocates. This multi-tiered approach will pose different challenges at each different level.

For instance, the mental models of implementers—i.e. their ways of conceptualising the information they are receiving from the algorithmic decision-support system—may, in some cases, largely be shaped by their accumulation of domain know-how and by the filter of on-the-job expertise that they have developed over long periods of practice. These users may have a predisposition to automation distrust or aversion bias, and this should be taken into account when you are formulating appropriate means of explanation delivery.

In other contexts, the opposite may be the case. Where implementers tend to over-rely on or over-comply with automated systems, the means of explanation delivery must anticipate a different sort of mental model and adjust the presentation of information accordingly.

In any event, you will need to have a good empirical understanding of your implementer's decision-making context and maintain such knowledge through ongoing assessment. In both bias risk areas, the conveyance and communication of the assessments generated by algorithmic decision-support systems should attempt to bolster each user's practical judgment in ways that mitigate the possibility of either sort of bias. These assessments should present results as evidence-based reasons that support and better capacitate the objectivity of these implementers' reasoning processes.

The story is different with regard to the cognitive life of the technically inclined user. The mental models of system operators, who are natives in the technical vocabulary and epistemic representations of the statistical results, may be adept at the model-based problem-solving tasks that arise during implementation but less familiar with identifying and responding to the cognitive needs and limitations of non-technical stakeholders. Incorporating ongoing communication exercises and training into their roles in the delivery process may capacitate them to better facilitate implementers' and decision subjects' understanding of the technical details of the assessments generated by algorithmic decision-support systems. These ongoing development activities will not only helpfully enrich operators' mental models, they may also inspire them to develop deeper, more responsive, and more effective ways of communicating the technical yields of the analytics they oversee.

Finally, the mental models of decision subjects and their advocates will show the broadest range of conceptualisation capacities, so your delivery strategy should (1) prioritise the facilitation of optimal explanation at the baseline level of the needs of the most disadvantaged of them and (2) build the depth of your multi-tiered approach to providing effective explanations into the delivery options presented to decision subjects and their advocates. This latter suggestion entails that, beyond provision of the baseline explanation of the algorithmically generated result, options should be given to decision subjects and their advocates to view more detailed and technical presentations of the sort available to implementers and operators (with the proviso that reasonable limitations be placed on transparency in accordance with the need to protect the confidential personal and organisational information and to prevent gaming of the system).

- **How can non-technical stakeholders be adequately prepared to gain baseline knowledge of the kinds of statistical and probabilistic reasoning that have factored into the technical interpretation of the system's output, so that they are able to comprehend it on its own technical terms? How can the technical components be presented in a way that will enable explainees to easily translate the statistical inferences and meanings of the results into understandable and rationally assessable terms? What are the best available media for presenting the technical results in engaging and comprehensible ways?**

To meet these challenges, you should consider supplementing your implementation platform with knowledge-building and enrichment resources that will provide non-technical stakeholders with access to basic technical concepts and vocabulary. At a minimum, you should consider building a plain language glossary of basic terms and concepts that will include all of the technical ideas covered by the algorithmic component of a given explanation. If your explanation platforms are digital, you should also make them as user friendly as possible by hyperlinking the technical terms used in the explanations to their plain language glossary elaborations.

Where possible, explanatory demonstrations of technical concepts (like performance metrics, formal fairness criteria, confidence intervals, etc.) should be provided to users and decision subjects in an engaging and easy-to-comprehend way, and

graphical and visualisation techniques should be consistently used to make potentially difficult ideas more accessible. Moreover, the explanation interfaces themselves should be as simple, learnable, and usable as possible. They should be tested to measure the ease with which those with neither technical experience nor domain knowledge are able to gain proficiency in their use and in understanding their content.

**Substance of the technical content to be delivered:** The overall interpretability of your AI system will largely hinge on the effectiveness and even-handedness of your technical content delivery. You will have to strike a balance between (1) determining how best to convey and communicate the rationale of the statistical results so that they may be treated appropriately as decision supporting and clarifying reasons and (2) being clear about the limitations of and potential uncertainties in the statistical results themselves so that the explanations you offer will not mislead implementers and decision subjects. These are not easy tasks and will require substantial forethought as you map out the content clarification aspect of your delivery process.

To assist you in this, here is a non-exhaustive list of recommendations that you should consider as you map out the execution of the technical content delivery component of the responsible implementation of your AI project (This list will, for the sake of specificity, assume the predictive risk assessment example):

- Each explanation should be presented in plain, non-technical language and in an optimally understandable way so that the results provided can enable the affordance of better judgment on the part of implementers and optimal understanding on the part of decision subjects. On the implementer's side, the primary goal of the explanation should be to support the user's ability to offer solid, coherent, and reasonable justifications of their determinations of decision outcomes. On the decision subject's side, the primary goal of the explanation should be to make maximally comprehensible the rationale behind the algorithmic component of the decision process, so that the decision subject can undertake a properly informed critical evaluation of the decision outcome as a whole.
- Each explanation should present its results as facts or evidence in as sparse but complete and sound a manner as possible with a clear indication of what components in the explanation are operating as premises, what components are operating as conclusions, and what the inferential rationale is that is connecting the premises to the conclusions. Each explanation should therefore make explicit the rational criteria for its determination whether this be, for example, global inferences drawn from the population-based reasoning of a demographic analysis or more locally or instance-based inferences drawn from the indication of feature significance by a proxy model. In all cases, the optimisation criteria of the operative algorithmic system should be specified, made explicit, and connected to the logic and rationale of the decision.
- Each explanation should make available the records and activity-monitoring results that the design and development processes of your AI project yielded. Building this link between the process transparency dimension of your project and its outcome transparency will help to make its result, as a whole, more sufficiently interpretable. This

can be done by simply linking or including the public-facing component of the process log of your PBG Framework.

- Each explanation provided to an implementer should come with a standard **Implementation Disclaimer** that may read as follows:

**Implementation Disclaimer:**

These results are intended to assist you in making an evidence-based judgment. They are meant neither to replace your reasoned deliberations nor to constitute the sole evidentiary basis of your judgement. These results are also derived from statistical analysis. This means (1) that there are unavoidable possibilities of error and uncertainty in their results, which are specified in the performance measures and confidence intervals provided and (2) that these results are based on population-level data that do not refer specifically to the actual circumstances and abilities of the individual subject of their prediction. The inferences you draw directly from them will therefore be based on statistical generalisation not on an understanding of the life context or concrete potential of the individual person, who will be impacted by your decision.

- Each explanation should specify and make explicit its governing performance metrics together with the acceptability criteria used to select those metrics and any standard benchmarks followed in establishing that criteria. Where appropriate and possible, fuller information about model validation measurement (including confusion matrix and ROC curve results) and any external validation results should be made available.
- Each explanation should provide confirmatory information that the formal fairness criteria specified in your project's Fairness Policy Statement has been met.
- Each explanation should include clear representations of confidence intervals and error bars. These certainty estimates should make as quantitatively explicit as possible the confidence range of specific predictions, so that users and decision subjects can more fully understand their reliability and the levels of uncertainty surrounding them.
- When an explanation offers categorically ordered scores (for instance, risk scores on a scale of 1 to 10), that explanation must also explicitly indicate the actual raw numerical probabilities for the labels (predicted outcomes) that have been placed into those categories. This will help your delivery process avoid producing confusion about the relative magnitudes of the categorical groupings under which the various scores fall. Information should also be provided about the relative distances between the risk scores of specific cases if the risk categories under which they are placed are unevenly distributed. It may be possible, for example, for two cases, which fall under the same high risk category (say, 9) to be farther apart in terms of the actual values of their risk probabilities than two other cases in two different categories (say 1 and 4). This may be misleading to the user.

- Each explanation should, where possible, include a counterfactual explanatory tool, so that implementers and affected individuals have the opportunity to gain a better contrastive understanding of the logic of the outcome and its alternative possibilities.

**Social component of responsible implementation:** We have now established the first step in the delivery of a responsible implementation process: making clear the rationale behind the technical content of an algorithmic model’s statistical results and determining how best to convey and communicate it so that these results may be appropriately treated as decision supporting and clarifying reasons. This leaves us with a second related task of content clarification, which is only implicit in the first step but must be made explicit and treated reflectively in a second.

Beyond translating statistically expressed results into humanly significant reasons, you will have to make sure that their ***socially meaningful content*** is clarified by implementers, so that they are able to thoughtfully apply these results to the real human lives they impact in terms of the specific societal and individual contexts in which those lives are situated.

This will involve explicitly translating that model’s technical machinery—its input and output variables, parameters, and functional rationale—*back* into the everyday language of the humanly relevant meanings, categories, and relationships that informed the formulation of its purpose, objectives, and intended elements of design in the first place. It will also involve training and preparing implementers to intentionally assist in carrying out this translation in each particular case, so that due regard for the dignity of decision subjects can be supported by the interpretive charity, reasonableness, empathy, and context-specificity of the determination of the outcomes that affect them.

Only through this re-translation will the internals, mechanisms, and output of the model become ***useably interpretable*** by implementers: Only then will they be able to apply input features of relevance to the specific situations and attributes of decision subjects. Only then will they be able to critically assess the manner of inference-making that led to its conclusion. And only then will they be able to adequately weigh the normative considerations (such as prioritising public interest or safeguarding individual well-being) that factored into the system’s original objectives.

Having clarified the socially meaningful content of the model’s results, the implementer will be able to more readily apply its evidentiary contribution to a more holistic and wide-angled consideration of the particular circumstances of the decision subject while, at the same time, weighing these circumstances against the greater purpose of the algorithmically assisted assessment. It is important to note here that the understanding enabled by the clarification of the social context and stakes of an algorithmically supported decision-making process goes hand-in-glove with fuller considerations of the moral justifiability of the outcome of that process.

A good starting point for considering how to integrate this clarification of the socially meaningful content of an algorithmic model’s output into your map of the delivery process is to consider what you might think of as your AI project’s **content lifecycle**.

**The content lifecycle:** The output of an algorithmic system does not begin and end with the computation. Rather, it begins with the very human purposes, ideas, and initiatives that lay behind the conceptualisation and design of that system. Creating technology is a shared public activity, and it is animated by human objectives and beliefs. An algorithmic system is brought into the world as the result of this collective enterprise of ingenuity, intention, action, and collaboration.

