

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

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.

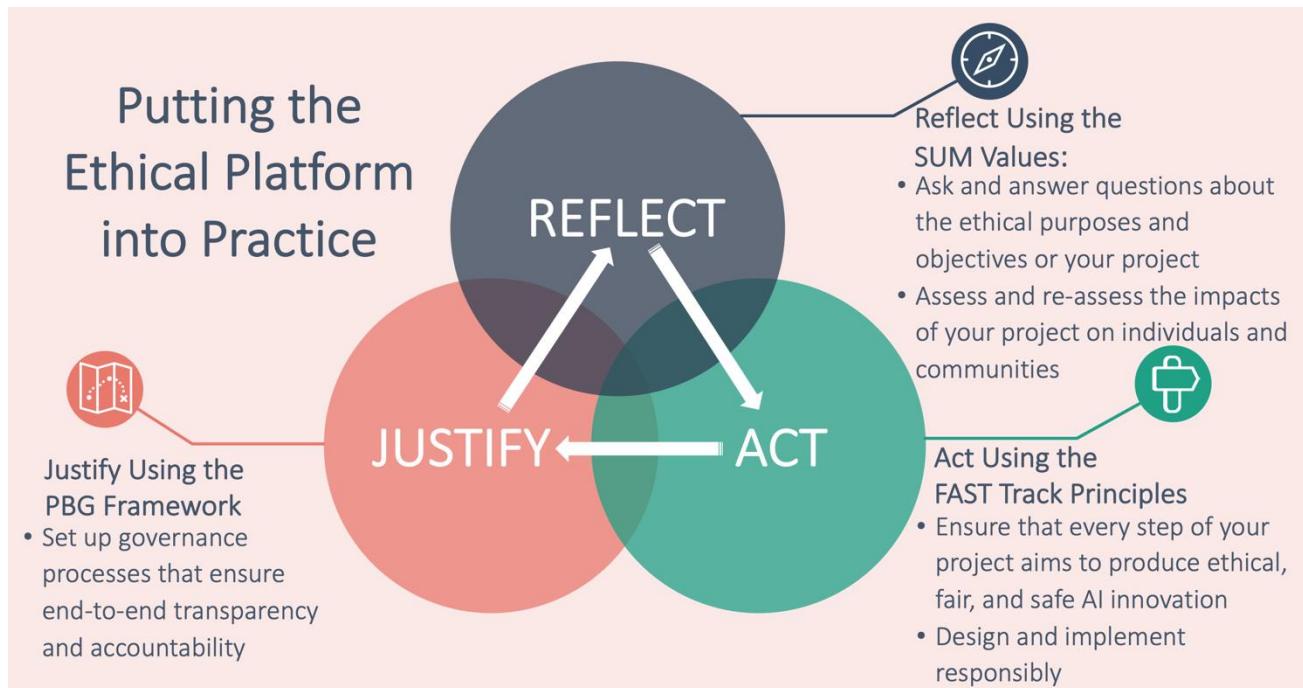
Here is a summary visualisation of these three building blocks of the platform:

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.

