

# The Alan Turing Institute

---

## Understanding artificial intelligence ethics and safety

A guide for the responsible  
design and implementation of AI  
systems in the public sector

Dr David Leslie  
Public Policy Programme



# The Alan Turing Institute

The Public Policy Programme at The Alan Turing Institute was set up in May 2018 with the aim of developing research, tools, and techniques that help governments innovate with data-intensive technologies and improve the quality of people's lives. We work alongside policy makers to explore how data science and artificial intelligence can inform public policy and improve the provision of public services. We believe that governments can reap the benefits of these technologies only if they make considerations of ethics and safety a first priority.

This document provides end-to-end guidance on how to apply principles of AI ethics and safety to the design and implementation of algorithmic systems in the public sector. We will shortly release a workbook to bring the recommendations made in this guide to life. The workbook will contain case studies highlighting how the guidance contained here can be applied to concrete AI projects. It will also contain exercises and practical tools to help strengthen the process-based governance of your AI project.

Please note, that this guide is a living document that will evolve and improve with input from users, affected stakeholders, and interested parties. We need your participation. Please share feedback with us at [policy@turing.ac.uk](mailto:policy@turing.ac.uk)

This work was supported exclusively by the Turing's Public Policy Programme. All research undertaken by the Turing's Public Policy Programme is supported entirely by public funds.

<https://www.turing.ac.uk/research/research-programmes/public-policy>

---

This work is licensed under the terms of the Creative Commons Attribution License 4.0 which permits unrestricted use, provided the original author and source are credited. The license is available at:  
<https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>

Cite this work as:

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute*.  
<https://doi.org/10.5281/zenodo.3240529>

# Table of Contents:

## What is AI ethics?

Intended audience and existing government guidance

AI ethics

## Why AI ethics?

An ethical platform for the responsible delivery of an AI project

Preliminary considerations about the ethical platform

Three building-blocks of a responsible AI project delivery ecosystem

## The SUM Values

### The FAST Track Principles

Fairness

Data fairness

Design fairness

Outcome fairness

Implementation fairness

Putting the principle of discriminatory non-harm into action

Accountability

Accountability deserves consideration both before and after model completion

Sustainability

Stakeholder Impact Assessment

Safety

Accuracy, reliability, security, and robustness

Risks posed to accuracy and reliability

Risks posed to security and robustness

Transparency

Defining transparent AI

Three critical tasks for designing and implementing transparent AI

Mapping AI transparency

## Process transparency: Establishing a Process-Based Governance Framework

## Outcome transparency: Explaining outcomes, clarifying content, implementing responsibly

Defining interpretable AI

Technical aspects of choosing, designing, and using an interpretable AI system

Guidelines for designing and delivering a sufficiently interpretable AI system

Guideline 1: Look first to context, potential impact, and domain-specific need

Guideline 2: Draw on standard interpretable techniques when possible

Guideline 3: Considerations for 'black box' AI systems

Guideline 4: Think about interpretability in terms of capacities for understanding

## Securing responsible delivery through human-centred implementation protocols and practices

Step 1: Consider aspects of application type and domain context to define roles

Step 2: Define delivery relations and map delivery processes

Step 3: Build an ethical implementation platform

## Conclusion

## Bibliography

## What is AI ethics?

### Intended audience and existing government guidance

The following guidance is designed to outline values, principles, and guidelines to assist department and delivery leads in ensuring that they develop and deploy AI ethically, safely, and responsibly. It is designed to complement and supplement the Data Ethics Framework. The [Data Ethics Framework](#) is a practical tool that should be used in any project initiation phase.

### AI ethics

A remarkable time of human promise has been ushered in by the convergence of the ever-expanding availability of big data, the soaring speed and stretch of cloud computing platforms, and the advancement of increasingly sophisticated machine learning algorithms.

This brave new digitally interconnected world is delivering rapid gains in the power of AI to better society. Innovations in AI are already dramatically improving the provision of essential social goods and services from healthcare, education, and transportation to food supply, energy, and environmental management. These bounties are, in fact, likely just the start. Because AI and machine learning systems organically improve with the enlargement of access to data and the growth of computing power, they will only become more effective and useful as the information age continues to develop apace. It may not be long before AI technologies become gatekeepers for the advancement of vital public interests and sustainable human development.

This prospect that progress in AI will help humanity to confront some of its most urgent challenges is exciting, but legitimate worries still abound. As with any new and rapidly evolving technology, a steep learning curve means that mistakes and miscalculations will be made and that both unanticipated and harmful impacts will inevitably occur. AI is no exception.

In order to manage these impacts responsibly and to direct the development of AI systems toward optimal public benefit, you will have to make considerations of **AI ethics and safety a first priority**.

This will involve integrating considerations of the social and ethical implications of the design and use of AI systems into **every stage** of the delivery of your AI project. It will also involve a **collaborative effort** between the data scientists, product managers, data engineers, domain experts, and delivery managers on your team to align the development of artificial intelligence technologies with ethical values and principles that safeguard and promote the wellbeing of the communities that these technologies affect.

By including a primer on AI ethics with the Guide, we are providing you with the conceptual resources and practical tools that will enable you to steward the responsible design and implementation of AI projects.

*AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies.*

These values, principles, and techniques are intended both to motivate morally acceptable practices and to prescribe the basic duties and obligations necessary to produce ethical, fair, and safe AI applications.

## Why AI ethics?

The field of AI ethics has largely emerged as a response to the range of individual and societal harms that the misuse, abuse, poor design, or negative unintended consequences of AI systems may cause. As a way to orient you to the importance of building a robust culture of AI ethics, here is a table that represents some of the most consequential forms that these potential harms may take:

### Potential Harms Caused by AI Systems

#### Bias and Discrimination

Because they gain their insights from the existing structures and dynamics of the societies they analyse, data-driven technologies can reproduce, reinforce, and amplify the patterns of marginalisation, inequality, and discrimination that exist in these societies.

Likewise, because many of the features, metrics, and analytic structures of the models that enable data mining are chosen by their designers, these technologies can potentially replicate their designers' preconceptions and biases.

Finally, the data samples used to train and test algorithmic systems can often be insufficiently representative of the populations from which they are drawing inferences. This creates real possibilities of biased and discriminatory outcomes, because the data being fed into the systems is flawed from the start.

#### Denial of Individual Autonomy, Recourse, and Rights

When citizens are subject to decisions, predictions, or classifications produced by AI systems, situations may arise where such individuals are unable to hold directly accountable the parties responsible for these outcomes.

AI systems automate cognitive functions that were previously attributable exclusively to accountable human agents. This can complicate the designation of responsibility in algorithmically generated outcomes, because the complex and distributed character of the design, production, and implementation processes of AI systems may make it difficult to pinpoint accountable parties.

In cases of injury or negative consequence, such an accountability gap may harm the autonomy and violate the rights of the affected individuals.

#### Non-transparent, Unexplainable, or Unjustifiable Outcomes

Many machine learning models generate their results by operating on high dimensional correlations that are beyond the interpretive capabilities of human scale reasoning. In these cases, the rationale of algorithmically produced outcomes that directly affect decision subjects remains opaque to those subjects. While in some use cases, this lack of explainability may be acceptable, in some applications, where the processed data could

harbour traces of discrimination, bias, inequity, or unfairness, the opaqueness of the model may be deeply problematic.

### Invasions of Privacy

Threats to privacy are posed by AI systems both as a result of their design and development processes, and as a result of their deployment. As AI projects are anchored in the structuring and processing of data, the development of AI technologies will frequently involve the utilisation of personal data. This data is sometimes captured and extracted without gaining the proper consent of the data subject or is handled in a way that reveals (or places under risk the revelation of) personal information.

On the deployment end, AI systems that target, profile, or nudge data subjects without their knowledge or consent could in some circumstances be interpreted as infringing upon their ability to lead a private life in which they are able to intentionally manage the transformative effects of the technologies that influence and shape their development. This sort of privacy invasion can consequently harm a person's more basic right to pursue their goals and life plans free from unchosen influence.

### Isolation and Disintegration of Social Connection

While the capacity of AI systems to curate individual experiences and to personalise digital services holds the promise of vastly improving consumer life and service delivery, this benefit also comes with potential risks. Excessive automation, for example, might reduce the need for human-to-human interaction, while algorithmically enabled hyper-personalisation, by limiting our exposure to worldviews different from ours, might polarise social relationships. Well-ordered and cohesive societies are built on relations of trust, empathy, and mutual understanding. As AI technologies become more prevalent, it is important that these relations be preserved.