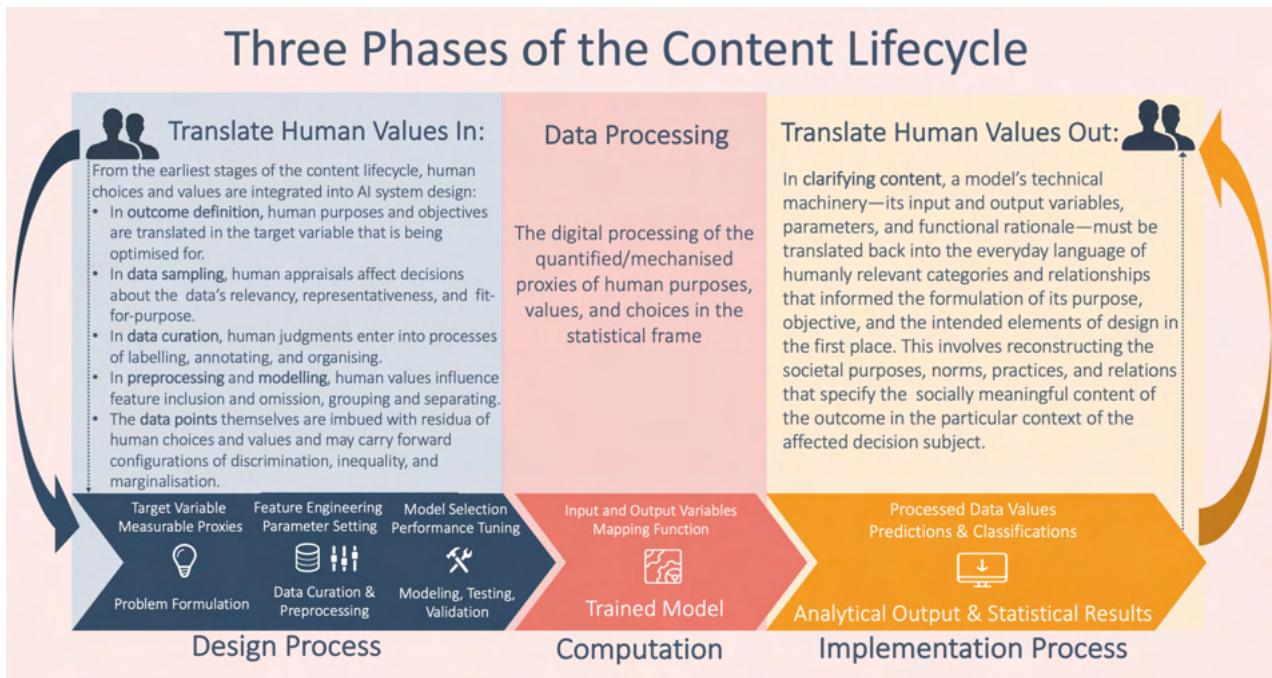
Human choices and values therefore punctuate the design and implementation of AI systems. These choices and values are inscribed in algorithmic models:

- At the very inception of an AI project, human choices and values come into play when we formulate the goals and objectives to be achieved by our algorithmic technologies. They come into play when we define the optimal outcome of our use of such technologies and when we translate these goals and objectives into target variables and their measurable proxies.
- Human choices and values come into play when decisions are made about the sufficiency, fit-for-purpose, representativeness, relevance, and appropriateness of the data sampled. They come into play in how we curate our data—in how we label, organise, and annotate them.
- Such choices and values operate as well when we make decisions about how we craft a feature space—how we select or omit and aggregate or segregate attributes. Determinations of what is relevant, reasonable, desirable, or undesirable will factor into what kinds of inputs we are going to include in the processing and how we are going to group and separate them.
- Moreover, the data points themselves are imbued with residua of human choices and values. They carry forward historical patterns of social and cultural activity that may contain configurations of discrimination, inequality, and marginalisation—configurations that must be thoughtfully and reflectively considered by implementers as they incorporate the analytics into their reasoned determinations.

Whereas all of these human choices and values are translated into the algorithmic systems we build, the responsible implementation of these systems requires that they be translated out. The rationale and logic behind an algorithmic model's output can be properly understood as it affects the real existence of a decision subject only when we transform its variables, parameters, and analytical structures back into the human currency of values, choices, and norms that shaped the construction of its purpose, its intended design, and its optimisation logic from the start.

It is only in virtue of this **re-translation** that an algorithmically supported outcome can afford stakeholders the degree of deliberation, dialogue, assessment, and mutual understanding that is necessary to make it fully comprehensible and justifiable to them. And, it is likewise only in virtue of this re-translation that the implementation process itself can, at once, secure end-to-end accountability and give due regard to the SUM values.

The content lifecycle of algorithmic systems therefore has three phases: (1) The **translation in** of human purposes, values, and choices during the design process; (2) The digital processing of the quantified/mechanised proxies of these purposes, values, and choices in the statistical frame; (3) The **translation out** of the purposes, values, and choices in clarifying the socially meaningful content of the result as it affects the life of the decision subject through the implementation process. Here is a visualisation of these three phases of the content lifecycle:



**The translation rule:** A beneficial result of framing the implementation process in terms of the content lifecycle is that it gives us a clear and context-sensitive measure by which to identify the explanatory needs of any given AI application. We can think of this measurement as the **translation rule**. It states that:

What is *translated in* to an algorithmic system with regard to the human choices and societal values that determine its content and purpose is directly proportional to what, in terms of the explanatory needs of clarification and justification, must be *translated out*.

The translation rule organically makes two distinctions that have great bearing on the delivery process for responsible implementation. First, it divides the question of what needs explaining into two parts: (1) issues of socially meaningful content in need of clarification (i.e., the explanatory need that comes from the **translation in** to the AI model of the categories, meanings, and relations that originate in social practices, beliefs, and intentions) (2) issues of normative rightness in need of justification (i.e. the explanatory need that comes from **translation in** to the AI model of choices and considerations that have bearing on its ethical permissibility, discriminatory non-harm, and public trustworthiness). These two parts line up with what we have [above](#) called **interpretable AI** and **justifiable AI** respectively, and what we have also identified as [tasks 2 and 3](#) of delivering transparent AI.

Secondly, the translation rule divides the two dimensions of translation (translation in and translation out) into aspects of **intention-in-design** and **intention-in-application**. *Translating in*

has to do with *intention-in-design*. It involves an active awareness of the human purposes, objectives, and intentions that factor into the construction of AI systems. *Translating out*, on the other hand, has directly to do with *intention-in-application*, or put differently, the intentional dimension of the implementation of an AI system by a user in a specific context and with direct consequences for a subject affected by its outcome.

In human beings, intention-in-design and intention-in-application are *united in intelligent action*, and it is precisely this unity that enables people to reciprocally hold each other accountable for the consequences of what they say and what they do. By contrast, in artificial intelligence systems, which fulfil surrogate cognitive functions in society but are themselves neither intentional nor accountable, design and application are divided. In these systems, intention-in-design and intention-in-application are and must remain *punctuation points of human involvement and responsibility* that manifest on either side of the vacant mechanisms of data processing. This is why translation is so important, and this is why enabling the implementer's capacity to *intentionally translate out the social and normative content* of the model's results is such a critical element of the responsible delivery of your AI project.

It might be helpful to think more concretely about the translation rule by considering it in action. Let's compare two hypothetical examples: (1) a use case about an early cancer detection system in radiomics (a machine learning application that uses high throughput computing to identify features of pathology that are undetectable to the trained radiological eye); and (2) a use case about a predictive risk assessment application that supports decision-making in child social care.

In the radiomics case, the *translating in* dimension involves minimal social content: the clinical goal inscribed in the model's objective is that of lesion detection and the features of relevance are largely voxels extracted from PET and CT scanner images. However, the normative aspect of *translating in* is, in this case, significant. Ethical considerations about looking after patient wellbeing and clinical safety are paramount and wider justice concerns about improving healthcare for all and health equity factor in as well.

The explanatory needs of the physician/implementer receiving clinical decision support and of the clinical decision subject will thus lean less heavily on the dimension of the clarification of socially meaningful content than it will on the normative dimension of justifying the safety of the system, the priority of the patient's wellbeing, and the issues of improved delivery and equitable access. The technical content of the decision support may be crucial here (Issues surrounding the reproducibility of the results and the robustness of the system may, in fact, be of great concern in the assessment of the validity of the outcome.), but the *translating out* component of the implementation remains directly proportional to the minimal social content and to the substantial ethical concerns and objectives that were *translated in* and that thus inform the explanatory and justificatory needs of the result in general.

The explanatory demands in the child social care risk assessment use case are entirely different. The social content of the *translating in* dimension is intricate, multi-layered, and extensive. The chosen target variable may be child safety or the prevention of severe mistreatment and the measurable proxy, home removal of at-risk children within a certain timeframe. Selected features that are deemed relevant may include the age of the at-risk

children, public health records, previous referrals, family history of violent crime, welfare records, juvenile criminal records, demographic information, and mental health records. Complex socioeconomic and cultural formations may additionally influence the representativeness and quality of the dataset as well as the substance of the data itself.

The normative aspect of *translating in* here is also subtle and complicated. Ethical considerations about protecting the welfare of children at risk are combined with concerns that parents and guardians be treated fairly and without discrimination. Objectives of providing evidence-based decision support are also driven by hopes that accurate results and well-reasoned determinations will preserve the integrity and sanctity of familial relations where just, safe, and appropriate. Other goals and purposes may be at play as well such as making an overburdened system of service provision more efficient or accelerating real-time decision-making without harming the quality of the decisions themselves.

In this case of predictive risk assessment, the *translating out* burdens of the frontline social worker are immense both in terms of clarifying content and in terms of moral justification. If, for example, analytical results yielding a high risk score were based on the relative feature importance of demographic information, welfare records, mental health records, and criminal history, the implementer would have to scrutinise the particular decision subject's situation, so that the socially meaningful content of these factors could be clarified in terms of the living context, relevant relationships, and behavioural patterns of the stakeholders directly affected. Only then could the features of relevance be thoroughly and deliberatively assessed.

The effective interpretability of the model's result would, in this case, heavily depend on the implementer's ability to apply domain-knowledge in order to reconstruct the meaningful social formations, intentions, and relationships that constituted the concrete form of life in which the predictive risk modelling applies. The implementer's well-reasoned decision here would involve a careful weighing of this socially clarified content against the wider predictive patterns in the data distribution yielded by the model's results—patterns that may have otherwise gone unnoticed.

Such a weighing process would, in turn, be informed by the normative-explanatory need to *translate out* the morally implicating choices, concerns, and objectives that influenced and informed the predictive risk assessment model's development in the first place. Again, the interpretive burden of the frontline social worker would be immense here. First, this implementer would have to deliberate with a critically informed awareness of the legacies of discrimination and inequity that tend to feed forward in the kinds of evidentiary sources drawn upon by the analytics. Such an active reflexivity is crucial for retaining the punctuating role of human involvement and responsibility in these sensitive and high-stakes environments.

Just as importantly, the frontline social worker would have to evaluate the real impact of ethical objectives at the point of delivery. Not only would the results of the analytics have to be aligned with the ethical concerns and purposes that fostered the construction of the model, this implementer would have to reflectively align their own potentially diverging ethical point of view both with those results and with those objectives. This *normative*

***triangulation*** between the original intention-in-design, the implementer's intention-in-application, and the content clarification of the AI system's results is, in fact, a crucial safeguard to the delivery of justifiable AI. It again enables a reanimation of moral involvement and responsibility at the most critical juncture of the content lifecycle.

### Step 3: Build an ethical implementation platform:

- (1) **Train ethical implementation.** The continuous challenges of translation, content clarification, and normative explanation should inform how you set up your implementation training to achieve optimal outcome transparency. In addition to the necessary [training to prevent implementation biases in the users of your AI system](#) (discussed above), you should prepare and train the implementers to be stewards of interpretable and justifiable AI. This entails that they be able to:
  - Rationally evaluate and critically assess the logic and rationale behind the outputs of the AI systems;
  - Convey and communicate their algorithmically assisted decisions to the individuals affected by them in plain language. This includes explaining to them in an everyday, non-technical, and accessible way how and why the decision-supporting model performed the way it did in a specific context and how that result factored into the final outcome of the implementation;
  - Apply the conclusions reached by the AI model to a more focused consideration of the particular social circumstances and life context of the decision subject and other affected parties;
  - Treat the inferences drawn from the results of the model's computation as evidentiary contributions to a broader, more rounded, and coherent understanding of the individual situations of the decision subject and other affected parties;
  - Weigh the interpretive understanding gained by integrating the model's insights into this rounded picture of the life context of the decision subject against the greater purpose and societal objective of the algorithmically assisted assessment;
  - Justify the ethical permissibility, the discriminatory non-harm, and the public trustworthiness both of the AI system's outcome and of the processes behind its design and use
- (2) **Make your implementation platform a relevant part and capstone of the sustainability track of your project.** An important element of gauging the impacts of your AI technology on the individuals and communities it touches is having access to the frontlines of its potentially transformative and long-term effects. Your implementation platform should assist you in gaining this access by being a *two-way medium of application and communication*. It should both enable you to sustainably achieve the objectives and goals you set for your project through responsible implementation, but it should also be a sounding board as well as a site for feedback and cooperative sense-checking about the real-life effects of your system's use.