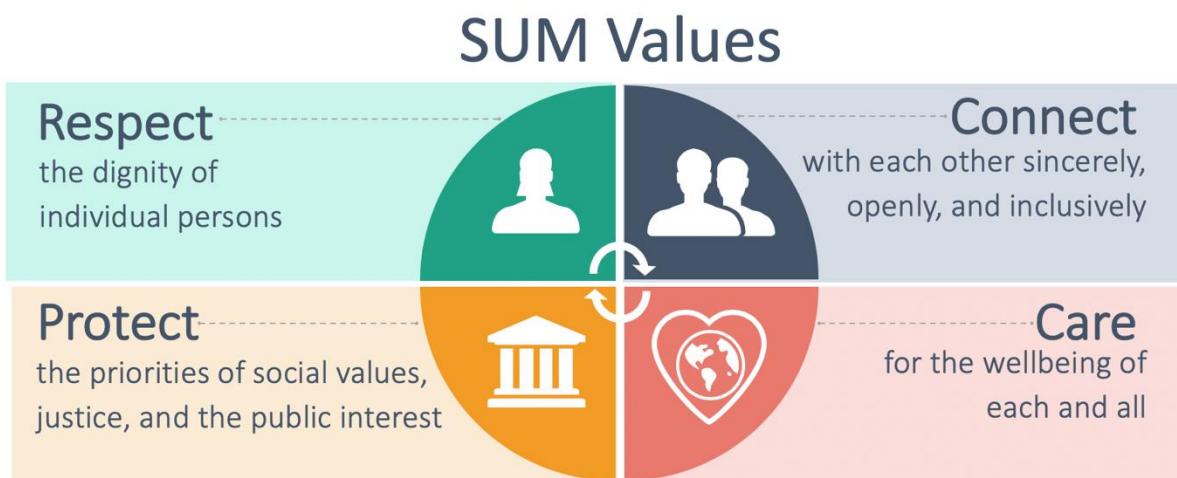
Whereas bioethics largely stresses the normative values that underlie the safeguarding of **individuals** in instances where technological practices affect their interests and wellbeing, human rights discourse mainly focuses on the set of **social, political, and legal entitlements** that are due to 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.

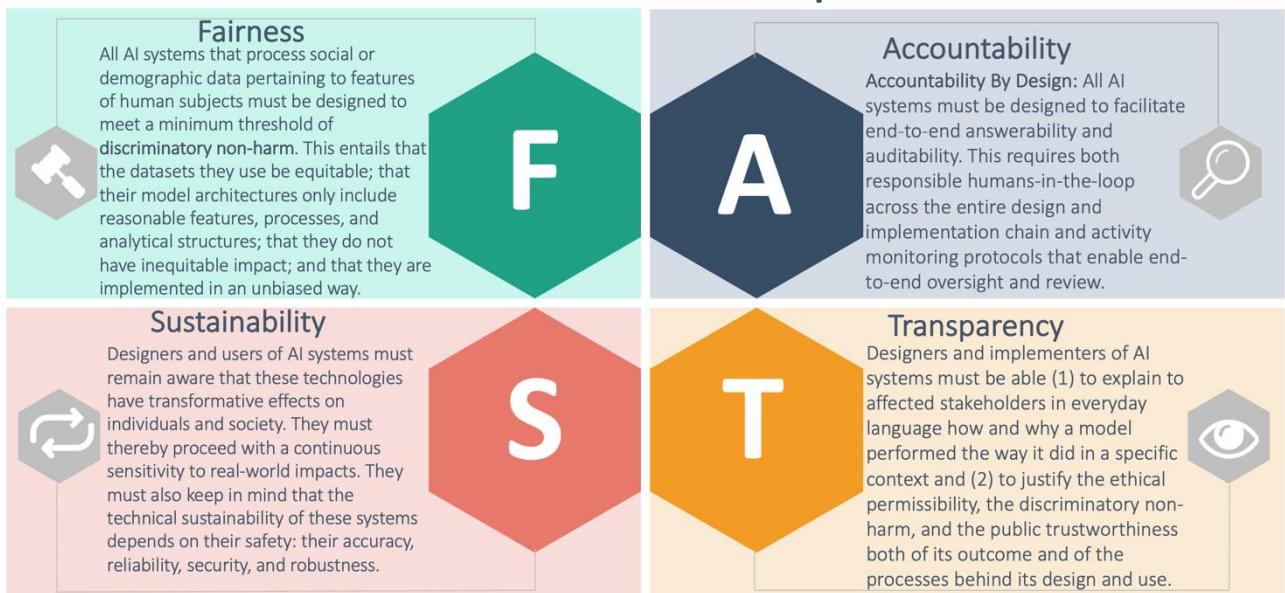
The FAST Track Principles: Fairness, Accountability, Sustainability, and Transparency