### Unreliable, Unsafe, or Poor-Quality Outcomes

Irresponsible data management, negligent design and production processes, and questionable deployment practices can, each in their own ways, lead to the implementation and distribution of AI systems that produce unreliable, unsafe, or poor-quality outcomes. These outcomes can do direct damage to the wellbeing of individual persons and the public welfare. They can also undermine public trust in the responsible use of societally beneficial AI technologies, and they can create harmful inefficiencies by virtue of the dedication of limited public resources to inefficient or even detrimental AI technologies.

## An ethical platform for the responsible delivery of an AI project

Building a project delivery environment, which enables the ethical design and deployment of AI systems, requires a multidisciplinary team effort. It demands the active cooperation of all team members both in maintaining a **deeply ingrained culture of responsibility** and in executing a **governance architecture that adopts ethically sound practices at every point in the innovation and implementation lifecycle**.

This task of uniting an in-built culture of responsible innovation with a governance architecture that brings the values and principles of ethical, fair, and safe AI to life, will require that you and your team accomplish several goals:

- You will have to ensure that your AI project is ***ethically permissible*** by considering the impacts it may have on the wellbeing of affected stakeholders and communities.
- You will have to ensure that your AI project is ***fair and non-discriminatory*** by accounting for its potential to have discriminatory effects on individuals and social groups, by mitigating biases that may influence your model's outputs, and by being aware of the issues surrounding fairness that come into play at every phase of the design and implementation pipeline.
- You will have to ensure that your AI project is ***worthy of public trust*** by guaranteeing to the extent possible the safety, accuracy, reliability, security, and robustness of its product.
- You will have to ensure that your AI project is ***justifiable*** by prioritising both the transparency of the process by which your model is designed and implemented, and the transparency and interpretability of its decisions and behaviours.

We call this governance architecture an ***ethical platform*** for two important reasons. First, it is intended to provide you with a solid, process-based footing of values, principles, and protocols—*an ethical platform to stand on*—so that you and your team are better able to design and implement AI systems ethically, equitably, and safely. Secondly, it is intended to help you facilitate a culture of responsible AI innovation—to *help you provide an ethical platform to stand for*—so that your project team can be united in a collaborative spirit to develop AI technologies for the public good.

### Preliminary considerations about the ethical platform

Our aim for the remainder of this document is to provide you with guidance that is as comprehensive as possible in its presentation of the values, principles, and governance mechanisms necessary to serve the purpose of responsible innovation. Keep in mind, however, that not all issues discussed in this document will apply equally to each project. Clearly, a machine learning algorithm trained to detect spam emails will present fewer ethical challenges compared to one trained to detect cancer in blood samples. Similarly, image recognition systems used for sorting and routing mail raise fewer ethical dilemmas compared to the facial recognition technologies used in law enforcement.

Low-stakes AI applications that are not safety-critical, do not directly impact the lives of people, and do not process potentially sensitive social and demographic data will need less proactive ethical stewardship than high-stakes projects. You and your project team will need to evaluate the scope and possible impacts of your project on affected individuals and communities, and you will have to apply reasonable assessments of the risks posed to individual wellbeing and public welfare in order to formulate proportional governance procedures and protocols.

Be that as it may, you should also keep in mind that all AI projects have social and ethical impacts on stakeholders and communities even if just by diverting or redistributing limited intellectual, material, and economic resources away from other concerns and possibilities for socially beneficial innovation. Ethical considerations and principles-based policy formation should therefore play a salient role in every prospective AI project.

## Three building-blocks of a responsible AI project delivery ecosystem

Setting up an ethical platform for responsible AI project delivery involves not only *building from the cultural ground up*; it involves providing your team with the means to accomplish the goals of establishing the ethical permissibility, fairness, trustworthiness, and justifiability of your project. It will take three building-blocks to make such an ethical platform possible:

1. At the most basic level, it necessitates that you gain a working knowledge of a framework of **ethical values** that *Support, Underwrite, and Motivate* a responsible data design and use ecosystem. These will be called **SUM Values**, and they will be composed of four key notions: *Respect, Connect, Care, and Protect*. The objectives of these SUM Values are (1) to provide you with an accessible framework to start thinking about the moral scope of the societal and ethical impacts of your project and (2) to establish well-defined criteria to evaluate its ethical permissibility.
2. At a second and more concrete level, an ethical platform for responsible AI project delivery requires a set of **actionable principles** that facilitate an orientation to the responsible design and use of AI systems. These will be called **FAST Track Principles**, and they will be composed of four key notions: *Fairness, Accountability, Sustainability, and Transparency*. The objectives of these FAST Track Principles are to provide you with the moral and practical tools (1) to make sure that your project is bias-mitigating, non-discriminatory, and fair, and (2) to safeguard public trust in your project's capacity to deliver safe and reliable AI innovation.
3. At a third and most concrete level, an ethical platform for responsible AI project delivery requires a **process-based governance framework (PBG Framework)** that **operationalises the SUM Values and the FAST Track Principles** across the entire AI project delivery workflow. The objective of this PBG Framework is to set up transparent processes of design and implementation that safeguard and enable the justifiability of both your AI project and its product.

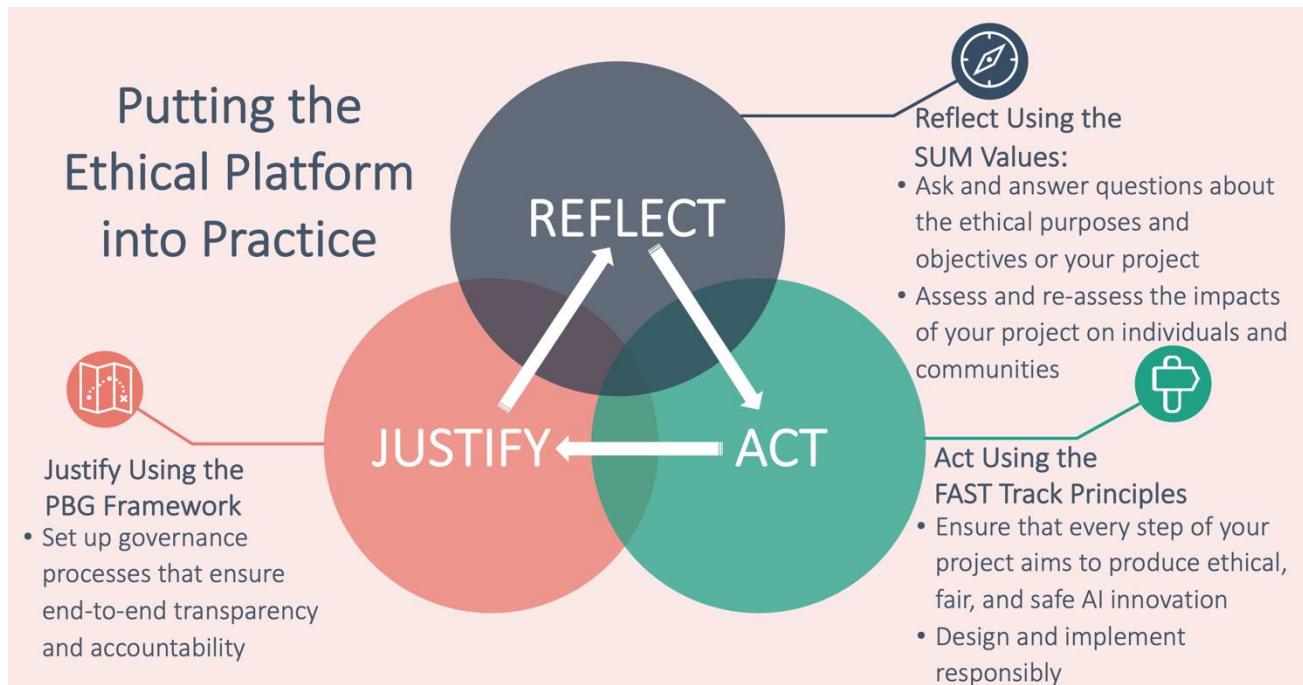
Here is a summary visualisation of these three building blocks of the platform:

### Ethical Platform for the Responsible Delivery of an AI Project



## *How to use this guide*

This guide is intended to assist you in stewarding practices of responsible AI innovation. This entails that the ethical platform be put into practice at every step of the design and implementation workflow. Turning the SUM Values, the FAST Track Principles, and the PBG Framework into practice will require that you and your team continuously **reflect, act, and justify**:



## The SUM Values

### Background

The challenge of creating a culture of responsible innovation begins with the task of building an **accessible moral vocabulary** that will allow team members to explore and discuss the ethical stakes of the AI projects that they are involved in or are considering taking on.