Your implementation platform should be dialogically and collaboratively connected to the stakeholders it effects. It should be bound to the communities it serves as part of a shared project to advance their immediate and long-run wellbeing.

- (3) **Provide a model sheet to implementers and establish protocols for implementation reporting.** As part of the roll-out of your AI project, you should prepare a summary/model sheet for implementers, which includes summation information about the system's technical specifications and all of the relevant details indicated above in the section on [substance of the technical content to be delivered](#). This should include relevant information about performance metrics, formal fairness criteria and validation, the implementation disclaimer, links or summaries to the relevant information from the process logs of your PBG Framework, and links or summary information from the Stakeholder Impact Assessment.

You should also set up protocols for implementation reporting that are proportional to the potential impacts and risks of the system's use.

- (4) **Foster outcome understanding through dialogue.** Perhaps the single most important aspect of building a platform for ethical implementation is the awareness that the realisation of interpretable and justifiable AI is a dialogical and collaborative effort. Because all types of explanation are mediated by language, each and every explanatory effort is a participatory enterprise where understanding can be reached only through acts of communication. The interpretability and justifiability of AI systems depend on this shared human capacity to give and ask for reasons in the ends of reaching mutual understanding. Implementers and decision subjects are, in this respect, first and foremost participants in an explanatory dialogue, and the success of their exchange will hinge both on a reciprocal readiness take the other's perspective and on a willingness to enlarge their respective mental models in accordance with new, communicatively achieved, insights and understandings.

For these reasons, your implementation platform should encourage open, mutually respectful, sincere, and well-informed dialogue. Reasons from all affected voices must be heard and considered as demands for explanation arise, and manners of response and expression should remain clear, straightforward, and optimally accessible. Deliberations that have been inclusive, unfettered, and impartial tend to generate new ideas and insights as well as better and more inferentially sound conclusions, so approaching the interpretability and justifiability of your AI project in this manner will not only advance its responsible implementation, it will likely encourage further improvements in its design, delivery, and performance.

## Conclusion:

In 1936, a 23-year-old mathematician from Maida Vale named Alan Turing sat down with pencil and paper. Using just the image of a linear tape divided evenly into squares, a list of symbols, and a few basic rules, he drew a sketch to show the step-by-step process of how a human being can carry out any calculation, from the simplest operation of arithmetic to the most complex nonlinear differential equation.

Turing's remarkable invention (now known simply as the Turing machine) solved the perplexing and age-old mathematical question of *what an effective calculation is*—the question of *how to define an algorithm*. Not only did Turing show what it means to compute a number by showing *how humans do it*, he created, in the process, the idea behind the modern general purpose computer. Turing's astonishingly humble innovation ushered in the digital age.

Just over eight decades later, as we step forward together into the open horizons of a rapidly evolving digital future, it is difficult to image that what started as a thought experiment in a small room at Kings College, Cambridge has now become such a humanly defining force. We live in an increasingly dynamic and integrated computational reality where connected devices containing countless sensors and actuators intermingle with omnipresent algorithmic systems and cloud computing platforms.

With the rise of the Internet of Things, edge computing, and the expanding smart automation of infrastructure, industry, and the workplace, AI systems are progressively more coming to comprise the cyber-physical frame and fabric of our networked society. For better or worse, artificial intelligence is not simply becoming a general purpose technology (like steam power or electricity). It is, more essentially, becoming a gatekeeper technology that uniquely holds the key both to the potential for the exponential advancement of human wellbeing and to possibilities for the emergence of significant risks for society's future. It is, as yet, humankind that must ultimately choose which direction the key will turn.

This choice leaves difficult questions in the lap of the moral agency of the present: What shape will the data-driven society of tomorrow take? How will the values and motivations that are currently driving the gathering energies of technological advancement in artificial intelligence come both to influence that future society's ways of life and to transform the identities of its warm-blooded subjects?

This guide on understanding AI ethics and safety has offered you one way to move forward in answering these questions. In a significant sense, it has attempted to prepare you to take Turing's lead: to see the design and implementation of algorithmic models as an eminently *human activity*—an activity guided by our purposes and values, an activity for which, each of us, who is involved in the development and deployment of AI systems, is morally and socially responsible.

This starting point in human action and intention is a crucial underpinning of responsible innovation. For, it is only when we prioritise considerations of the ethical purposes and values behind the trajectories of our technological advancement, that we, as vested societal stakeholders, will be able to take the reins of innovation and to steer the course of our algorithmic creations in accordance with a shared vision of what a better human future should look like.

## Acknowledgments:

Writing this guide would simply not have been possible without the hard work, dedication, and insight of so many interlocutors both within The Alan Turing Institute and through the meaningful partnerships that the Turing’s Public Policy Programme has formed with stakeholders from across the UK Government.

To take the latter group first, the Office for Artificial Intelligence (OAI) and the Government Digital Service (GDS)’s keen vision and their commitment to responsible AI innovation have been an enabling condition of the development of this work. In particular, the patience and incisiveness of OAI’s Sébastien Krier and Jacob Beswick, and GDS’ Bethan Charnley have been instrumental in bringing the project to its completion.

I am also incredibly grateful for the impact that our interactions with the Ministry of Justice (MoJ)’s Data Science Hub has had on developing the framing for this guide. Input from the MoJ’s Megan Whewell, Philip Howard, Jonathan Roberts, Olivia Lewis, Ross Wyatt, and from its Data Science Innovation Board have left a significant mark on the research.

Last, but not least, our ongoing partnership with the Information Commissioner’s Office on Project ExplA/n—and, in particular, with ICO colleagues Carl Wiper and Alex Hubbard—has been a key contributor to this guide’s focus on fairness, transparency, and accountability. Project ExplA/n aims to provide practical guidance for organisations on explaining AI supported decisions to the subjects of those decisions. Taking inspiration from our work on Project ExplA/n and from the input gathered over the course of the two citizens’ juries we held in Manchester and Coventry, the current guide emphasises the importance of communication and attempts to build out a vision of human-centred and context-sensitive implementation.

As the Ethics Fellow within the Public Policy Programme at the Turing, I have benefited tremendously from being surrounded by an immensely talented group of thinkers and doers, whose commitment to making the connected world a better place through interdisciplinary research and advisory intervention is an inspiration every day. Programme Director, Helen Margetts, and Deputy Director, Cosmina Dorobantu, have been crucial and inimitable supports of this project from its inception as has my small but brilliant team of researchers, Josh Cowls and Christina Hitrova. My involvement with the Turing’s Data Ethics Group has also been a tremendous source of insight and inspiration for this project. Given the ambitious deadlines that accompanied this guide’s final stages of production, heroic efforts to review its contents as a whole or in parts were made by Florian Ostmann, Michael Veale, David Watson, Mark Briers, Evelina Gabsova, Alexander Harris, and Anna FitzMaurice. Their perceptive feedback notwithstanding, any unclarities that appear in *Understanding Artificial Intelligence Ethics and Safety* reflect the faults of its author alone.

## Bibliography and Further Readings

Included here is a bibliography organised into the main themes covered in this guide. Please use this as a starting point for further exploration of these complex topics. Many thanks to the tireless efforts of Jess Morley and Corianna Moffatt without whom this bibliography could not have been compiled.

[The SUM Values](#)

[General fairness](#)

[Data fairness](#)

[Design fairness](#)

[Outcome fairness](#)

[Implementation fairness](#)

[Accountability](#)

[Stakeholder Impact Assessment](#)

[Safety: Accuracy, reliability, security, and robustness](#)

[Transparency](#)

[Process-Based Governance](#)

[Interpretable AI](#)

[Responsible delivery through human-centred implementation protocols and practices](#)

[Individual and societal impacts of machine learning and algorithmic systems](#)

### The SUM Values

Access Now. (2018). *The Toronto declaration: Protecting the rights to equality and non-discrimination in machine learning systems*. Retrieved from [https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration\\_ENG\\_08-2018.pdf](https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf)

Adamson, G., Havens, J. C., & Chatila, R. (2019). Designing a value-driven future for ethical autonomous and intelligent systems. Proceedings of the IEEE, 107(3), 518–525. <https://doi.org/10.1109/JPROC.2018.2884923>

American Medical Association. (2001). AMA code of medical ethics. Retrieved from <https://www.ama-assn.org/sites/ama-assn.org/files/corp/media-browser/principles-of-medical-ethics.pdf>

American Psychological Association. (2016). *Ethical principles of psychologists and code of conduct*. Retrieved from <https://www.apa.org/ethics/code/>

Article 19. (2019). *Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence*. Retrieved from <https://www.article19.org/resources/governance-with-teeth-how-human-rights-can-strengthen-fat-and-ethics-initiatives-on-artificial-intelligence/>

Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics*. 6th edition. Oxford University Press, USA.

- Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), 20180080. <https://doi.org/10.1098/rsta.2018.0080>
- Cowls, J., & Floridi, L. (2018). Prolegomena to a White Paper on an Ethical Framework for a Good AI Society. <http://dx.doi.org/10.2139/ssrn.3198732>
- European Commission. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- European Group on Ethics in Science and New Technologies. (2018). *Artificial intelligence, robotics, and 'autonomous' systems*. Retrieved from [https://ec.europa.eu/research/ege/pdf/ege\\_ai\\_statement\\_2018.pdf](https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf)
- Felten, E. (2016). Preparing for the future of artificial intelligence. *Washington DC: The White House*.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707. Retrieved from <https://link.springer.com/article/10.1007/s11023-018-9482-5>
- Floridi, L., & Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160360. <https://doi.org/10.1098/rsta.2016.0360>
- Future of Life Institute. (2017). *Asilomar AI principles*. Retrieved from <https://futureoflife.org/ai-principles/>
- Global Future Council on Human Rights 2016-2018. (2018). How to prevent discriminatory outcomes in machine learning. *World Economic Forum*. Retrieved from [http://www3.weforum.org/docs/WEF\\_40065\\_White\\_Paper\\_How\\_to\\_Prevent\\_Discriminatory\\_Outcomes\\_in\\_Machine\\_Learning.pdf](http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf)
- House of Lords Select Committee on Artificial Intelligence. (2018). *AI in the UK: Ready, willing and able?*. Retrieved from <https://publications.parliament.uk/pa/ld201719/ldselect/l dai/100/100.pdf>
- IEEE. (2018). *The IEEE Global Initiative on ethics of autonomous and intelligent systems*. Retrieved from [https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead\\_v2.pdf](https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf)
- Latonero, M. (2018). Governing artificial intelligence: Upholding human rights & dignity. *Data & Society*. Retrieved from [https://datasociety.net/wp-content/uploads/2018/10/DataSociety\\_Governing\\_Artificial\\_Intelligence\\_Upholding\\_Human\\_Rights.pdf](https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf)
- The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, D.C.: United States Government Printing Office.
- Nuffield Council on Bioethics. (2015). *The collection, linking and use of data in biomedical research and health care: ethical issues*. Retrieved from <http://nuffieldbioethics.org/wp-content/uploads/Biodata-a-guide-to-the-report-PDF.pdf>
- Nuffield Council on Bioethics. (2018). *Artificial intelligence (AI) in healthcare and research*. Retrieved from <http://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf>
- Pielemeier, J. (2018). The advantages and limitations of applying the international human rights framework to artificial intelligence. *Data & Society: Points*. Retrieved from <https://points.datasociety.net/the-advantages-and-limitations-of-applying-the-international-human-rights-frame-work-to-artificial-291a2dfe1d8a>
- Ramesh, S. (2017). A checklist to protect human rights in artificial-intelligence research. *Nature*, 552(7685), 334–334. <https://doi.org/10.1038/d41586-017-08875-1>
- Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). Artificial intelligence & human rights: Opportunities & risks. *Berkman Klein Center Research Publication*, (2018-6). Retrieved from [https://cyber.harvard.edu/sites/default/files/2018-09/2018-09\\_AIHumanRightsSmall.pdf](https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf)
- Reform. (2018). *Thinking on its own: AI in the NHS*. Retrieved from [https://reform.uk/sites/default/files/2018-11/AI%20in%20Healthcare%20report\\_WEB.pdf](https://reform.uk/sites/default/files/2018-11/AI%20in%20Healthcare%20report_WEB.pdf)