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
Demographic/ Statistical Parity	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.
True Positive Rate Parity	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.
False Positive Rate Parity	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.
Positive Predictive Value Parity	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.
Individual Fairness	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.
Counterfactual Fairness	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/ (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:

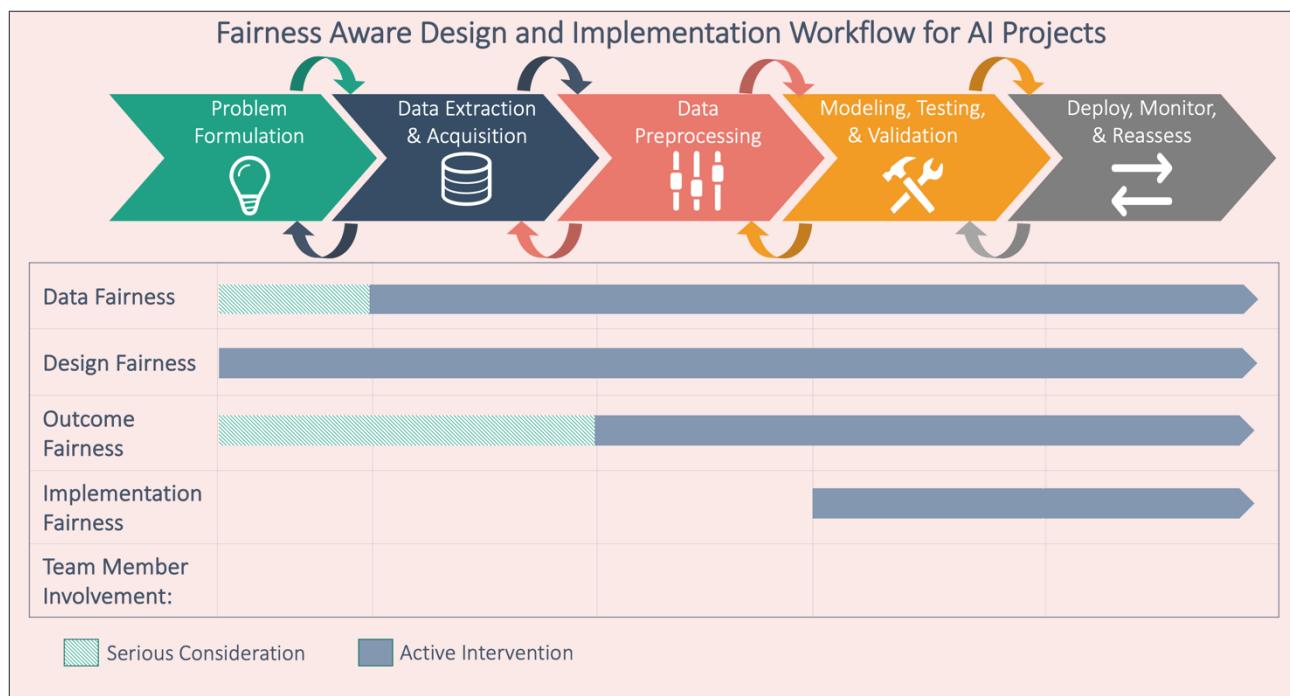
- Training of implementers should include the conveyance of basic knowledge about the statistical and probabilistic character of machine learning and about the limitations of AI and automated decision-support technologies. This training should avoid any anthropomorphic

(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 1st 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 2nd 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 3rd section should provide answers to the additional questions relevant to pre-implementation evaluation. The 4th 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:

<u>Stakeholder Impact Assessment for (Project Name)</u>	
<p>1. Alpha Phase (Problem Formulation) General Questions</p> <p>Completed on this Date:</p>	<p>I. Identifying Affected Stakeholders</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>II. Goal-Setting and Objective-Mapping</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>III. Possible Impacts on the Individual</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>IV. Possible Impacts on Society and Interpersonal Relationships</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>2. Alpha Phase (Problem Formulation) Sector-Specific and Use Case-Specific Questions</p> <p>Completed on this Date:</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>3. From Alpha to Beta (Pre-Implementation)</p> <p>Completed on this Date:</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>Dates of Public Consultation on Beta-Phase Impacts:</p> <p>Date of Planned Beta Phase Re-Assessment:</p>
<p>4. Beta Phase (Re-Assessment)</p> <p>Completed on this Date:</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>Dates of Public Consultation on Beta-Phase Impacts:</p> <p>Date of Next Planned Beta Phase Re-Assessment:</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:

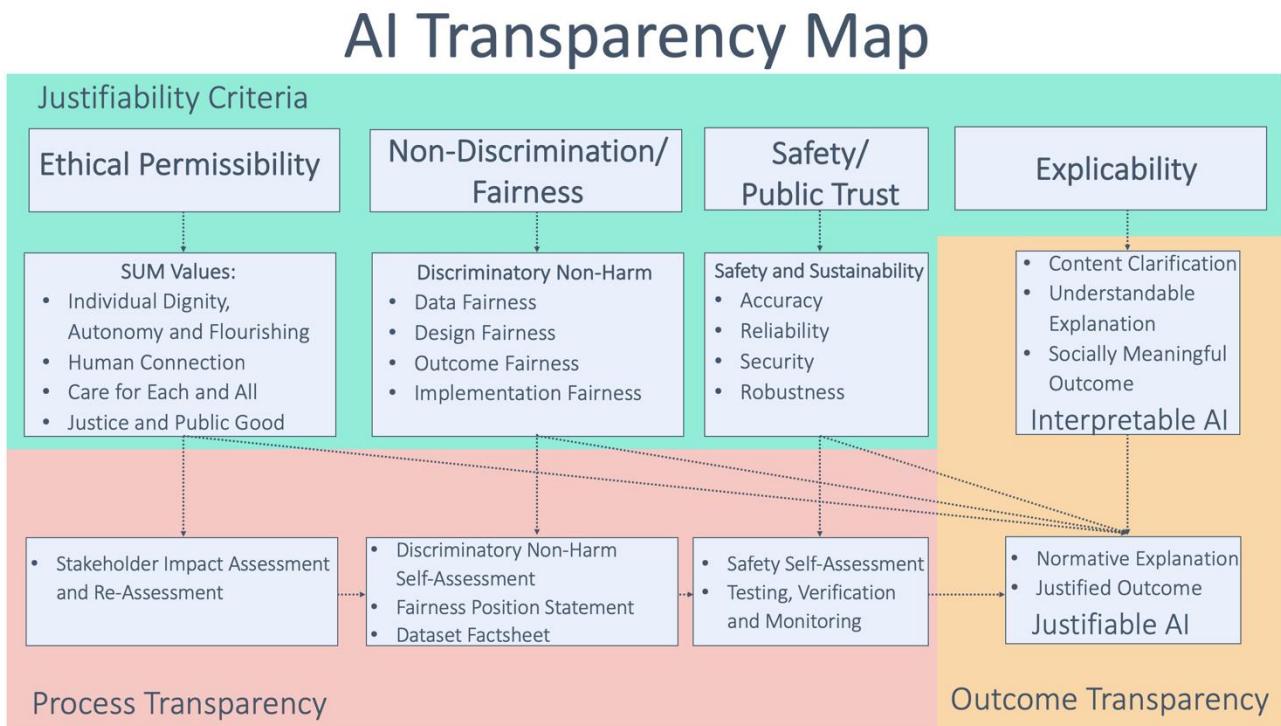
- **Process Transparency, Task 1: Justify Process.** In offering an explanation to affected stakeholders, you should be able to demonstrate that considerations of ethical permissibility, non-discrimination/fairness, and safety/public trustworthiness were operative end-to-end in the design and implementation 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.