In the field of AI ethics, this moral vocabulary draws primarily on two traditions of moral thinking: (1) **bioethics** and (2) **human rights discourse**. **Bioethics** is the study of the ethical impacts of biomedicine and the applied life sciences. **Human rights discourse** draws inspiration from the UN Declaration of Human Rights. It is anchored in a set of universal principles that build upon the idea that all humans have an equal moral status as bearers of intrinsic human dignity.

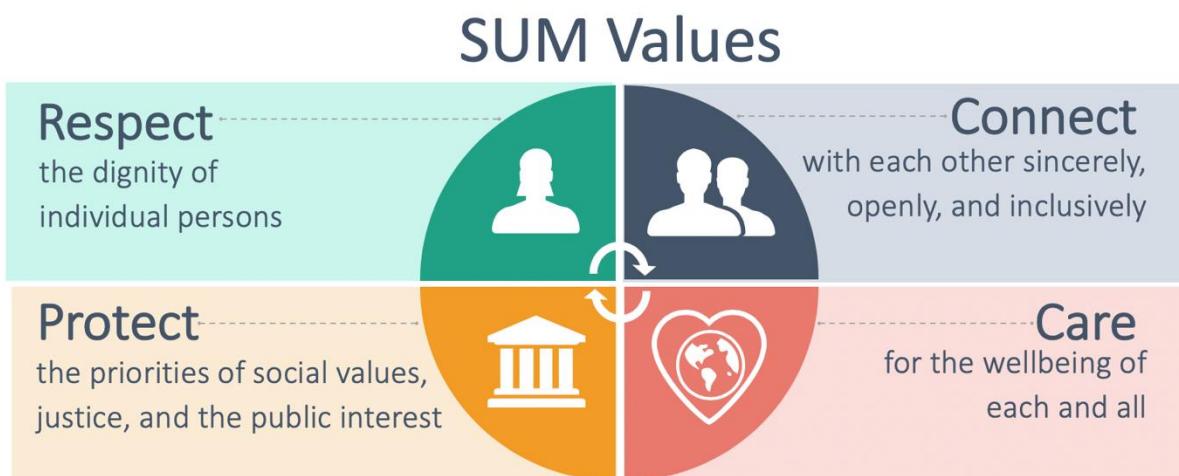
Whereas bioethics largely stresses the normative values that underlie the safeguarding of **individuals** in instances where technological practices affect their interests and wellbeing, human rights discourse mainly focuses on the set of **social, political, and legal entitlements** that are due to all human beings under a universal framework of juridical protection and the rule of law.

The main principles of bioethics include respecting the autonomy of the individual, protecting people from harm, looking after the well-being of others, and treating all individuals equitably and justly. The main tenets of human rights include the entitlement to equal freedom and dignity under the law, the protection of civil, political, and social rights, the universal recognition of personhood, and the right to free and unencumbered participation in the life of the community.

#### *The SUM Values: Respect, Connect, Care, and Protect*

While the SUM Values incorporate conceptual elements from both bioethics and human rights discourse, they do so with an eye to applying the most critical of these elements to the specific social and ethical problems raised by the potential misuse, abuse, poor design, or harmful unintended consequences of AI systems.

They are also meant to be utilised as guiding values throughout the innovation lifecycle: from the preliminary steps of project evaluation, planning, and problem formulation, through processes of design, development, and testing, to the stages of implementation and reassessment. The SUM Values can be visualised as follows:



#### Key Concept: Normativity/Normative

In the context of practical ethics, the word '**normativity**' means that a given concept, value, or belief puts a moral demand on one's practices, i.e. that such a concept, value, or belief indicates what one '**should**' or '**ought to**' do in circumstances where that concept, value, or belief applies. For example, if I hold the moral belief that helping people in need is a good thing, then, when confronted with a sick person in the street who requires help, I should help them. My belief puts a normative demand on me to act in accordance with what it is indicating that I ought to do, namely to come to the needy person's aid.

In order to focus in on a more detailed exploration of each of the values' meanings, their contents will be presented individually. Formulating it as a question: What are each of these values charging you to do?

→ **RESPECT the dignity of individual persons:**

- Ensure their abilities to make free and informed decisions about their own lives
- Safeguard their autonomy, their power to express themselves, and their right to be heard
- Secure their capacities to make well-considered and independent contributions to the life of the community
- Support their abilities to flourish, to fully develop themselves, and to pursue their passions and talents according to their own freely determined life plans

→ **CONNECT with each other sincerely, openly, and inclusively:**

- Safeguard the integrity of interpersonal dialogue, meaningful human connection, and social cohesion
- Prioritise diversity, participation, and inclusion at all points in the design, development, and deployment processes of AI innovation.
- Encourage all voices to be heard and all opinions to be weighed seriously and sincerely throughout the production and use lifecycle
- Use the advancement and proliferation of AI technologies to strengthen the developmentally essential relationship between interacting human beings.
- Utilise AI innovations *pro-socially* so as to enable bonds of interpersonal solidarity to form and individuals to be socialised and recognised by each other
- Use AI technologies to foster this capacity to connect so as to reinforce the edifice of trust, empathy, reciprocal responsibility, and mutual understanding upon which all ethically well-founded social orders rest

→ **CARE for the wellbeing of each and all:**

- Design and deploy AI systems to foster and to cultivate the welfare of all stakeholders whose interests are affected by their use
- Do no harm with these technologies and minimise the risks of their misuse or abuse

- Prioritise the safety and the mental and physical integrity of people when scanning horizons of technological possibility and when conceiving of and deploying AI applications

→ PROTECT the priorities of social values, justice, and the public interest:

- Treat all individuals equally and protect social equity
- Use digital technologies as an essential support for the protection of fair and equal treatment under the law
- Prioritise social welfare, public interest, and the consideration of the social and ethical impacts of innovation in determining the legitimacy and desirability of AI technologies
- Use AI to empower and to advance the interests and well-being of as many individuals as possible
- Think big-picture about the wider impacts of the AI technologies you are conceiving and developing. Think about the ramifications of their effects and externalities for others around the globe, for future generations, and for the biosphere as a whole

As a general rule, these SUM Values should orient you in deliberating about the **ethical permissibility** of a prospective AI project. They should also provide you with a framework of concepts to consider the **ethical impacts of an AI system across the design, use, and monitoring lifecycle**.

Taking these SUM Values as a starting point of conversation, you should also encourage discussion within your team of how to weigh the values against one another and how to consider trade-offs should use case specific circumstances arise when the values come into tension with each other.

## The FAST Track Principles:

### Background

While the SUM Values are intended to provide you with some general normative guideposts and moral motivations for thinking through the social and ethical aspects of AI project delivery, they are not specifically catered to the actual processes involved in developing and deploying AI systems.

To make clear what is needed for this next step toward a more actionable orientation to the responsible design and use of AI technologies, it would be helpful to briefly touch upon what has necessitated the emergence of AI ethics in the first place.

Marvin Minsky, the great cognitive scientist and AI pioneer, defined AI as follows: ‘Artificial Intelligence is the science of ***making computers do things that require intelligence*** when done by humans.’ This standard definition should key us in to the principal motivation that has driven the development of the field of the applied ethics of artificial intelligence:

When humans do ‘things that require intelligence,’ we hold them responsible for the accuracy, reliability, and soundness of their judgements. Moreover, we demand of them that their actions and decisions be supported by good reasons, and we hold them accountable for the fairness, equity, and reasonableness of how they treat others.

What creates the need for principles tailored to the design and use of AI systems is that their emergence and expanding power ‘to do things that require intelligence’ has heralded a shift of a wide array of cognitive functions to algorithmic processes that themselves can be held neither directly responsible nor immediately accountable for the consequences of their behaviour.

As inert and program-based machinery, AI systems are not morally accountable agents. This has created an ethical breach in the sphere of the applied science of AI that the growing number of frameworks for AI ethics are currently trying to fill. Targeted principles such as fairness, accountability, sustainability, and transparency are meant to ‘fill the gap’ between the new ‘smart agency’ of machines and their fundamental lack of moral responsibility.

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

By becoming well-acquainted with the FAST Track Principles, *all members* of your project delivery team will be better able to support a responsible environment for data innovation.

Issues of fairness, accountability, sustainability, and transparency operate at every juncture and at every level of the AI project delivery workflow and demand the cooperative attention and deliberative involvement of those with technical expertise, domain knowledge, project/product management skill, and policy competence. Ethical AI innovation is a team effort from start to finish.