- Royal Society. (2017). *Machine learning: The power and promise of computers that learn by example*. Retrieved from <https://royalsociety.org/~media/policy/projects/machine-learning/publications/machine-learning-report.pdf>
- Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, 361(6404), 751–752. <https://doi.org/10.1126/science.aat5991>
- UK Statistics Authority. (2017). *Code of practice for statistics: Ensuring public confidence in statistics*. Retrieved from <https://www.statisticsauthority.gov.uk/wp-content/uploads/2017/07/DRAFT-Code-2.pdf>
- UNESCO. (2017). *Report of COMEST on robotics ethics*. Retrieved from <http://unesdoc.unesco.org/images/0025/002539/253952E.pdf>
- Université de Montréal. (2017). *Montreal declaration for responsible AI*. Retrieved from <https://www.montrealdeclaration-responsibleai.com/the-declaration>
- US Department of Homeland Security. (2012). *The Menlo report: Ethical principles guiding information and communication technology research*. Retrieved from [https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803\\_1.pdf](https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf)
- US National Science and Technology Council. (2016). *Preparing for the future of artificial intelligence*. Retrieved from [https://obamawhitehouse.archives.gov/sites/default/files/whitehouse\\_files/microsites/ostp/NSTC/preparing\\_for\\_the\\_future\\_of\\_ai.pdf](https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf)
- Villani, C. (2018). For a meaningful artificial intelligence: Towards a French and European strategy. *AI For Humanity*. Retrieved from [https://www.aiforhumanity.fr/pdfs/MissionVillani\\_Report\\_ENG-VF.pdf](https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf).
- Yuste, R., Goering, S., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., ... & Kellmeyer, P. (2017). Four ethical priorities for neurotechnologies and AI. *Nature News*, 551(7679), 159. Retrieved from <https://www.nature.com/news/four-ethical-priorities-for-neurotechnologies-and-ai-1.22960>

## General fairness

- Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. *arXiv:1712.03586*. Retrieved from <https://arxiv.org/abs/1712.03586>
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 377). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3173951>
- Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M., & Wallach, H. (2018). Improving fairness in machine learning systems: What do industry practitioners need?. *ArXiv:1812.05239*. <https://doi.org/10.1145/3290605.3300830>
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31(4), 611–627. <https://doi.org/10.1007/s13347-017-0279-x>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 205395171667967. <https://doi.org/10.1177/2053951716679679>
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59-68). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287598>
- Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv:1901.10002*. Retrieved from <https://arxiv.org/abs/1901.10002>
- Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 440). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3174014>

## Data fairness

- Abadi, D., Agrawal, R., Ailamaki, A., Balazinska, M., Bernstein, P. A., Carey, M. J., ... & Gehrke, J. (2016). The Beckman report on database research. *Communications of the ACM*, 59(2), 92-99. Retrieved from <https://dl.acm.org/citation.cfm?id=2845915>
- Abiteboul, S., & Stoyanovich, J., & Weikum, G. (2015). Data, Responsibly. *ACM Sigmod Blog*. Retrieved from <http://wp.sigmod.org/?p=1900>
- Alper, P., Becker, R., Satagopam, V., Grouès, V., Lebioda, J., Jarosz, Y., ... & Schneider, R. (2018). *Provenance-enabled stewardship of human data in the GDPR era*. <https://doi.org/10.7490/f1000research.1115768.1>
- Ambacher, B., Ashley, K., Berry, J., Brooks, C., Dale, R. L., & Flecker, D. (2007). Trustworthy repositories audit & certification: Criteria and checklist. *Center for Research Libraries, Chicago/Illinois*. Retrieved from [https://www.crl.edu/sites/default/files/d6/attachments/pages/trac\\_0.pdf](https://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf)
- Antignac, T., Sands, D., & Schneider, G. (2016). Data minimisation: A language-based approach (long version). *ArXiv:1611.05642*. Retrieved from <http://arxiv.org/abs/1611.05642>
- Bell, D., L'Hours, H., Lungley, D., Cunningham, & N., Corti, L. (n.d.). Scaling up: digital data services for the social sciences. *UK Data Service*. Retrieved from <https://www.ukdataservice.ac.uk/media/604995/ukds-case-studies-scaling-up.pdf>
- Bower, A., Niss, L., Sun, Y., & Vargo, A. (2018). Debiasing representations by removing unwanted variation due to protected attributes. *arXiv:1807.00461*. Retrieved from <https://arxiv.org/abs/1807.00461>
- Custers, B. (2013). Data dilemmas in the information society: Introduction and overview. In *Discrimination and Privacy in the Information Society* (pp. 3-26). Springer, Berlin, Heidelberg. Retrieved from [https://link.springer.com/chapter/10.1007/978-3-642-30487-3\\_1](https://link.springer.com/chapter/10.1007/978-3-642-30487-3_1)
- Custers, B. H., & Schermer, B. W. (2014). Responsibly innovating data mining and profiling tools: A new approach to discrimination sensitive and privacy sensitive attributes. In *Responsible Innovation 1* (pp. 335-350). Springer, Dordrecht. Retrieved from [https://link.springer.com/chapter/10.1007/978-94-017-8956-1\\_19](https://link.springer.com/chapter/10.1007/978-94-017-8956-1_19)
- Dai, W., Yoshigoe, K., & Parsley, W. (2018). Improving data quality through deep learning and statistical models. *ArXiv:1810.07132*, 558, 515–522. [https://doi.org/10.1007/978-3-319-54978-1\\_66](https://doi.org/10.1007/978-3-319-54978-1_66)
- Davidson, S. B., & Freire, J. (2008). Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 1345-1350). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=1376772>
- European Commission Expert Group on FAIR Data. (2018). Turning FAIR into reality. *European Union*. Retrieved from [https://ec.europa.eu/info/sites/info/files/turning\\_fair\\_into\\_reality\\_1.pdf](https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf)
- Faundeen, J. (2017). Developing criteria to establish trusted digital repositories. *Data Science Journal*, 16. Retrieved from <https://datascience.codata.org/article/10.5334/dsj-2017-022/>
- Joshi, C., Kaloskampis, I., & Nolan, L. (2019). Generative adversarial networks (GANs) for synthetic dataset generation with binary classes. *Data Science Campus*. Retrieved from <https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset-generation-with-binary-classes/>
- L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with Big Data: Challenges and approaches. *IEEE Access*, 5, 7776-7797. Retrieved from <https://ieeexplore.ieee.org/abstract/document/7906512/>
- Ruggieri, S., Pedreschi, D., & Turini, F. (2010). DCUBE: Discrimination discovery in databases. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 1127-1130). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=1807298>
- Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. *LREC*, 859–866.
- Stoyanovich, J., Howe, B., Abiteboul, S., Miklau, G., Sahuguet, A., & Weikum, G. (2017). Fides: Towards a platform for responsible data science. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (p. 26). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3085530>

- Swingler, K. (2011). *The perils of ignoring data suitability: The suitability of data used to train neural networks deserves more attention*. Presented at the NCTA 2011 - Proceedings of the International Conference on Neural Computation Theory and Applications. Retrieved from <http://hdl.handle.net/1893/3950>
- Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big Data*, 5(3), 246-255. Retrieved from <https://www.liebertpub.com/doi/abs/10.1089/big.2016.0051>
- Vidgen, B., Nguyen, D., Tromble, R., Hale, S., Margetts, H., Harris, A. (2019) 'Challenges and frontiers in abusive content detection', *Forthcoming ACL 2019*.
- Zheng, X., Wang, M., & Ordieres-Meré, J. (2018). Comparison of data preprocessing approaches for applying deep learning to human activity recognition in the context of industry 4.0. *Sensors*, 18(7), 2146. Retrieved from <https://www.mdpi.com/1424-8220/18/7/2146>

## Design fairness

- Barocas, S., & Selbst, A. D. (2016). Big Data's disparate impact. *Calif. L. Rev.*, 104, 671. Retrieved from [https://heinonline.org/hol-cgi-bin/get\\_pdf.cgi?handle=hein.journals/calr104&section=25](https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/calr104&section=25)
- Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2), 277–292. <https://doi.org/10.1007/s10618-010-0190-x>
- Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems* (pp. 3992-4001). Retrieved from <http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention>
- d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2), 120-134. <https://doi.org/10.1089/big.2016.0048>
- Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125-2126). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2945386>
- Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1-33. Retrieved from <https://link.springer.com/article/10.1007/s10115-011-0463-8>
- Lehr, D., & Ohm, P. (2017). Playing with the data: What legal scholars should learn about machine learning. *UCDL Rev.*, 51, 653. Retrieved from [https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2\\_Lehr\\_Ohm.pdf](https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf)
- Passi, S., & Barocas, S. (2019). Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 39-48). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287567>
- Singh, J., & Sane, S. S. (2014). Preprocessing technique for discrimination prevention in data mining. *The IJES*, 3(6), 12-16. Retrieved from [https://www.academia.edu/6994180/Pre-Processing\\_Approach\\_for\\_Discrimination\\_Prevention\\_in\\_Data\\_Mining](https://www.academia.edu/6994180/Pre-Processing_Approach_for_Discrimination_Prevention_in_Data_Mining)
- Singhal, S., & Jena, M. (2013). A study on WEKA tool for data preprocessing, classification and clustering. *International Journal of Innovative technology and exploring engineering (IJItee)*, 2(6), 250-253. Retrieved from [https://pdfs.semanticscholar.org/095c/fd6f1a9dc6eaac7cc3100\\_a16cca9750ff9d8.pdf](https://pdfs.semanticscholar.org/095c/fd6f1a9dc6eaac7cc3100_a16cca9750ff9d8.pdf)
- van der Aalst, W. M., Bichler, M., & Heinzl, A. (2017). Responsible data science. *Springer Fachmedien Wiesbaden*. <https://doi.org/10.1007/s12599-017-0487-z>