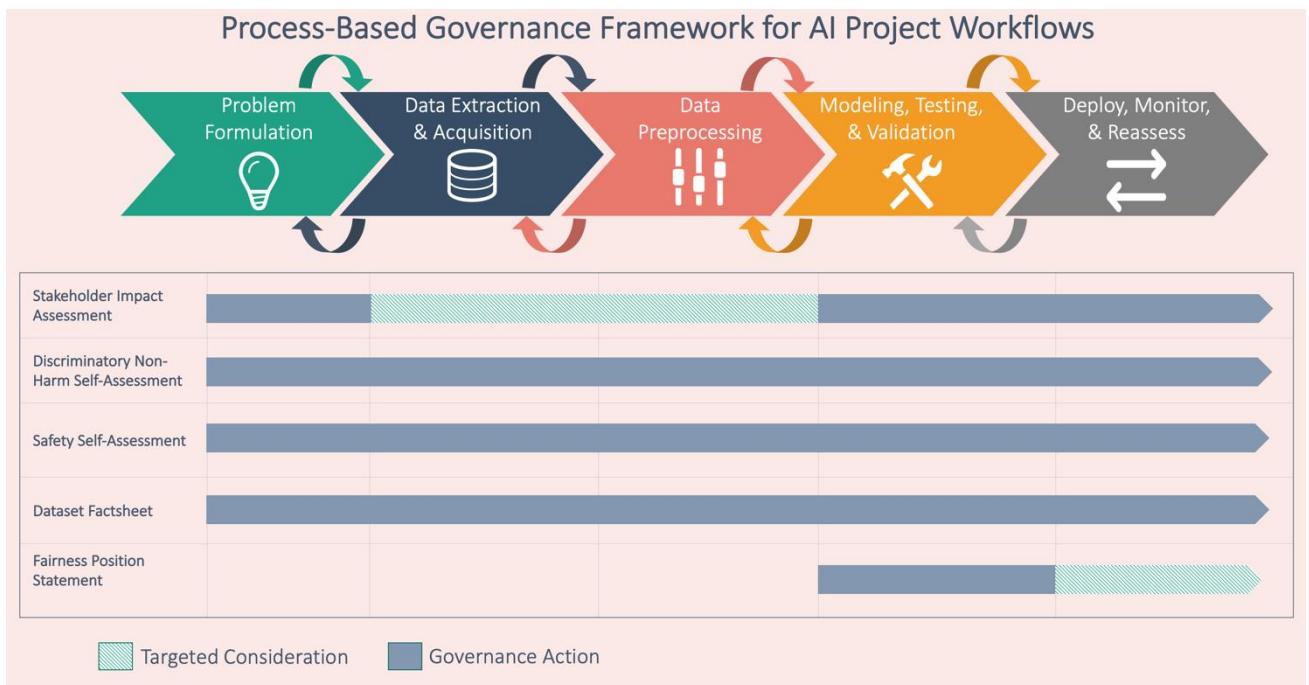
2. **Process-Based Governance Framework:** So far, this guide has presented some of the main steps that are necessary for establishing responsible innovation practices in your AI project. Perhaps the most vital of these measures is the effective operationalisation of the values and principles that underpin 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:

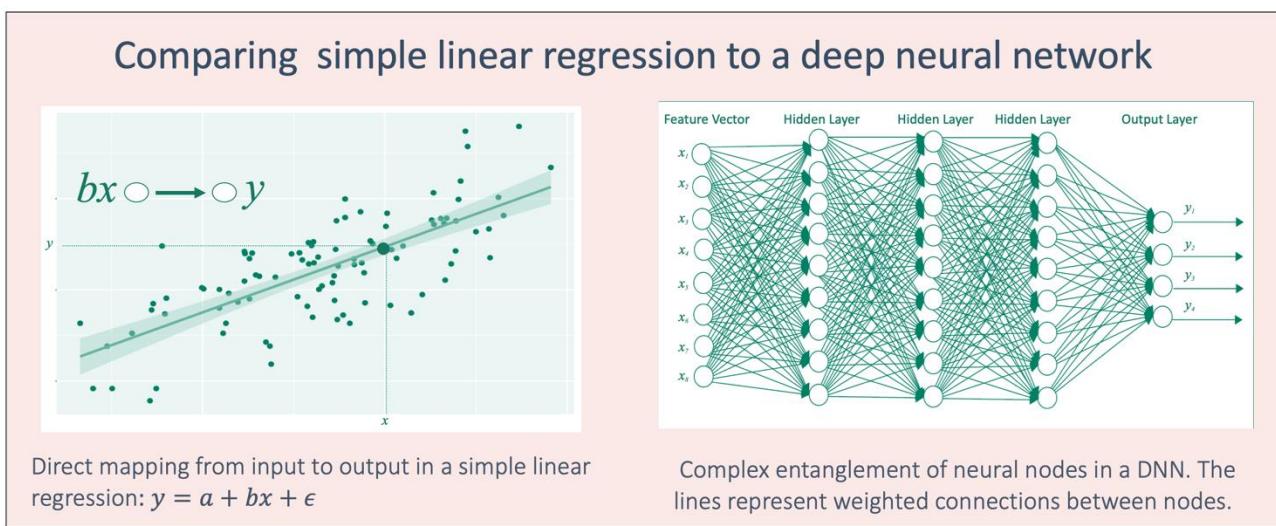
- **Linearity:** Any change in the value of the predictor variable is directly reflected in a change in the value of the response variable at a constant rate b . The interpretable prediction yielded by the model can therefore be directly inferred. This linearity dimension of predictive models has been 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.

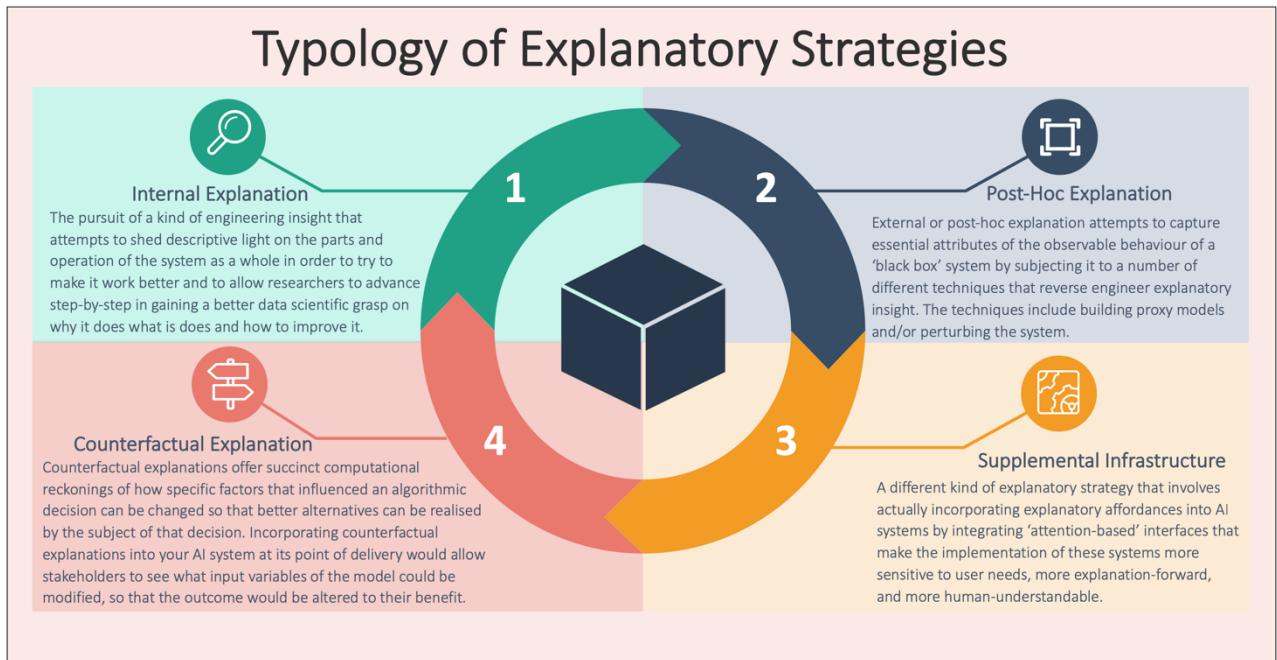
- d) ***Counterfactual explanation:*** While counterfactual explanation is a kind of post-hoc approach, it deserves special attention insofar as it moves beyond other post-hoc explanations to provide affected stakeholders with clear and precise options for actionable recourse and practical remedy.