To introduce you to the scope of the FAST Track Principles, here is a summary visualisation of them:

## FAST Track Principles



You should keep in mind, initially, that while fairness, accountability, sustainability, and transparency are grouped together in the FAST acronym, they do not necessarily relate to each other on the same plane or as equivalents. The principles of accountability and transparency are ***end-to-end governing principles***. Accountability entails that humans are answerable for the parts they play across the entire AI design and implementation workflow. It also demands that the results of this work are traceable from start to finish. The principle of transparency entails that design and implementation processes are justifiable through and through. It demands as well that an algorithmically influenced outcome is interpretable and made understandable to affected parties.

The governing roles of accountability and transparency are very different from the more dependent roles of fairness and sustainability. These latter two are *qualities* of algorithmic systems for which their designers and implementers are ***held accountable*** through the ***transparency both of the outcomes of their practices and of the practices themselves***. According to the principle of fairness, designers and implementers are held accountable for being equitable and for not harming anyone through bias or discrimination. According to the principle of sustainability, designers and implementers are held accountable for producing AI innovation that is safe and ethical in its outcomes and wider impacts.

Whereas the principles of transparency and accountability thus provide the procedural mechanisms and means through which AI systems can be justified and by which their producer and implementers can be held responsible, fairness and sustainability are the crucial aspects of the design, implementation, and outcomes of these systems which establish the normative criteria for such governing constraints. These four principles are therefore all deeply interrelated, but they are not equal.

There is one more important thing to keep in mind before we delve into the details of the FAST Track principles. Transparency, accountability, and fairness are *also data protection principles*, and where algorithmic processing involves personal data, complying with them is not simply a matter of ethics or good practice, but a legal requirement, which is enshrined in the General Data Protection Regulation (GDPR) and the Data Protection Act of 2018 (DPA 2018). For more detailed information about the specific meanings of transparency, accountability, and fairness as data protection principles in the context of the GDPR and the DPA 2018, please refer to the [Guide to Data Protection](#) produced by the Information Commissioner's Office.

## Fairness

When thinking about fairness in the design and deployment of AI systems, it is important to always keep in mind that these technologies, no matter how neutral they may seem, are designed and produced by human beings, who are bound by the limitations of their contexts and biases.

Human error, prejudice, and misjudgement can enter into the innovation lifecycle and create biases at any point in the project delivery process from the preliminary stages of data extraction, collection, and pre-processing to the critical phases of problem formulation, model building, and implementation.

Additionally, data-driven technologies achieve accuracy and efficacy by building inferences from datasets that record complex social and historical patterns, which themselves may contain culturally crystallised forms of bias and discrimination. There is no silver bullet when it comes to remediating the dangers of discrimination and unfairness in AI systems. The problem of fairness and bias mitigation in algorithmic design and use therefore has no simple or strictly technical solution.

That said, best practices of fairness-aware design and implementation (both at the level of non-technical self-assessment and at the level of technical controls and means of evaluation) hold great promise in terms of securing just, morally acceptable, and beneficial outcomes that treat affected stakeholders fairly and equitably.

While there are different ways to characterise or define fairness in the design and use of AI systems, you should consider the **principle of discriminatory non-harm** as a minimum required threshold of fairness. This principle directs us to do no harm to others through the biased or discriminatory outcomes that may result from practices of AI innovation:

**Principle of Discriminatory Non-Harm:** The designers and users of AI systems, which process social or demographic data pertaining to features of human subjects, societal patterns, or cultural formations, should prioritise the mitigation of bias and the exclusion of discriminatory influences on the outputs and implementations of their models. Prioritising discriminatory non-harm implies that the designers and users of AI systems ensure that the decisions and behaviours of their models do not generate discriminatory or inequitable impacts on affected individuals and communities. This entails that these designers and users ensure that the AI systems they are developing and deploying:

1. Are trained and tested on properly representative, relevant, accurate, and generalisable datasets (**Data Fairness**)
2. Have model architectures that do not include target variables, features, processes, or analytical structures (correlations, interactions, and inferences) which are unreasonable, morally objectionable, or unjustifiable (**Design Fairness**)
3. Do not have discriminatory or inequitable impacts on the lives of the people they affect (**Outcome Fairness**)
4. Are deployed by users sufficiently trained to implement them responsibly and without bias (**Implementation Fairness**)

### *Data fairness*

Responsible data acquisition, handling, and management is a necessary component of algorithmic fairness. If the results of your AI project are generated by biased, compromised, or skewed datasets, affected stakeholders will not adequately be protected from discriminatory harm. Your project team should keep in mind the following key elements of data fairness:

- **Representativeness:** Depending on the context, either underrepresentation or overrepresentation of disadvantaged or legally protected groups in the data sample may lead to the systematic disadvantaging of vulnerable stakeholders in the outcomes of the trained model. To avoid such kinds of sampling bias, domain expertise will be crucial to assess the fit between the data collected or procured and the underlying population to be modelled. Technical team members should, if possible, offer means of remediation to correct for representational flaws in the sampling.
- **Fit-for-Purpose and Sufficiency:** An important question to consider in the data collection and procurement process is: Will the amount of data collected be sufficient for the intended purpose of the project? The quantity of data collected or procured has a significant impact on the accuracy and reasonableness of the outputs of a trained model. A data sample not large enough to represent with sufficient richness the significant or qualifying attributes of the members of a population to be classified may lead to unfair outcomes. Insufficient datasets may not equitably reflect the qualities that should rationally be weighed in producing a justified outcome that is consistent with the desired purpose of the AI system. Members of the project team with technical and policy competences should collaborate to determine if the data quantity is, in this respect, sufficient and fit-for-purpose.
- **Source Integrity and Measurement Accuracy:** Effective bias mitigation begins at the very commencement of data extraction and collection processes. Both the sources and instruments of measurement may introduce discriminatory factors into a dataset. When incorporated as inputs in the training data, biased prior human decisions and judgments—such as prejudiced scoring, ranking, interview-data or evaluation—will become the ‘ground truth’ of the model and replicate the bias in the outputs of the system. In order to secure discriminatory non-harm, you must do your best to make sure your data sample has optimal source integrity. This involves securing or confirming that the data gathering processes involved suitable, reliable, and impartial sources of measurement and sound methods of collection.
- **Timeliness and Recency:** If your datasets include outdated data then changes in the underlying data distribution may adversely affect the generalisability of your trained model. Provided these distributional drifts reflect changing social relationship or group dynamics, this loss of accuracy with regard to the actual characteristics of the underlying population may introduce bias into your AI system. In preventing discriminatory outcomes, you should scrutinise the timeliness and recency of all elements of the data that constitute your datasets.
- **Relevance, Appropriateness and Domain Knowledge:** The understanding and utilisation of the most appropriate sources and types of data are crucial for building a robust and unbiased AI system. Solid domain knowledge of the underlying population distribution and of the predictive or classificatory goal of the project is instrumental for choosing optimally relevant measurement inputs that contribute to the reasonable determination of the defined solution. You should make sure that domain experts collaborate closely with your technical team to assist in the determination of the optimally appropriate categories and sources of measurement.

To ensure the uptake of best practices for responsible data acquisition, handling, and management across your AI project delivery workflow, you should initiate the creation of a **Dataset Factsheet** at the alpha stage of your project. This factsheet should be maintained diligently throughout the design and implementation lifecycle in order to secure optimal data quality, deliberate bias-mitigation aware practices, and optimal auditability. It should include a **comprehensive record of data provenance, procurement, pre-processing, lineage, storage, and security** as well as qualitative input from team members about determinations made with regard to data representativeness, data sufficiency, source integrity, data timeliness, data relevance, training/testing/validating splits, and unforeseen data issues encountered across the workflow.

### *Design Fairness*

Because human beings have a hand in all stages of the construction of AI systems, fairness-aware design must take precautions across the AI project workflow to prevent bias from having a discriminatory influence:

- **Problem Formulation:** At the initial stage of problem formulation and outcome definition, technical and non-technical members of your team should work together to translate project goals into measurable targets. This will involve the use of both domain knowledge and technical understanding to define what is being optimised in a formalisable way and to translate the project's objective into a target variable or measurable proxy, which operates as a statistically actionable rendering of the defined outcome.

At each of these points, choices must be made about the design of the algorithmic system that may introduce structural biases which ultimately lead to discriminatory harm. Special care must be taken here to identify affected stakeholders and to consider how vulnerable groups might be negatively impacted by the specification of outcome variables and proxies. Attention must also be paid to the question of whether these specifications are reasonable and justifiable given the general purpose of the project and the potential impacts that the outcomes of the system's use will have on the individuals and communities involved.