## Outcome fairness

- Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *ArXiv:1803.02453*. Retrieved from <http://arxiv.org/abs/1803.02453>
- Albarghouthi, A., & Vinitsky, S. (2019). Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 211-219). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287588>
- Chiappa, S., & Gillam, T. P. (2018). Path-specific counterfactual fairness. *arXiv:1802.08139*. Retrieved from <https://arxiv.org/abs/1802.08139>
- Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv:1610.07524*. Retrieved from <http://arxiv.org/abs/1610.07524>
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *ArXiv:1701.08230*. <https://doi.org/10.1145/3097983.309809>
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2090255>
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259-268). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2783311>
- Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 329-338). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287589>
- Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law* (Vol. 1, p. 2). Retrieved from <http://www.mlandthelaw.org/papers/grgic.pdf>
- Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2017). On Fairness, Diversity and Randomness in Algorithmic Decision Making. *arXiv:1706.10208*. Retrieved from <https://arxiv.org/abs/1706.10208>
- Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*. Retrieved from <http://mlg.eng.cam.ac.uk/adrian/AAAI18-BeyondDistributiveFairness.pdf>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323). Retrieved from <http://papers.nips.cc/paper/6373-equality-of-opportunity-in-supervised-learning>
- Johansson, F. D., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. *ArXiv:1605.03661*. Retrieved from <http://arxiv.org/abs/1605.03661>
- Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In P. A. Flach, T. De Bie, & N. Cristianini (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 7524, pp. 35–50). [https://doi.org/10.1007/978-3-642-33486-3\\_3](https://doi.org/10.1007/978-3-642-33486-3_3)
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *ArXiv:1609.05807*. Retrieved from <http://arxiv.org/abs/1609.05807>
- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066-4076). Retrieved from <http://papers.nips.cc/paper/6995-counterfactual-fairness>
- Russell, C., Kusner, M. J., Loftus, J., & Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 6414–6423). Retrieved from <http://papers.nips.cc/paper/7220-when-worlds-collide-integrating-different-counterfactual-assumptions-in-fairness.pdf>

- Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 10-19). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287566>
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1-7). IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8452913>
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *ArXiv:1711.00399*. Retrieved from <http://arxiv.org/abs/1711.00399>
- Wexler, J. (2018). The what-if tool: Code-free probing of machine. *Google AI Blog*. Retrieved from <https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html>
- Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. *arXiv:1507.05259*. Retrieved from <https://arxiv.org/abs/1507.05259>
- Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1171-1180). International World Wide Web Conferences Steering Committee. Retrieved from <https://dl.acm.org/citation.cfm?id=3052660>
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning* (pp. 325-333). Retrieved from <http://proceedings.mlr.press/v28/zemel13.pdf>
- Zhang, J., & Bareinboim, E. (2018). *Fairness in decision-making the causal explanation formula*. Presented at the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018. Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16949>
- Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4), 1060-1089. Retrieved from <https://link.springer.com/article/10.1007/s10618-017-0506-1>

## Implementation fairness

- Alexander, V., Blinder, C., & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior*, 89, 279-288. <https://doi.org/10.1016/j.chb.2018.07.026>
- Bahner, J. E., Hüper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, 66(9), 688-699. <https://doi.org/10.1016/j.ijhcs.2008.06.001>
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211-233. [https://doi.org/10.1016/0001-6918\(80\)90046-3](https://doi.org/10.1016/0001-6918(80)90046-3)
- Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions, *Cognition*, Vol. 181, 21-34. <https://doi.org/10.1016/j.cognition.2018.08.003>
- Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency* (No. ARL-TR-6905). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory. Retrieved from <https://apps.dtic.mil/docs/citations/ADA600351>
- Crocill, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 34, No. 19, pp. 1524-1528). Los Angeles, CA: SAGE Publications. <https://doi.org/10.1177/154193129003401922>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114. Retrieved from [https://repository.upenn.edu/cgi/viewcontent.cgi?article=1392&context=fnce\\_papers](https://repository.upenn.edu/cgi/viewcontent.cgi?article=1392&context=fnce_papers)
- Domeinski, J., Wagner, R., Schoebel, M., & Manzey, D. (2007). Human redundancy in automation monitoring: Effects of social loafing and social compensation. In *Proceedings of the Human Factors and Ergonomics*

- Society 51st Annual Meeting (pp. 587–591). Santa Monica, CA: Human Factors and Ergonomics Society. <https://doi.org/10.1177/154193120705101004>
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1), 79–94. <https://doi.org/10.1518/0018720024494856>
- Gigerenzer, G., & Todd, P. A. (1999). *Simple heuristics that make us smart*. London, England: Oxford University Press.
- Gilovich, Thomas (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press.
- Kahneman, D. (2000). Evaluation by moments: Past and future. *Choices, values, and frames*, 693-708. Retrieved from <http://www.vwl.tuwien.ac.at/hanappi/TEI/momentfull.pdf>
- Kahneman, D. (2011). Thinking, fast and slow. London, England: Allen Lane.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgement under uncertainty: Heuristics and biases. New York, NY: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, 80(4), 237. Retrieved from <https://web.archive.org/web/20160518202232/https://faculty.washington.edu/jmiyamot/p466/kahneman%20psych%20o%20prediction.pdf>
- Karau, S. J., & Williams, K. D. (1993). Social-loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4), 681–706. Retrieved from <https://psycnet.apa.org/buy/1994-33384-001>
- Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological review*, 107(4), 852. <http://dx.doi.org/10.1037/0033-295X.107.4.852>
- Lee, J. D., & See, J. (2004). Trust in automation and technology: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1). <https://doi.org/10.1177/2053951718756684>
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90-103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11), 2098. <http://dx.doi.org/10.1037/0022-3514.37.11.2098>
- McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, 48, 656–665. <https://doi.org/10.1518/001872006779166334>
- Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors*, 58(3), 401-415. <https://doi.org/10.1177/0018720815621206>
- Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics*, 31, 175–178. Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science*, 1, 354–365. [https://doi.org/10.1016/S0169-8141\(02\)00194-4](https://doi.org/10.1016/S0169-8141(02)00194-4)
- Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other?. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and application* (pp. 201–220). Mahwah, NJ: Erlbaum.
- Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision-making and performance in hightech cockpits. *International Journal of Aviation Psychology*, 8, 47–63. [https://doi.org/10.1207/s15327108ijap0801\\_3](https://doi.org/10.1207/s15327108ijap0801_3)
- Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22, 390 – 409. <http://dx.doi.org/10.1002/bdm.637>

- Packin, N. G. (2019). Algorithmic Decision-Making: The Death of Second Opinions?. *New York University Journal of Legislation and Public Policy, Forthcoming*. Retrieved from [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3361639](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3361639)
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3), 381-410. <https://doi.org/10.1177/0018720810376055>
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, 3(1), 1-23. [https://doi.org/10.1207/s15327108ijap0301\\_1](https://doi.org/10.1207/s15327108ijap0301_1)
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, 49(1), 76-87. <https://doi.org/10.1518/001872007779598082>
- Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., & Keim, D. A. (2015). The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1), 240-249. [https://bib.dbvis.de/uploadedFiles/uncertainty\\_trust.pdf](https://bib.dbvis.de/uploadedFiles/uncertainty_trust.pdf)
- Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, 43, 573–583. <https://doi.org/10.1518/001872001775870403>
- Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3), 377-400. <https://pdfs.semanticscholar.org/629b/f1f076f8d5bc203c573d4ba1dad5bb6743cf.pdf>
- Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & cognition*, 21(4), 546-556. <https://doi.org/10.3758/BF03197186>
- Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755-769. <https://doi.org/10.1111/j.1540-5907.2006.00214.x>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157), 1124-1131. Retrieved from <https://science.scienmag.org/content/185/4157/1124>
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453-458. Retrieved from <https://science.scienmag.org/content/211/4481/453>

## Accountability

- AI Now Institute. (2018). *Algorithmic Accountability Policy Toolkit*. Retrieved from <https://ainowinstitute.org/aap-toolkit.pdf>
- Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & Technology*, 31(4), 543-556. Retrieved from <https://link.springer.com/article/10.1007/s13347-017-0263-5>
- Cavoukian, A., Taylor, S., & Abrams, M. E. (2010). Privacy by Design: essential for organizational accountability and strong business practices. *Identity in the Information Society*, 3(2), 405–413. <https://doi.org/10.1007/s12394-010-0053-z>
- Center for Democracy & Technology. (n.d.). *Digital decisions*. Retrieved from <https://cdt.org/issue/privacy-data/digital-decisions/>
- Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism*, 3(3), 398-415. <https://doi.org/10.1080/21670811.2014.976411>
- Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., ... & Wilson, C. (2017). Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML*. Retrieved from <https://www.fatml.org/resources/principles-for-accountable-algorithms>
- Donovan, J., Caplan, R., Hanson, L., & Matthews, J. (2018). Algorithmic accountability: A primer. *Data & Society Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality*. Retrieved from <https://datasociety.net/output/algorithmic-accountability-a-primer/>

- ICO. (2017). *Big Data, artificial intelligence, machine learning and data protection*. Retrieved from <https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf>
- Janssen, M., & Kuk, G. (2016). The challenges and limits of Big Data algorithms in technocratic governance. *Government Information Quarterly*, 33(3), 371–377. <https://doi.org/10.1016/j.giq.2016.08.011>
- Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *U. Pa. L. Rev.*, 165, 633. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/pnlr165&div=20&id=&page=&t=1559932490>
- Malgieri, G., & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*. Retrieved from <https://academic.oup.com/idpl/article-abstract/7/4/243/4626991?redirectedFrom=fulltext>
- O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., ... & Ashrafi, H. (2019). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 15(1), e1968. <https://doi.org/10.1002/rcs.1968>
- Reed, C. (2018). How should we regulate artificial intelligence?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128), 20170360. Retrieved from <https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2017.0360>
- Stahl, B. C., & Wright, D. (2018). Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security & Privacy*, 16(3), 26–33. <https://doi.org/10.1109/MSP.2018.2701164>
- Veale, M., Binns, R., & Edwards, L. (2018). Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133). <https://doi.org/10.1098/rsta.2018.0083>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017a). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, 2(6). <https://doi.org/10.1126/scirobotics.aan6080>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017b). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. <https://doi.org/10.1093/idpl/ixp005>
- Zook, M., Barocas, S., boyd, danah, Crawford, K., Keller, E., Gangadharan, S. P., ... Pasquale, F. (2017). Ten simple rules for responsible Big Data research. *PLOS Computational Biology*, 13(3). <https://doi.org/10.1371/journal.pcbi.1005399>

## Stakeholder Impact Assessment

- AI Now Institute. (2018). Algorithmic Impact Assessments: Toward Accountable Automation in Public Agencies. Retrieved from <https://medium.com/@AINowInstitute/algorithmic-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde>
- Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H. V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C., & Zevenbergen, B. (n.d.). Principles for accountable algorithms and a social impact statement for algorithms. Fairness, Accountability, and Transparency in Machine Learning. Retrieved from: <http://www.fatml.org/resources/principles-for-accountable-algorithms>
- Karlin, M. (2018). A Canadian algorithmic impact assessment. Retrieved from <https://medium.com/@supergovernance/a-canadian-algorithmic-impact-assessment-128a2b2e7f85>
- Karlin, M., & Corriveau, N. (2018). The government of Canada's algorithmic impact assessment: Take two. Retrieved from <https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f>
- Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now institute*. Retrieved from: <https://ainowinstitute.org/aiareport2018.pdf>

Vallor, S. (2018) An ethical toolkit for engineering/design practice. Retrieved from: <https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/>