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.

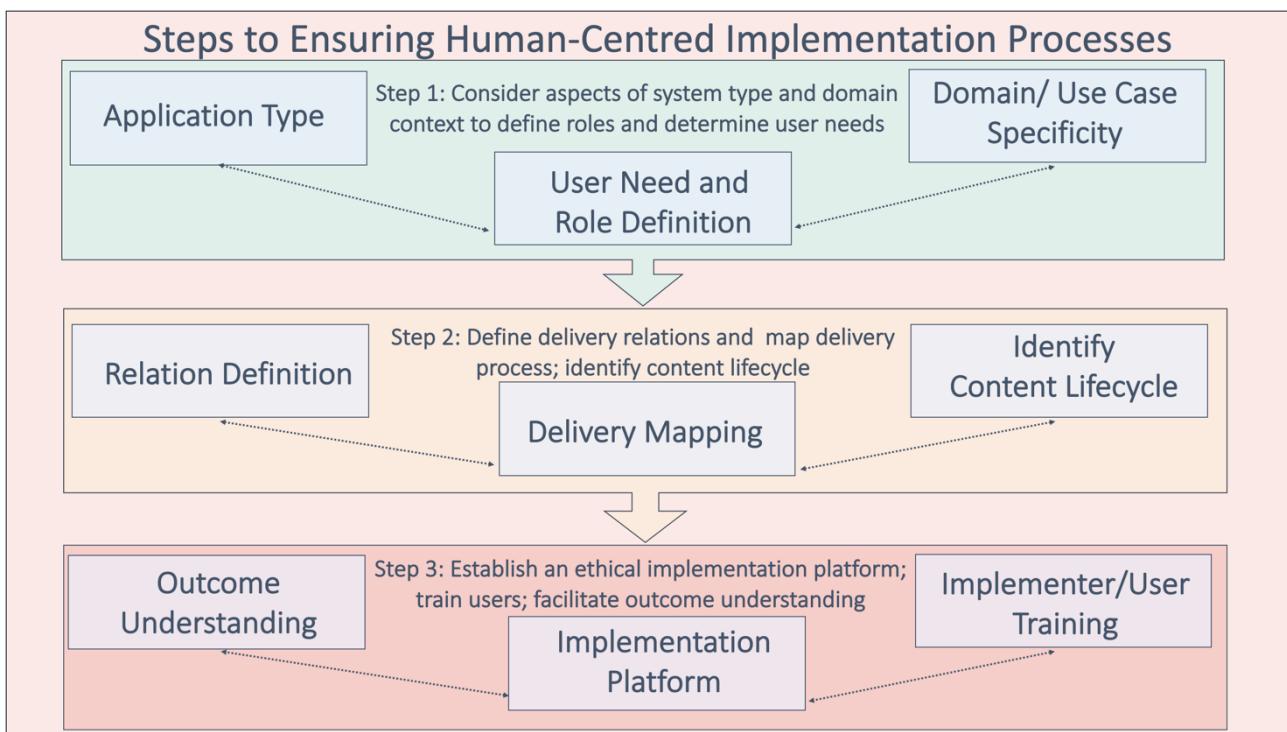
Seeing interpretability as a continuum of comprehensibility that is dependent on the capacities and limits of the individual human interpreter should key you in to what is needed in order to deliver an interpretable AI system. Such limits to consider should include not only cognitive

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.