These challenges of fairness aware design at the problem formulation stage show the need for making diversity and inclusive participation a priority from the start of the AI project lifecycle. This involves both the collaboration of the entire team and the attainment of stakeholder input about the acceptability of the project plan. This also entails collaborative deliberation across the project team and beyond about the ethical impacts of the design choices made.

- **Data Pre-Processing:** Human judgment enters into the process of algorithmic system construction at the stage of labelling, annotating, and organising the training data to be utilised in building the model. Choices made about how to classify and structure raw inputs must be taken in a fairness aware manner with due consideration given to the sensitive social contexts that may introduce bias into such acts of classification. Similar fairness aware processes should be put in place to review automated or outsourced classifications. Likewise, efforts should be made to attach solid contextual information and ample metadata to the datasets, so that downstream analyses of data processing have access to properties of concern in bias mitigation.

- **Feature Determination and Model-Building:** The constructive task of selecting the attributes or features that will serve as input variables for your model involves human decisions being made about what sorts of information may or may not be relevant or rationally required to yield an accurate *and* unbiased classification or prediction. Moreover, the feature engineering tasks of aggregating, extracting, or decomposing attributes from datasets may introduce human appraisals that have biasing effects. For this reason, discrimination awareness should play a large role at this stage of the AI model-building workflow as should domain knowledge and policy expertise. Your team should proceed in the modelling stage aware that choices made about grouping or separating and including or excluding features as well as more general judgements about the comprehensiveness or coarseness of the total set of features may have significant consequences for vulnerable or protected groups.

The process of tuning hyperparameters and setting metrics at the modelling, testing, and evaluation stages also involves human choices that may have discriminatory effects in the trained model. Your technical team should proceed with an attentiveness to bias risk, and continual iterations of peer review and project team consultation should be encouraged to ensure that choices made in adjusting the dials and metrics of the model are in line with bias mitigation and discriminatory non-harm.

- **Evaluating Analytical Structures:** Design fairness also demands close assessment of the existence in the trained model of lurking or hidden proxies for discriminatory features that may act as significant factors in its output. Including such hidden proxies in the structure of the model may lead to implicit ‘redlining’ (the unfair treatment of a sensitive group on the basis of an unprotected attribute or interaction of attributes that ‘stands in’ for a protected or sensitive one).

Designers must additionally scrutinise the moral justifiability of the significant correlations and inferences that are determined by the model’s learning mechanisms themselves. In cases of the processing of social or demographic data related to human features, where the complexity and high dimensionality of machine learning models preclude the confirmation of the discriminatory non-harm of these inferences (for reason of their uninterpretability by human assessors), these models should be avoided. In AI systems that process and draw analytics from data arising from human relationships, societal patterns, and complex socioeconomic and cultural formations, designers must prioritise a degree of interpretability that is sufficient to ensure that the inferences produced by these systems are non-discriminatory. In cases where this is not possible, a different, more transparent and explainable model or portfolio of models should be chosen.

Analytical structures must also be confirmed to be *procedurally fair*. Any rule or procedure employed in an AI system should be consistently and uniformly applied to every decision subject whose information is being processed by that system. Your team should be able to certify that when a rule or procedure has been used to render an outcome for any given individual, the same rule or procedure will be applied to any other individual in the same way regardless of that other subject’s similarities with or differences from the first.

Implementers, in this respect, should be able to show that any algorithmic output is replicable when the same rules and procedures are applied to the same inputs. Such a uniformity of the application of rules and procedures secures the equal procedural treatment of decision subjects and precludes any rule-changes in the algorithmic processing targeted at a specific person that may disadvantage that individual vis-à-vis any other.

### *Outcome fairness*

As part of this minimum safeguarding of discriminatory non-harm, forethought and well-informed consideration must be put into *how you are going to define and measure the fairness of the impacts and outcomes of the AI system you are developing*.

There is a great diversity of beliefs in the area of **outcome fairness** as to how to properly classify what makes the consequences of an algorithmically supported decision equitable, fair, and allocatively just. Different approaches—detailed below—stress different principles: some focus on demographic parity, some on individual fairness, others on error rates equitably distributed across subpopulations.

Your determination of outcome fairness should heavily depend both on the **specific use case for which the fairness of outcome is being considered** and the **technical feasibility of incorporating your chosen criteria into the construction of the AI system**. (Note that different fairness-aware methods involve different types of technical interventions at the pre-processing, modelling, or post-processing stages of production). Again, this means that determining your fairness definition should be a **cooperative and multidisciplinary effort across the project team**.

You will find below a summary table of some of the main definitions of outcome fairness that have been integrated by researchers into formal models as well as a list of current articles and technical resources, which should be consulted to orient your team to the relevant knowledge base. (Note that this is a rapidly developing field, so your technical team should keep updated about further advances.) The first four fairness types fall under the category of group fairness and allow for comparative criteria of non-discrimination to be considered in model construction and evaluation. The final two fairness types focus instead on cases of individual fairness, where context-specific issues of effective bias are considered and assessed at the level of the individual agent.

Take note, though, that these technical approaches have limited scope in terms of the bigger picture issues of algorithmic fairness that we have already stressed. Many of the formal approaches work only in use cases that have *distributive or allocative consequences*. In order to carry out group comparisons, these approaches require access to data about sensitive/protected attributes (that may often be unavailable or unreliable) as well as accurate demographic information about the underlying population distribution. Furthermore, there are unavoidable trade-offs and inconsistencies between these technical definitions that must be weighed in determining which of them are best fit for your use case. Consult those on your project team with the technical expertise to consider the use case appropriateness of a desired formal approach.

Some Formalisable Definitions of Outcome Fairness	
Type of Fairness	Definition
<b>Demographic/ Statistical Parity</b>	An outcome is fair if each group in the selected set receives benefit in equal or similar proportions, i.e. if there is no correlation between a sensitive or protected attribute and the allocative result. This approach is intended to prevent <i>disparate impact</i> , which occurs when the outcome of an algorithmic process disproportionately harms members of disadvantaged or protected groups.
<b>True Positive Rate Parity</b>	An outcome is fair if the ‘true positive’ rates of an algorithmic prediction or classification are equal across groups. This approach is intended to align the goals of bias mitigation and accuracy by ensuring that the accuracy of the model is equivalent between relevant population subgroups. This method is also referred to as ‘equal opportunity’ fairness because it aims to secure equalised odds of an advantageous outcome for qualified individuals in a given population regardless of the protected or disadvantaged groups of which they are members.
<b>False Positive Rate Parity</b>	An outcome is fair if it does not disparately mistreat people belonging to a given social group by misclassifying them at a higher rate than the members of a second social group, for this would place the members of the first group at an unfair disadvantage. This approach is motivated by the position that sensitive groups and advantaged groups should have similar error rates in outcomes of algorithmic decisions.
<b>Positive Predictive Value Parity</b>	An outcome is fair if the rates of positive predictive value (the fraction of correctly predicted positive cases out of all predicted positive cases) are equal across sensitive and advantaged groups. Outcome fairness is defined here in terms of a parity of precision, where the probability of members from different groups actually having the quality they are predicted to have is the same across groups.
<b>Individual Fairness</b>	An outcome is fair if it treats individuals with similar relevant qualifications similarly. This approach relies on the establishment of a similarity metric that shows the degree to which pairs of individuals are alike with regard to a specific task.
<b>Counterfactual Fairness</b>	An outcome is fair if an automated decision made about an individual belonging to a sensitive group would have been the same were that individual a member of a different group in a closest possible alternative (or counterfactual) world. Like the individual fairness approach, this method of defining fairness focuses on the specific circumstances of an affected decision subject, but, by using the tools of contrastive explanation, it moves beyond individual fairness insofar as it brings out the causal influences behind the algorithmic output. It also presents the possibility of offering the subject of an automated decision knowledge of what factors, if changed, could have influenced a different outcome. This could provide them with actionable recourse to change an unfavourable decision.