## Hong Kong

The Information Accountability Foundation. (2018a). *Ethical accountability framework for Hong Kong, China: A report prepared for the Office of the Privacy Commission for Personal Data*. Retrieved from [https://www.pcpd.org.hk/mis/files/Ethical\\_Accountability\\_Framework.pdf](https://www.pcpd.org.hk/mis/files/Ethical_Accountability_Framework.pdf)

The Information Accountability Foundation. (2018b). *Data stewardship accountability, data impact assessments and oversight models: Detailed support for an ethical accountability framework*. Retrieved from [https://www.pcpd.org.hk/mis/files/Ethical\\_Accountability\\_Framework\\_Detailed\\_Support.pdf](https://www.pcpd.org.hk/mis/files/Ethical_Accountability_Framework_Detailed_Support.pdf)

## Canada

Treasury Board of Canada Secretariat. (2019). *Algorithmic impact assessment*. Retrieved from <https://open.canada.ca/data/en/dataset/748a97fb-6714-41ef-9fb8-637a0b8e0da1>

## Safety: Accuracy, reliability, security, and robustness

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv:1606.06565*. Retrieved from <https://arxiv.org/abs/1606.06565>

Auerhammer, K., Kolagari, R. T., & Zoppelt, M. (2019). Attacks on Machine Learning: Lurking Danger for Accountability [PowerPoint Slides]. Retrieved from <https://safeai.webs.upv.es/wp-content/uploads/2019/02/3.SafeAI.pdf>

Demšar, J., & Bosnić, Z. (2018). Detecting concept drift in data streams using model explanation. *Expert Systems with Applications*, 92, 546–559. <https://doi.org/10.1016/j.eswa.2017.10.003>

Google. (2019). *Perspectives on issues in AI governance*. Retrieved from <https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf>

Göpfert, J. P., Hammer, B., & Wersing, H. (2018). Mitigating concept drift via rejection. In *International Conference on Artificial Neural Networks* (pp. 456-467). Springer, Cham. [https://doi.org/10.1007/978-3-030-01418-6\\_45](https://doi.org/10.1007/978-3-030-01418-6_45)

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill*, 4(2). <https://doi.org/10.23915/distill.00014>

Kohli, P., Dvijotham, K., Uesato, J., & Gowal, S. (2019). Towards a robust and verified AI: Specification testing, robust training, and formal verification. *DeepMind Blog*. Retrieved from <https://deepmind.com/blog/robust-and-verified-ai/>

Kolter, Z., & Madry, A. (n.d.). Materials for tutorial adversarial robustness: Theory and practice. Retrieved from <https://adversarial-ml-tutorial.org/>

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv:1801.00631*. Retrieved from <https://arxiv.org/abs/1801.00631>

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C., & Roli, F. (2017, November). Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 27-38). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3140451>

Nicolae, M. I., Sinn, M., Tran, M. N., Rawat, A., Wistuba, M., Zantedeschi, V., ... & Edwards, B. (2018). Adversarial Robustness Toolbox v0.4.0. *arXiv:1807.01069*. Retrieved from <https://arxiv.org/abs/1807.01069>

- Ortega, P. A., & Maini, V. (2018). Building safe artificial intelligence: specification, robustness, and assurance. *DeepMind Safety Research Blog, Medium*. Retrieved from <https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1>
- Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. (2017). Improving network robustness against adversarial attacks with compact convolution. *arXiv:1712.00699*. Retrieved from <https://arxiv.org/abs/1712.00699>
- Ratasich, D., Khalid, F., Geissler, F., Grosu, R., Shafique, M., & Bartocci, E. (2019). A roadmap toward the resilient internet of things for cyber-physical systems. *IEEE Access*, 7, 13260-13283. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8606923>
- Salay, R., & Czarnecki, K. (2018). Using machine learning safely in automotive software: An assessment and adaption of software process requirements in iso 26262. *arXiv:1808.01614*. Retrieved from <https://arxiv.org/abs/1808.01614>
- Shi, Y., Erpek, T., Sagduyu, Y. E., & Li, J. H. (2018). Spectrum data poisoning with adversarial deep learning. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)* (pp. 407-412). IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8599832/>
- Song, Q., Jin, H., Huang, X., & Hu, X. (2018). Multi-Label Adversarial Perturbations. In *2018 IEEE International Conference on Data Mining (ICDM)* (pp. 1242-1247). IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8594975>
- Warde-Farley, D., & Goodfellow, I. (2016). Adversarial perturbations of deep neural networks. In T. Hazan, G. Papandreou, & D. Tarlow (Eds.), *Perturbations, Optimization, and Statistics*, 311. Cambridge, MA: The MIT Press.
- Webb, G. I., Lee, L. K., Goethals, B., & Petitjean, F. (2018). Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32(5), 1179-1199. Retrieved from <https://link.springer.com/article/10.1007/s10618-018-0554-1>
- Zantedeschi, V., Nicolae, M. I., & Rawat, A. (2017). Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 39-49). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3140449>
- Zhao, M., An, B., Yu, Y., Liu, S., & Pan, S. J. (2018). Data poisoning attacks on multi-task relationship learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*. Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16073>
- Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2019). Adversarial attacks on deep learning models in natural language processing: A survey. 1(1). *arXiv:1901.06796*. <https://arxiv.org/abs/1901.06796>

## Transparency

- ACM US Public Policy Council. (2017). *Statement on algorithmic transparency and accountability*. Retrieved from [https://www.acm.org/binaries/content/assets/public-policy/2017\\_usacm\\_statement\\_algorithms.pdf](https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf)
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989. Retrieved from <https://journals.sagepub.com/doi/abs/10.1177/1461444816676645>
- Antunes, N., Balby, L., Figueiredo, F., Lourenco, N., Meira, W., & Santos, W. (2018). Fairness and transparency of machine learning for trustworthy cloud services. *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 188–193. <https://doi.org/10.1109/DSN-W.2018.00063>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 205395171562251. <https://doi.org/10.1177/2053951715622512>
- Citron, D. K. (2008). Technological due process. *Washington University Law Review*, 85(6). Retrieved from [https://heinonline.org/hol-cgi-bin/get\\_pdf.cgi?handle=hein.journals/walq85&section=38](https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/walq85&section=38)

- Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89, 1. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/washlr89&div=4&id=&page=&t=1560014586>
- Crawford, K., & Schultz, J. (2014). Big Data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, 55, 93. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/bclr55&div=5&id=&page=&t=1560014537>
- Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16, 18. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/dltr16&div=3&id=&page=&t=1560014649>
- Kemper, J., & Kolkman, D. (2018). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, 1-16. Retrieved from <https://www.tandfonline.com/doi/full/10.1080/1369118X.2018.1477967>
- Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, 11(2), 105–112. <https://doi.org/10.1007/s10676-009-9187-9>
- Weller, A. (2017). Challenges for transparency. arXiv preprint arXiv:1708.01870. Retrieved from <https://arxiv.org/abs/1708.01870>

## Process-Based Governance

- Andrews, L., Benbouzid, B., Brice, J., Bygrave, L. A., Demortain, D., Griffiths, A., ... & Yeung, K. (2017). Algorithmic Regulation. *The London School of Economics and Political Science*. Retrieved from <https://www.kcl.ac.uk/law/research/centres/telos/assets/DP85-Algorithmic-Regulation-Sep-2017.pdf>
- Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Reimer, D., Olteanu, A., Tsay, J., & Varshney, K. R & Piorkowski, D. (2018). FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *arXiv:1808.07261*. Retrieved from <https://arxiv.org/abs/1808.07261>
- Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587-604. Retrieved from [https://www.mitpressjournals.org/doi/abs/10.1162/tacl\\_a\\_00041](https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00041)
- Calo, R. (2017). Artificial Intelligence policy: a primer and roadmap. *UCDL Rev.*, 51, 399. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/davlr51&div=18&id=&page=&t=1560015127>
- D'Agostino, M., & Durante, M. (2018). Introduction: The governance of algorithms. *Philosophy & Technology*, 31(4), 499–505. <https://doi.org/10.1007/s13347-018-0337-z>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv:1803.09010*. Retrieved from <https://arxiv.org/abs/1803.09010>
- Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *arXiv:1805.03677*. Retrieved from <https://arxiv.org/abs/1805.03677>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287596>
- Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., ... & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*, 162(1), W1-W73. Retrieved from <https://annals.org/aim/fullarticle/2088542>
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. *arXiv:1905.06876*. Retrieved from <https://arxiv.org/abs/1905.06876>

- Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: a practical framework for public agency accountability. *AI Now*. Retrieved from <https://ainowinstitute.org/aiareport2018.pdf>
- Saurwein, F., Just, N., & Latzer, M. (2015). Governance of algorithms: options and limitations. *info*, 17(6), 35–49. Retrieved from <https://www.emeraldinsight.com/doi/abs/10.1108/info-05-2015-0025>
- Tutt, A., (2016). An FDA for algorithms. 69 *Admin. L. Rev.* 83 (2017). <http://dx.doi.org/10.2139/ssrn.2747994>
- Wachter, S., & Mittelstadt, B. D. (2018). A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI. *Columbia Business Law Review*. Retrieved from [https://ora.ox.ac.uk/objects/uuid:d53f7b6a-981c-4f87-91bc-743067d10167/download\\_file?file\\_format=pdf&safe\\_filename=Wachter%2Band%2BMittelstadt%2B2018%2B-%2BA%2Bright%2Bto%2Breasonable%2Binferences%2B-%2BVersion%2B6%2Bssrn%2Bversion.pdf&type\\_of\\_work=Journal+article](https://ora.ox.ac.uk/objects/uuid:d53f7b6a-981c-4f87-91bc-743067d10167/download_file?file_format=pdf&safe_filename=Wachter%2Band%2BMittelstadt%2B2018%2B-%2BA%2Bright%2Bto%2Breasonable%2Binferences%2B-%2BVersion%2B6%2Bssrn%2Bversion.pdf&type_of_work=Journal+article)

## Interpretable AI

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8466590>
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1), 8753-8830. Retrieved from <http://www.jmlr.org/papers/volume18/17-716/17-716.pdf>
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7), e0130140. <https://doi.org/10.1371/journal.pone.0130140>
- Bathaei, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, 31(2), 889. Retrieved from <https://www.questia.com/library/journal/1G1-547758123/the-artificial-intelligence-black-box-and-the-failure>
- Bibal, A., & Frénay, B. (2016). *Interpretability of Machine Learning Models and Representations: an Introduction*. Retrieved from [https://www.researchgate.net/profile/Adrien\\_Bibal/publication/326839249\\_Interpretability\\_of\\_Machine\\_Learning\\_Models\\_and\\_Representations\\_an\\_Introduction/links/5b6861caa6fdcc87df6d58e4/Interpretability-of-Machine-Learning-Models-and-Representations-an-Introduction.pdf](https://www.researchgate.net/profile/Adrien_Bibal/publication/326839249_Interpretability_of_Machine_Learning_Models_and_Representations_an_Introduction/links/5b6861caa6fdcc87df6d58e4/Interpretability-of-Machine-Learning-Models-and-Representations-an-Introduction.pdf)
- Bracamonte, V. (2019). *Challenges for transparent and trustworthy machine learning* [Power Point]. KDDI Research, Inc. Retrieved from [https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20190121/Documents/Vanessa\\_Bracamonte\\_Presentation.pdf](https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20190121/Documents/Vanessa_Bracamonte_Presentation.pdf)
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>
- Card, D. (2017). The “black box” metaphor in machine learning. *Towards Data Science*. Retrieved from <https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0>
- Caruana, R., Kangaroo, H., Dionisio, J. D., Sinha, U., & Johnson, D. (1999). Case-based explanation of non-case-based learning methods. *Proceedings. AMIA Symposium*, 212–215. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232607/>
- Chen, C., Li, O., Tao, C., Barnett, A., Su, J., & Rudin, C. (2018). This looks like that: deep learning for interpretable image recognition. *arXiv:1806.10574*. Retrieved from <https://arxiv.org/abs/1806.10574>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*. Retrieved from <https://arxiv.org/abs/1702.08608>
- Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., ... & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv:1711.01134*. Retrieved from <https://arxiv.org/abs/1711.01134>

- Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. <https://doi.org/10.23919/MIPRO.2018.8400040>
- Eisenstadt, V., & Althoff, K. (2018). A Preliminary Survey of Explanation Facilities of AI-Based Design Support Approaches and Tools. *LWDA*. Presented at the LWDA. [https://www.researchgate.net/profile/Viktor\\_Eisenstadt/publication/327339350\\_A\\_Preliminary\\_Survey\\_of\\_Explanation\\_Facilities\\_of\\_AI-Based\\_Design\\_Support\\_Approaches\\_and\\_Tools/links/5b891ecd299bf1d5a7338b1a/A-Preliminary-Survey-of-Explanation-Facilities-of-AI-Based-Design-Support-Approaches-and-Tools.pdf](https://www.researchgate.net/profile/Viktor_Eisenstadt/publication/327339350_A_Preliminary_Survey_of_Explanation_Facilities_of_AI-Based_Design_Support_Approaches_and_Tools/links/5b891ecd299bf1d5a7338b1a/A-Preliminary-Survey-of-Explanation-Facilities-of-AI-Based-Design-Support-Approaches-and-Tools.pdf)
- Feldmann, F. (2018). *Measuring machine learning model interpretability*. Retrieved from [https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/860270201/felix\\_feldmann\\_eml2018\\_report.pdf](https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/860270201/felix_feldmann_eml2018_report.pdf)
- Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3429-3437). Retrieved from [http://openaccess.thecvf.com/content\\_iccv\\_2017/html/Fong\\_Interpretable\\_Explanations\\_of\\_ICCV\\_2017\\_paper.html](http://openaccess.thecvf.com/content_iccv_2017/html/Fong_Interpretable_Explanations_of_ICCV_2017_paper.html)
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. Explaining explanations: An approach to evaluating interpretability of machine. *arXiv:1806.00069*. Retrieved from <https://arxiv.org/abs/1806.00069>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 93. Retrieved from <https://dl.acm.org/citation.cfm?id=3236009>
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics*. <https://doi.org/10.1093/qje/qjx032>
- Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), 20180084. <https://doi.org/10.1098/rsta.2018.0084>
- Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1675-1684). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2939874>
- Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 275–284. <https://doi.org/10.1145/3097983.3098066>
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4), 611-627. <https://doi.org/10.1007/s13347-017-0279-x>
- Li, O., Liu, H., Chen, C., & Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*. Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/17082>
- Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv:1606.03490*. Retrieved from <https://arxiv.org/abs/1606.03490>
- Lipton, Z. C., & Steinhardt, J. (2018). Troubling trends in machine learning scholarship. *arXiv:1807.03341*. Retrieved from <https://arxiv.org/abs/1807.03341>
- Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13*, 623. <https://doi.org/10.1145/2487575.2487579>
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *ArXiv:1705.07874*. Retrieved from <http://arxiv.org/abs/1705.07874>
- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287574>

- Molnar, C. (2018). Interpretable machine learning. A guide for making black box models explainable. *Leanpub*. Retrieved from <https://christophm.github.io/interpretable-ml-book/>
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. *arXiv:1901.04592*. Retrieved from <https://arxiv.org/abs/1901.04592>
- Olhede, S. C., & Wolfe, P. J. (2018). The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128). <https://doi.org/10.1098/rsta.2017.0364>
- Park, D. H., Hendricks, L. A., Akata, Z., Schiele, B., Darrell, T., & Rohrbach, M. (2016). Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv:1612.04757*. Retrieved from <https://arxiv.org/abs/1612.04757>
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S., & Turini, F. (2018). Open the black box data-driven explanation of black box decision systems. *arXiv:1806.09936*. Retrieved from <https://arxiv.org/abs/1806.09936>
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. *AAAI Press*.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability. *ArXiv:1802.07810*. Retrieved from <http://arxiv.org/abs/1802.07810>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv:1606.05386*. Retrieved from <https://arxiv.org/abs/1606.05386>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM. Retrieved from <https://dl.acm.org/citation.cfm?Id=2939778>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Thirty-second AAAI Conference on Artificial Intelligence*. Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16982>
- Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *arXiv:1811.10154*. Retrieved from <https://arxiv.org/abs/1811.10154>
- Rudin, C., & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5), 449-466. <https://doi.org/10.1287/inte.2018.0957>
- Shmueli, G. (2010). To explain or to predict?. *Statistical science*, 25(3), 289-310. Retrieved from <https://projecteuclid.org/euclid.ss/1294167961>
- Shaywitz, D. (2018). AI doesn't ask why – But physicians and drug developers want to know. *Forbes*. Retrieved from <https://www.forbes.com/sites/davidshaywitz/2018/11/09/ai-doesnt-ask-why-but-physicians-and-drug-developers-want-to-know/>
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *ArXiv:1704.02685*. Retrieved from <http://arxiv.org/abs/1704.02685>
- Simonite, T. (2017). AI experts want to end "black box" algorithms in government. *Wired Business*, 10, 17. Retrieved from <https://www.wired.com/story/ai-experts-want-to-end-black-box-algorithms-in-government/>
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *ArXiv:1312.6034*. Retrieved from <http://arxiv.org/abs/1312.6034>
- Sokol, K., & Flach, P. (2018). Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 5868–5870. <https://doi.org/10.24963/ijcai.2018/865>
- Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3), 349-391. Retrieved from: <https://link.springer.com/article/10.1007/s10994-015-5528-6>
- Zhang, Q., & Zhu, S. (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1), 27–39. <https://doi.org/10.1631/FITEE.1700808>

## Responsible delivery through human-centred implementation protocols and practices

- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 582). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3174156>
- Antaki, C., & Leudar, I. (1992). Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2), 181-194. <https://doi.org/10.1002/ejsp.2420220206>
- Arioua, A., & Croitoru, M. (2015). Formalizing explanatory dialogues. In *International Conference on Scalable Uncertainty Management* (pp. 282-297). Springer, Cham. [https://doi.org/10.1007/978-3-319-23540-0\\_19](https://doi.org/10.1007/978-3-319-23540-0_19)
- Bex, F., & Walton, D. (2016). Combining explanation and argumentation in dialogue. *Argument & Computation*, 7(1), 55-68. Retrieved from <https://content.iospress.com/articles/argument-and-computation/aac001>
- Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (Vol. 8). Retrieved from [http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17\\_XAI\\_WS\\_Proceedings.pdf#page=8](http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17_XAI_WS_Proceedings.pdf#page=8)
- Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. *arXiv:1901.03729*. Retrieved from <https://arxiv.org/abs/1901.03729>
- Ginet, C. (2008). In defense of a non-causal account of reasons explanations. *The Journal of Ethics*, 12(3-4), 229-237. <https://doi.org/10.1007/s10892-008-9033-z>
- Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., ... & Holzinger, A. (2018). Explainable AI: the new 42?. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 295-303). Springer, Cham. [https://doi.org/10.1007/978-3-319-99740-7\\_21](https://doi.org/10.1007/978-3-319-99740-7_21)
- Habermas, J. (1993). Remarks on discourse ethics. *Justification and application: Remarks on discourse ethics*, 44, 313-314. Cambridge, UK: Polity Press.
- Habermas, J. (2003). Rightness versus truth: on the sense of normative validity in moral judgments and norms. *Truth and justification*, 248. Cambridge, UK: Polity Press.
- Hoffman, R. R., Mueller, S. T., & Klein, G. (2017). Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems*, 32(4), 78-86. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8012316>
- Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). A Grounded Interaction Protocol for Explainable Artificial Intelligence. *arXiv:1903.02409*. Retrieved from <https://arxiv.org/abs/1903.02409>
- McCarthy, T. (1974). The operation called Verstehen: Towards a redefinition of the problem. In *PSA 1972* (pp. 167-193). Springer, Dordrecht. [https://doi.org/10.1007/978-94-010-2140-1\\_12](https://doi.org/10.1007/978-94-010-2140-1_12)
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Rapanta, C., & Walton, D. (2016). The use of argument maps as an assessment tool in higher education. *International Journal of Educational Research*, 79, 211-221. <https://doi.org/10.1016/j.ijer.2016.03.002>
- Springer, A., & Whittaker, S. (2018). Progressive disclosure: Designing for effective transparency. *arXiv:1811.02164*. Retrieved from <https://arxiv.org/abs/1811.02164>
- Taylor, C. (1973). Interpretation and the sciences of man. In *Explorations in Phenomenology* (pp. 47-101). Springer, Dordrecht. [https://doi.org/10.1007/978-94-010-1999-6\\_3](https://doi.org/10.1007/978-94-010-1999-6_3)
- Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv:1806.07552*. Retrieved from <https://arxiv.org/abs/1806.07552>

- Tsai, C. H., & Brusilovsky, P. (2019). Designing explanation interfaces for transparency and beyond. In *Joint Proceedings of the ACM IUI 2019 Workshops*. Retrieved from <http://ceur-ws.org/Vol-2327/IUI19WS-IUIATEC-4.pdf>
- Von Wright, G. H. (2004). *Explanation and understanding*. Ithaca, NY: Cornell University Press.
- Walton, D. (2004). A new dialectical theory of explanation. *Philosophical Explorations*, 7(1), 71-89. <https://doi.org/10.1080/1386979032000186863>
- Walton, D. (2005). Dialectical Explanation in AI. *Argumentation Methods for Artificial Intelligence in Law*, 173-212. [https://doi.org/10.1007/3-540-27881-8\\_6](https://doi.org/10.1007/3-540-27881-8_6)
- Walton, D. (2007). Dialogical Models of Explanation. *ExaCt*, 2007, 1-9. Retrieved from <https://www.aaai.org/Papers/Workshops/2007/WS-07-06/WS07-06-001.pdf>
- Walton, D. (2011). A dialogue system specification for explanation. *Synthese*, 182(3), 349-374. <https://doi.org/10.1007/s11229-010-9745-z>
- Walton, D. (2016). Some artificial intelligence tools for argument evaluation: An introduction. *Argumentation*, 30(3), 317-340. <https://doi.org/10.1007/s10503-015-9387-x>
- Weld, D. S., & Bansal, G. (2018). The challenge of crafting intelligible intelligence. *arXiv:1803.04263*. Retrieved from <https://arxiv.org/abs/1803.04263>
- Walton, D., Toniolo, A., & Norman, T. (2016). Speech acts and burden of proof in computational models of deliberation dialogue. In *Proceedings of the First European Conference on Argumentation*, ed. D. Mohammed and M. Lewinski, London, College Publications (Vol. 1, pp. 757-776). Retrieved from [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2852054](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852054)
- Wendt, A. (1998). On constitution and causation in international relations. *Review of international studies*, 24(5), 101-118. <https://doi.org/10.1017/S0260210598001028>
- Winikoff, M. (2017). Debugging agent programs with why?: Questions. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (pp. 251-259). International Foundation for Autonomous Agents and Multiagent Systems. Retrieved from <https://dl.acm.org/citation.cfm?id=3091166>
- Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G. M. (2018). Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)* (pp. 1-8). IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8490433>

### Individual and societal impacts of machine learning and algorithmic systems

- Amoore, L. (2018a). Cloud geographies: Computing, data, sovereignty. *Progress in Human Geography*, 42(1), 4-24. <https://doi.org/10.1177/0309132516662147>
- Amoore, L. (2018b). Doubtful algorithms: of machine learning truths and partial accounts. *Theory, culture & society*. Retrieved from <http://dro.dur.ac.uk/26913/1/26913.pdf>
- Amoore, L., & Raley, R. (2017). Securing with algorithms: Knowledge, decision, sovereignty. *Security Dialogue*, 48(1), 3-10. <https://doi.org/10.1177/0967010616680753>
- Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, 41(1), 93-117. <https://doi.org/10.1177/0162243915606523>
- Anderson, B. (2010). Preemption, precaution, preparedness: Anticipatory action and future geographies. *Progress in Human Geography*, 34(6), 777-798. <https://doi.org/10.1177/0309132510362600>
- Anderson, B. (2010). Security and the future: Anticipating the event of terror. *Geoforum*, 41(2), 227-235. <https://doi.org/10.1016/j.geoforum.2009.11.002>
- Anderson, S. F. (2017). *Technologies of vision: The war between data and images*. MIT Press.