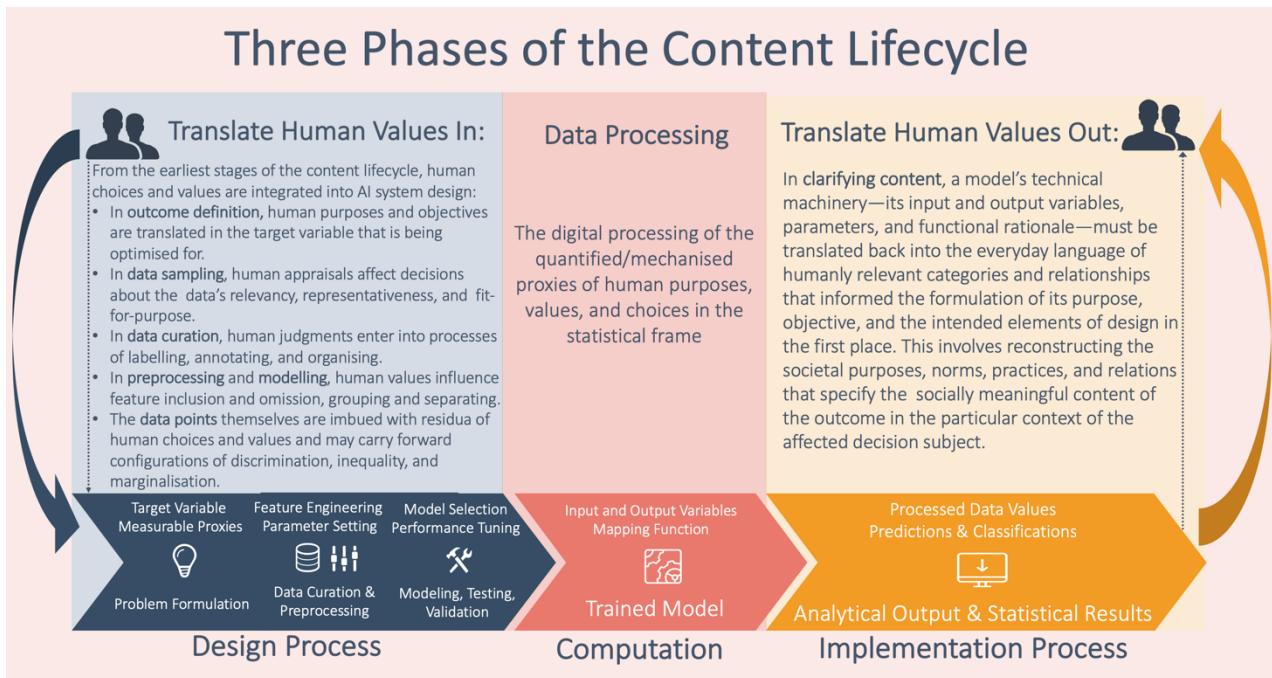
Human choices and values therefore punctuate the design and implementation of AI systems. These choices and values are inscribed in algorithmic models:

- 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 in to 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 to 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

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
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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., Leboda, 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
- 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
- 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
- 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
- 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
- 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§ion=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
- 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
- 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/fd6f1a9dc6eaac7cc3100a16cca9750ff9d8.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
- 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>
- Crocoll, 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
- 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.wvl.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.obhd.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
- 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

- 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
- 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
- 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/ipx005>
- 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/misc/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/misc/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

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/@deephindsafetyresearch/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
- 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§ion=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
- 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

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
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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
- 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=155993_4122
- 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
- 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
- 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., & Baracas, 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