## Selected References and Technical Resources

- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM. (Statistical Parity and Individual Fairness)
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, 325–333. (Demographic Parity)
- Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323. (Equality of Opportunity)
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2):153–163. (Balancing Error Rates)
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM. (Test for Disparate Impact)
- Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. International World Wide Web Conferences Steering Committee. (Disparate Mistreatment)
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, 1-7. Fairware '18. (Summary and Comparison)
- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4066–4076. (Counterfactual Fairness)
- Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10-19. (Extension of Counterfactual Fairness)

### Technical Resources for Exploring Fairness Tools:

- <https://dsapp.uchicago.edu/projects/aequitas/> (University of Chicago's open source bias audit toolkit for machine learning developers)
- <http://www.fairness-measures.org/> and [https://github.com/megantosh/fairness\\_measures\\_code/](https://github.com/megantosh/fairness_measures_code/) (Datasets and software for detecting algorithmic discrimination from TU Berlin and Eurecat)
- <https://github.com/columbia/fairtest> (Fairtest unwarranted association discovery platform from Columbia University)
- <http://aif360.mybluemix.net/#> (IBM's Fairness 360 open source toolkit)

### Fairness Position Statement:

Once you and your project team have thoroughly considered the use case appropriateness as well as technical feasibility of the formal models of fairness most relevant for your system and have incorporated the model into your application, you should prepare a **Fairness Position Statement (FPS)** in which the fairness criteria being employed in the AI system is made explicit and explained in plain and non-technical language. This FPS should then be made publicly available for review by all affected stakeholders.

### *Implementation fairness*

When your project team is approaching the beta stage, you should begin to build out your plan for implementation training and support. This plan should include adequate preparation for the responsible and unbiased deployment of the AI system by its on-the-ground users. Automated

decision-support systems present novel risks of bias and misapplication at the point of delivery, so special attention should be paid to preventing harmful or discriminatory outcomes at this critical juncture of the AI project lifecycle.

In order to design an optimal regime of implementer training and support, you should pay special attention to the unique pitfalls of bias-in-use to which the deployment of AI technologies give rise. These can be loosely classified as decision-automation bias (more commonly just ‘automation bias’) and automation-distrust bias:

- **Decision-Automation Bias/The Technological Halo Effect:** Users of automated decision-support systems may tend to become hampered in their critical judgment, rational agency, and situational awareness as a result of their faith in the perceived objectivity, neutrality, certainty, or superiority of the AI system.

This may lead to **over-reliance** or **errors of omission**, where implementers lose the capacity to identify and respond to the faults, errors, or deficiencies, which might arise over the course of the use of an automated system, because they become complacent and overly deferent to its directions and cues. Decision-automation bias may also lead to **over-compliance** or **errors of commission** where implementers defer to the perceived infallibility of the system and thereby become unable to detect problems emerging from its use for reason of a failure to hold the results against available information.

Both over-reliance and over-compliance may lead to what is known as out-of-loop syndrome where the degradation of the role of human reason and the deskilling of critical thinking hampers the user’s ability to complete the tasks that have been automated. This condition may bring about a loss of the ability to respond to system failure and may lead both to safety hazards and to dangers of discriminatory harm.

To combat risks of decision-automation bias, you should operationalise strong regimes of accountability at the site of user deployment to steer human decision-agents to act on the basis of good reasons, solid inferences, and critical judgment.

- **Automation-Distrust Bias:** At the other extreme, users of an automated decision-support system may tend to disregard its salient contributions to evidence-based reasoning either as a result of their distrust or skepticism about AI technologies in general or as a result of their over-prioritisation of the importance of prudence, common sense, and human expertise. An aversion to the non-human and amoral character of automated systems may also influence decision subjects’ hesitation to consult these technologies in high impact contexts such as healthcare, transportation, and law.

In order to secure and safeguard fair implementation of AI systems by users well-trained to utilise the algorithmic outputs as tools for making evidence-based judgements, you should consider the following measures:

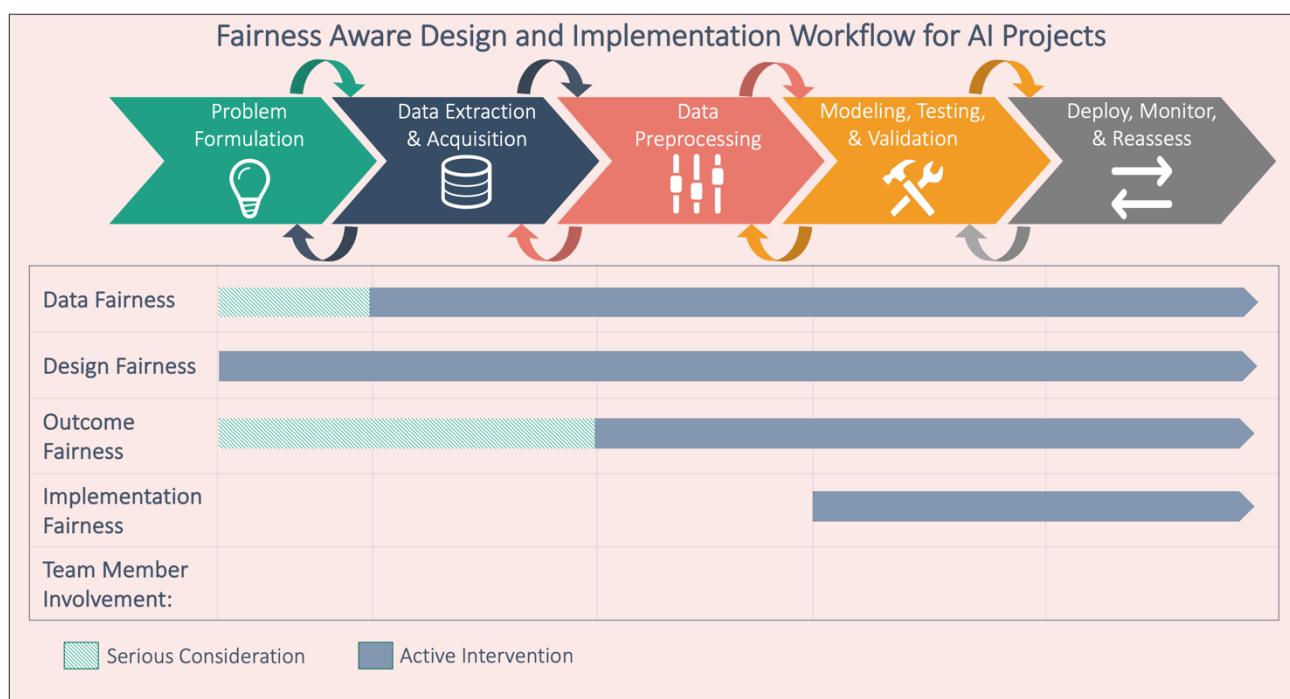
- Training of implementers should include the conveyance of basic knowledge about the statistical and probabilistic character of machine learning and about the limitations of AI and automated decision-support technologies. This training should avoid any anthropomorphic

(or human-like) portrayals of AI systems and should encourage users to view the benefits and risks of deploying these systems in terms of their role in assisting human judgment rather than replacing it.

- Forethought should be given in the design of the user-system interface about human factors and about possibilities for implementation biases. The systems should be *designed into* processes that encourage active user judgment and situational awareness. The interface between the user and the system should be designed to make clear and accessible to the user the system's rationale, compliance to fairness standards, and confidence level. Ideally this should happen in a 'runtime' manner.
- Training of implementers should include a pre-emptive exploration of the cognitive and judgmental biases that may occur across deployment contexts. This should be done in a use case based manner that highlights the particular misjudgements that may occur when people weigh statistical evidence. Examples of the latter may include overconfidence in prediction based on the historical consistency of data, illusions that any clustering of data points necessarily indicates significant insights, and discounting of societal patterns that exist beyond the statistical results.

#### *Putting the principle of discriminatory non-harm into action*

When you are considering how to put the principle of discriminatory non-harm into action, you should come together with all the managers on the project team to map out team member involvement at each stage of the AI project pipeline from alpha through beta. Considering fairness aware design and implementation from a workflow perspective will allow you, as a team, to concretise and make explicit end-to-end paths of accountability in a clear and peer-reviewable manner. This is essential for establishing a robust accountability framework. Here is a schematic representation of the fairness aware workflow. You will have to complete the final row.



Considering fairness aware design and implementation from such a workflow perspective will also assist you in pinpointing risks of bias or downstream discrimination and streamlining possible solutions in a proactive, pre-emptive, and anticipatory way. At each stage of the AI project pipeline (i.e. at each column of the above table), you and the relevant members of your team should carry out a collaborative self-assessment with regard to the applicable dimension of fairness. This is a three-step process:

#### Discriminatory Non-Harm Self-Assessment

Step 1: Identify the fairness and bias mitigation dimensions that apply to the specific stage under consideration (for example, at the data pre-processing stage, dimensions of data fairness, design fairness, and outcome fairness may be at issue).

Step 2: Scrutinise how your particular AI project might pose risks or have unintended vulnerabilities in each of these areas.