- Arnoldi, J. (2016). Computer algorithms, market manipulation and the institutionalization of high frequency trading. *Theory, Culture & Society*, 33(1), 29-52. <https://doi.org/10.1177/0263276414566642>
- Beer, D. (2013). Algorithms: Shaping tastes and manipulating the circulations of popular culture. In *Popular Culture and New Media* (pp. 63-100). Palgrave Macmillan, London. [https://doi.org/10.1057/9781137270061\\_4](https://doi.org/10.1057/9781137270061_4)
- Beer, D. (2017). The social power of algorithms. In *Information, Communication & Society*, (20), 1-13. <https://doi.org/10.1080/1369118X.2016.1216147>
- Bodó, B., Helberger, N., Irion, K., Zuiderveen Borgesius, F., Moller, J., van de Velde, B., ... & de Vreese, C. (2017). Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents. *Yale JL & Tech.*, 19, 133. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/yjolt19&div=4&id=&page=&t=1560029464>
- Bogost, I. (2015). The cathedral of computation. *The Atlantic*, 15. Retrieved from <https://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/>
- Bolin, G., & Andersson Schwarz, J. (2015). Heuristics of the algorithm: Big Data, user interpretation and institutional translation. *Big Data & Society*, 2(2). <https://doi.org/10.1177/2053951715608406>
- Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *NIPS*. Retrieved from <http://papers.nips.cc/paper/6227-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings>
- Browne, S. (2015). *Dark matters: On the surveillance of blackness*. Duke University Press.
- Bucher, T. (2017). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1), 30-44. <https://doi.org/10.1080/1369118X.2016.1154086>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186. Retrieved from <https://science.sciencemag.org/content/356/6334/183>
- Cheney-Lippold, J. (2011). A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society*, 28(6), 164-181. <https://doi.org/10.1177/0263276411424420>
- Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance & Society*, 15(5), 609-625. <https://doi.org/10.24908/ss.v15i5.6433>
- Crandall, J. (2006). Precision + guided + seeing. *CTheory*, 1-10. Retrieved from <https://journals.uvic.ca/index.php/ctheory/article/view/14468/5310>
- Crandall, J. (2010). The geospatialization of calculative operations: Tracking, sensing and megacities. *Theory, Culture & Society*, 27(6), 68-90. <https://doi.org/10.1177/0263276410382027>
- Crawford, K. (2014). The anxieties of Big Data. *The New Inquiry*, 30, 2014. Retrieved from <https://thenewinquiry.com/the-anxieties-of-big-data/>
- Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature News*, 538(7625), 311. Retrieved from <https://www.nature.com/news/there-is-a-blind-spot-in-ai-research-1.20805>
- Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2019). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, 46(2), 185-209. Retrieved from <https://journals.sagepub.com/doi/abs/10.1177/0093854818811379>
- Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., ... & Sandvig, C. (2015). I always assumed that I wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 153-162). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2702556>
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
- Ferguson, A. G. (2017). Policing Predictive Policing. *Washington University Law Review*, 94(5). Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/walq94&div=35&id=&page=&t=1559934122>

- Geiger, R. S. (2014). Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society*, 17(3), 342-356. <https://doi.org/10.1080/1369118X.2013.873069>
- Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society*. Cambridge, MA: The MIT Press.
- Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, 3(2). <https://doi.org/10.1177/2053951716674238>
- Jasanoff, S. (2015). Future imperfect: Science, technology, and the imaginations of modernity. In S. Jasanoff & S. Kim (Eds.), *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. Chicago, IL: The University of Chicago Press.
- Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14-29. <https://doi.org/10.1080/1369118X.2016.1154087>
- Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *ArXiv:1805.04508*. Retrieved from <http://arxiv.org/abs/1805.04508>
- Kushner, S. (2013). The freelance translation machine: Algorithmic culture and the invisible industry. *New Media & Society*, 15(8), 1241-1258. <https://doi.org/10.1177/1461444812469597>
- Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2016). The tyranny of data? The bright and dark sides of data-driven decision-making for social good. *ArXiv:1612.00323*. Retrieved from <http://arxiv.org/abs/1612.00323>
- Mackenzie, A. (2015a). Machine learning and genomic dimensionality: From features to landscapes. In S. Richardson & H. Stevens (Eds.), *Postgenomics: Perspectives on Biology after the Genome*. Durham, NC: Duke University Press.
- Mackenzie, A. (2015b). The production of prediction: What does machine learning want?. *European Journal of Cultural Studies*, 18(4-5), 429-445. <https://doi.org/10.1177/1367549415577384>
- Mackenzie, A., & McNally, R. (2013). Living multiples: How large-scale scientific data-mining pursues identity and differences. *Theory, Culture & Society*, 30(4), 72-91. <https://doi.org/10.1177/0263276413476558>
- Mackenzie, A., & Vurdubakis, T. (2011). Codes and codings in crisis: signification, performativity and excess. *Theory, Culture & Society*, 28(6), 3-23. <https://doi.org/10.1177/0263276411424761>
- Mager, A. (2012). Algorithmic ideology: How capitalist society shapes search engines. *Information, Communication & Society*, 15(5), 769-787. <https://doi.org/10.1080/1369118X.2012.676056>
- Manokha, I. (2018). Surveillance, panopticism, and self-discipline in the digital age. *Surveillance & Society*, 16(2), 219-237. <https://doi.org/10.24908/ss.v16i2.8346>
- Matzner, T. (2014). Why privacy is not enough privacy in the context of “ubiquitous computing” and “Big Data.” *Journal of Information, Communication and Ethics in Society*, 12(2), 93–106. <https://doi.org/10.1108/JICES-08-2013-0030>
- Mendoza, I., & Bygrave, L. A. (2017). The right not to be subject to automated decisions based on profiling. In *EU Internet Law* (pp. 77-98). Springer, Cham. Retrieved from [https://link.springer.com/chapter/10.1007/978-3-319-64955-9\\_4](https://link.springer.com/chapter/10.1007/978-3-319-64955-9_4)
- Mollicchi, S. (2017). Flatness versus depth: A study of algorithmically generated camouflage. *Security Dialogue*, 48(1), 78-94. <https://doi.org/10.1177/0967010616650227>
- Molnar, P., & Gill, L. (2018). Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada’s Immigration and Refugee System. *Citizen Lab and International Human Rights Program (Faculty of Law, University of Toronto)*. Retrieved from <https://tspace.library.utoronto.ca/handle/1807/94802>
- Monahan, T. (2018). Algorithmic fetishism. *Surveillance & Society*, 16(1), 1-5. <https://doi.org/10.24908/ss.v16i1.10827>
- Murphy, M. H. (2017). Algorithmic surveillance: The collection conundrum. *International Review of Law, Computers & Technology*, 31(2), 225–242. <https://doi.org/10.1080/13600869.2017.1298497>
- Napoli, P. M. (2014). Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication Theory*, 24(3), 340-360. <https://doi.org/10.1111/comt.12039>

- Neyland, D. (2015). On organizing algorithms. *Theory, Culture & Society*, 32(1), 119-132.  
<https://doi.org/10.1177/0263276414530477>
- Neyland, D. (2016). Bearing account-able witness to the ethical algorithmic system. *Science, Technology, & Human Values*, 41(1), 50-76. <https://doi.org/10.1177/0162243915598056>
- Neyland, D., & Möllers, N. (2017). Algorithmic IF... THEN rules and the conditions and consequences of power. *Information, Communication & Society*, 20(1), 45-62. <https://doi.org/10.1080/1369118X.2016.1156141>
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. nyu Press.
- O'Grady, N. (2015). A politics of redeployment: malleable technologies and the localisation of anticipatory calculation. In *Algorithmic Life* (pp. 86-100). Routledge. Retrieved from <http://eprints.uwe.ac.uk/id/eprint/33134>
- Plantin, J. C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, 20(1), 293-310.  
<https://doi.org/10.1177/1461444816661553>
- Redden, J., & Brand, J. (2017). Data Harm Record. *Data Justice Lab*. Retrieved from <http://orca-mwe.cf.ac.uk/107924/1/data-harm-record-djl2.pdf>
- Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online, Forthcoming*. Retrieved from [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3333423](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423)
- Roberge, J., & Seyfert, R. (2016). What are algorithmic cultures. *Algorithmic cultures: Essays on meaning, performance and new technologies*, 1-25. Retrieved from <https://www.taylorfrancis.com/books/e/9781315658698/chapters/10.4324/9781315658698-7>
- Schüll, N. D. (2018). Self in the Loop: Bits, Patterns, and Pathways in the Quantified Self. In *A Networked Self and Human Augmentics, Artificial Intelligence, Sentience* (pp. 41-54). New York, NY: Routledge.
- Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, 87, 1085. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/flr87&div=44&id=&page=&t=1560020999>
- Smith, G. (2018). High-tech redlining: AI is quietly upgrading institutional racism. *Fast Company*. Retrieved from <https://www.fastcompany.com/90269688/high-tech-redlining-ai-is-quietly-upgrading-institutional-racism>
- Striphas, T. (2015). Algorithmic culture. *European Journal of Cultural Studies*, 18(4-5), 395-412.  
<https://doi.org/10.1177/1367549415577392>
- Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), 197-208. <https://doi.org/10.24908/ss.v12i2.4776>
- Wilf, E., Cheney-Lippold, J., Duranti, A., Eisenlohr, P., Gershon, I., Mackenzie, A., ... & Wilf, E. (2013). Toward an anthropology of computer-mediated, algorithmic forms of sociality. *Current Anthropology*, 54(6), 000-000. Retrieved from <https://www.journals.uchicago.edu/doi/abs/10.1086/673321>
- Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society*, 20(1), 137-150.  
<https://doi.org/10.1080/1369118X.2016.1200645>
- Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1), 118-132. <https://doi.org/10.1177/0162243915605575>
- Ziewitz, M. (2016). Governing algorithms: Myth, mess, and methods. *Science, Technology, & Human Values*, 41(1), 3-16. <https://doi.org/10.1177/0162243915608948>
- Zuboff, S. (2019). *The age of surveillance capitalism: the fight for the future at the new frontier of power*. Profile Books.



**turing.ac.uk**  
**@turinginst**