Step 3: Take action to correct any existing problems that have been identified, strengthen areas of weakness that have possible discriminatory consequences, and take proactive bias-prevention measures in areas that have been identified to pose potential future risks.

## Accountability

When considering the role of accountability in the AI project delivery lifecycle, it is important first to make sure that you are taking a ‘best practices’ approach to data processing that is aligned with [Principle 6 of the Data Ethics Framework](#). Beyond following this general guidance, however, you should pay special attention to the new and unique challenges posed to public sector accountability by the design and implementation of AI systems.

Responsible AI project delivery requires that two related challenges to public sector accountability be confronted directly:

1. **Accountability gap:** As mentioned above, automated decisions are not self-justifiable. Whereas human agents can be called to account for their judgements and decisions in instances where those judgments and decisions affect the interests of others, the statistical models and underlying hardware that compose AI systems are not responsible in the same morally relevant sense. This creates an accountability gap that must be addressed so that clear and imputable sources of human answerability can be attached to decisions assisted or produced by an AI system.
2. **Complexity of AI production processes:** Establishing human answerability is not a simple matter when it comes to the design and deployment of AI systems. This is due to the complexity and multi-agent character of the development and use of these systems. Typically, AI project delivery workflows include department and delivery leads, technical

experts, data procurement and preparation personnel, policy and domain experts, implementers, and others. Due to this production complexity, it may become difficult to answer the question of who among these parties involved in the production of AI systems should bear responsibility if these systems' uses have negative consequences and impacts.

Meeting the special requirements of accountability, which are born out of these two challenges, call for a sufficiently fine-grained concept of what would make an AI project properly accountable. This concept can be broken down into two subcomponents of accountability: **answerability** and **auditability**:

- **Answerability:** The principle of accountability demands that the onus of justifying algorithmically supported decisions be placed on the shoulders of the human creators and users of those AI systems. This means that it is essential to establish a continuous chain of human responsibility across the whole AI project delivery workflow. Making sure that accountability is effective from end to end necessitates that no gaps be permitted in the answerability of responsible human authorities from first steps of the design of an AI system to its algorithmically steered outcomes.

Answerability also demands that explanations and justifications of both the content of algorithmically supported decisions and the processes behind their production be offered by competent human authorities in plain, understandable, and coherent language. These explanations and justifications should be based upon sincere, consistent, sound, and impartial reasons that are accessible to non-technical hearers.

- **Auditability:** Whereas the notion of answerability responds to the question of *who is accountable* for an automation supported outcome, the notion of auditability answers the question of *how the designers and implementers of AI systems are to be held accountable*. This aspect of accountability has to do with **demonstrating** both the **responsibility of design and use practices** and the **justifiability of outcomes**.

Your project team must ensure that every step of the process of designing and implementing your AI project is accessible for audit, oversight, and review. Successful audit requires builders and implementers of algorithmic systems to keep records and to make accessible information that enables monitoring of the soundness and diligence of the innovation processes that produced the AI system.

Auditability also requires that your project team keep records and make accessible information that enables monitoring of data provenance and analysis from the stages of collection, pre-processing, and modelling to training, testing, and deploying. This is the purpose of the previously mentioned Dataset Factsheet.

Moreover, it requires your team to enable peers and overseers to probe and to critically review the dynamic operation of the system in order to ensure that the procedures and operations which are producing the model's behaviour are safe, ethical, and fair. Practically transparent algorithmic models must be **built for auditability, reproducible**, and **equipped for end-to-end recording and monitoring** of their data processing.

The deliberate incorporation of both of these elements of accountability (answerability and auditability) into the AI project lifecycle may be called **Accountability-by-Design**:

**Accountability by Design:** All AI systems must be designed to facilitate end-to-end answerability and auditability. This requires both **responsible humans-in-the-loop** across the entire design and implementation chain as well as **activity monitoring protocols** that enable end-to-end oversight and review.

### *Accountability deserves consideration across the entire design and implementation workflow*

As a best practice, you should actively consider the different demands that accountability by design places on you before and after the roll out of your AI project. We will refer to the process of ensuring accountability during the design and development stages of your AI project as '**anticipatory accountability**.' This is because you are anticipating your AI project's accountability needs prior to it being completed. Following a similar logic, we will refer to the process of addressing accountability after the start of the deployment of your AI project as '**remedial accountability**.' This is because after the initial implementation of your system, you are remedying any of the issues that may be raised by its effects and potential externalities. These two subtypes of accountability are sometimes referred to as *ex-ante* (or before-the-event) accountability and *ex-post* (after-the-event) accountability respectively.

- **Anticipatory Accountability:** Treating accountability as an anticipatory principle entails that you take as of primary importance the decisions made and actions taken by your project delivery team prior to the outcome of an algorithmically supported decision process.

This kind of *ex ante* accountability should be prioritised over remedial accountability, which focuses instead on the corrective or justificatory measures that can be taken after that automation supported process had been completed.

By ensuring the AI project delivery processes are accountable prior to the actual application of the system in the world, you will bolster the soundness of design and implementation processes and thereby more effectively pre-empt possible harms to individual wellbeing and public welfare.

Likewise, by establishing strong regimes of anticipatory accountability and by making the design and delivery process as open and publicly accessible as possible, you will put affected stakeholders in a position to make better informed and more knowledgeable decisions about their involvement with these systems in advance of potentially harmful impacts. In doing so, you will also strengthen the public narrative and help to safeguard the project from reputational harm.

- **Remedial Accountability:** While remedial accountability should be seen, along these lines, as a necessary fallback rather than as a first resort for imputing responsibility in the design and deployment of AI systems, strong regimes of remedial accountability are no less important in

providing necessary justifications for the bearing these systems have on the lives of affected stakeholders.

**Putting in place comprehensive auditability regimes as part of your accountability framework and establishing transparent design and use practices, which are methodically logged throughout the AI project delivery lifecycle, are essential components for this sort of remedial accountability.**

One aspect of remedial accountability that you must pay close attention to is the need to provide **explanations** to affected stakeholders for algorithmically supported decisions. This aspect of accountable and transparent design and use practices will be called **explicability**, which literally means the ability to make explicit the meaning of the algorithmic model's result.

Offering explanations for the results of algorithmically supported decision-making involves furnishing decision subjects and other interested parties with an understandable account of the rationale behind the specific outcome of interest. It also involves furnishing the decision subject and other interested parties with an explanation of the ethical permissibility, the fairness, and the safety of the use of the AI system. These tasks of **content clarification** and **practical justification** will be explored in more detail below as part of the section on transparency.

## Sustainability

Designers and users of AI systems should remain aware that these technologies may have transformative and long-term effects on individuals and society. In order to ensure that the deployment of your AI system remains sustainable and supports the sustainability of the communities it will affect, you and your team should proceed with a continuous sensitivity to the real-world impacts that your system will have.

### *Stakeholder Impact Assessment*

You and your project team should come together to evaluate the social impact and sustainability of your AI project through a **Stakeholder Impact Assessment (SIA)**, whether the AI project is being used to deliver a public service or in a back-office administrative capacity. When we refer to 'stakeholders' we are referring primarily to affected individual persons, but the term may also extend to groups and organisations in the sense that individual members of these collectives may also be impacted as such by the design and deployment of AI systems. Due consideration to stakeholders should be given at both of these levels.

The purpose of carrying out an SIA is multidimensional. SIAs can serve several purposes, some of which include:

- (1) Help to build public confidence that the design and deployment of the AI system by the public sector agency has been done responsibly
- (2) Facilitate and strengthen your accountability framework
- (3) Bring to light unseen risks that threaten to affect individuals and the public good

- (4) Underwrite well-informed decision-making and transparent innovation practices
- (5) Demonstrate forethought and due diligence not only within your organisation but also to the wider public

Your team should convene to evaluate the social impact and sustainability of your AI project through the SIA at three critical points in the project delivery lifecycle:

- 1. Alpha Phase (Problem Formulation):** Carry out an initial Stakeholder Impact Assessment (SIA) to determine the ethical permissibility of the project. Refer to the SUM Values as a starting point for the considerations of the possible effects of your project on individual wellbeing and public welfare. In cases where you conclude that your AI project will have significant ethical and social impacts, you should open your initial SIA to the public so that their views can be properly considered. This will bolster the inclusion of a diversity of voices and opinions into the design and development process through the participation of a more representative range of stakeholders. You should also consider consulting with internal organisational stakeholders, whose input will likewise strengthen the openness, inclusivity, and diversity of your project.
- 2. From Alpha to Beta (Pre-Implementation):** Once your model has been trained, tested, and validated, you and your team should revisit your initial SIA to confirm that the AI system to be implemented is still in line with the evaluations and conclusions of your original assessment. This check-in should be logged on the pre-implementation section of the SIA with any applicable changes added and discussed. Before the launch of the system, this SIA should be made publicly available. At this point you must also set a timeframe for re-assessment once the system is in operation as well as a public consultation which predates and provides input for that re-assessment. Timeframes for these re-assessments should be decided by your team on a case-by-case basis but should be proportional to the scale of the potential impact of the system on the individuals and communities it will affect.
- 3. Beta Phase (Re-Assessment):** After your AI system has gone live, your team should intermittently revisit and re-evaluate your SIA. These check-ins should be logged on the re-assessment section of the SIA with any applicable changes added and discussed. Re-assessment should focus both on evaluating the existing SIA against real world impacts and on considering how to mitigate the unintended consequences that may have ensued in the wake of the deployment of the system. Further public consultation for input at the beta stage should be undertaken before the re-assessment so that stakeholder input can be included in re-assessment deliberations.

You should keep in mind that, in its specific focus on social and ethical sustainability, your Stakeholder Impact Assessment constitutes just one part of the governance platform for your AI project and should be a complement to your accountability framework and other auditing and activity-monitoring documentation.

Your SIA should be broken down into four sections of questions and responses. In the 1<sup>st</sup> section, there should be general questions about the possible big-picture social and ethical impacts of the use of the AI system you plan to build. In the 2<sup>nd</sup> section, your team should collaboratively formulate relevant sector-specific and use case-specific questions about the impact of the AI system on

affected stakeholders. The 3<sup>rd</sup> section should provide answers to the additional questions relevant to pre-implementation evaluation. The 4<sup>th</sup> section should provide the opportunity for members of your team to reassess the system in light of its real-world impacts, public input, and possible unintended consequences.

Here is a prototype of an SIA:

<b><u>Stakeholder Impact Assessment for (Project Name)</u></b>	
<p><b>1. Alpha Phase (Problem Formulation) General Questions</b></p> <p><b>Completed on this Date:</b></p>	<p><b>I. Identifying Affected Stakeholders</b></p> <p>Who are the stakeholders that this AI project is most likely to affect? What groups of these stakeholders are most vulnerable? How might the project negatively impact them?</p> <p><b>II. Goal-Setting and Objective-Mapping</b></p> <p>How are you defining the outcome (the target variable) that the system is optimising for? Is this a fair, reasonable, and widely acceptable definition?</p> <p>Does the target variable (or its measurable proxy) reflect a reasonable and justifiable translation of the project's objective into the statistical frame?</p> <p>Is this translation justifiable given the general purpose of the project and the potential impacts that the outcomes of its implementation will have on the communities involved?</p> <p><b>III. Possible Impacts on the Individual</b></p> <p>How might the implementation of your AI system impact the abilities of affected stakeholders to make free, independent, and well-informed decisions about their lives? How might it enhance or diminish their autonomy?</p> <p>How might it affect their capacities to flourish and to fully develop themselves?</p> <p>How might it do harm to their physical or mental integrity? Have risks to individual health and safety been adequately considered and addressed?</p> <p>How might it infringe on their privacy rights, both on the data processing end of designing the system and on the implementation end of deploying it?</p> <p><b>IV. Possible Impacts on Society and Interpersonal Relationships</b></p> <p>How might the implementation of your AI system adversely affect each stakeholder's fair and equal treatment under the law? Are there any aspects of the project that expose vulnerable communities to possible discriminatory harm?</p> <p>How might the use of your system affect the integrity of interpersonal dialogue, meaningful human connection, and social cohesion?</p>

	<p>Have the values of civic participation, inclusion, and diversity been adequately considered in articulating the purpose and setting the goals of the project? If not, how might these values be incorporated into your project design?</p> <p>Does the project aim to advance the interests and well-being of as many affected individuals as possible? Might any disparate socioeconomic impacts result from its deployment?</p> <p>Have you sufficiently considered the wider impacts of the system on future generations and on the planet as a whole?</p>
<p><b>2. Alpha Phase (Problem Formulation)</b> <b>Sector-Specific and Use Case-Specific Questions</b></p> <p><b>Completed on this Date:</b></p>	<p>In this section you should consider the sector-specific and use case-specific issues surrounding the social and ethical impacts of your AI project on affected stakeholders. Compile a list of the questions and concerns you anticipate. State how your team is attempting to address these questions and concerns.</p>
<p><b>3. From Alpha to Beta (Pre-Implementation)</b></p> <p><b>Completed on this Date:</b></p>	<p>After reviewing the results of your initial SIA, answer the following questions:</p> <p>Are the trained model's actual objective, design, and testing results still in line with the evaluations and conclusions contained in your original assessment? If not, how does your assessment now differ?</p> <p>Have any other areas of concern arisen with regard to possibly harmful social or ethical impacts as you have moved from the alpha to the beta phase?</p> <p>You must also set a reasonable timeframe for public consultation and beta phase re-assessment:</p> <p><b>Dates of Public Consultation on Beta-Phase Impacts:</b></p> <p><b>Date of Planned Beta Phase Re-Assessment:</b></p>
<p><b>4. Beta Phase (Re-Assessment)</b></p> <p><b>Completed on this Date:</b></p>	<p>Once you have reviewed the most recent version of your SIA and the results of the public consultation, answer the following questions:</p> <p>How does the content of the existing SIA compare with the real-world impacts of the AI system as measured by available evidence of performance, monitoring data, and input from implementers and the public?</p> <p>What steps can be taken to rectify any problems or issues that have emerged?</p> <p>Have any unintended harmful consequences ensued in the wake of the deployment of the system? If so, how might these negative impacts be mitigated and redressed?</p>

	<p>Have the maintenance processes for your AI model adequately taken into account the possibility of distributional shifts in the underlying population? Has the model been properly retuned and retrained to accommodate changes in the environment?</p> <p><b>Dates of Public Consultation on Beta-Phase Impacts:</b></p> <p><b>Date of Next Planned Beta Phase Re-Assessment:</b></p>
--	--

## Safety

Beyond safeguarding the sustainability of your AI project as it relates to its social impacts on individual wellbeing and public welfare, your project team must also confront the related challenge of **technical sustainability or safety**. A technically sustainable AI system is **safe, accurate, reliable, secure, and robust**. Securing these goals, however, is a difficult and unremitting task.

Because AI systems operate in a world filled with uncertainty, volatility, and flux, the challenge of building safe and reliable AI can be especially daunting. This job, however, must be met head-on. Only by making the goal of producing safe and reliable AI technologies central to your project, will you be able to mitigate risks of your system failing at scale when faced with real-world unknowns and unforeseen events. The issue of **AI safety** is of paramount importance, because these potential failures may both produce harmful outcomes and undermine public trust.

In order to safeguard that your AI system functions safely, you must prioritise the technical objectives of **accuracy, reliability, security, and robustness**. This requires that your technical team put careful forethought into how to construct a system that **accurately and dependably operates in accordance with its designers' expectations even when confronted with unexpected changes, anomalies, and perturbations**. Building an AI system that meets these safety goals also requires rigorous testing, validation, and re-assessment as well as the integration of adequate mechanisms of oversight and control into its real-world operation.

### *Accuracy, reliability, security, and robustness*

It is important that you gain a strong working knowledge of each of the safety relevant operational objectives (**accuracy, reliability, security, and robustness**):

- **Accuracy and Performance Metrics:** In machine learning, the accuracy of a model is the proportion of examples for which it generates a correct output. This performance measure is also sometimes characterised conversely as an **error rate** or the fraction of cases for which the model produces an incorrect output. Keep in mind that, in some instances, the choice of an acceptable error rate or accuracy level can be adjusted in accordance with the use case specific needs of the application. In other instances, it may be largely set by a domain established benchmark.

As a performance metric, accuracy should be a central component to establishing and nuancing your team's approach to safe AI. That said, specifying a reasonable performance level for your system may also often require you to refine or exchange your measure of accuracy. For instance, if certain errors are more significant or costly than others, a metric for total cost can be integrated into your model so that the cost of one class of errors can be weighed against that of another. Likewise, if the precision and sensitivity of the system in detecting uncommon events is a priority (say, in instances of the medical diagnosis of rare cases of a disease), you can use the technique of precision and recall. This method of addressing imbalanced classification would allow you to weigh the proportion of the system's correct detections—both of frequent and of rare outcomes—against the proportion of actual detections of the rare event (i.e. the ratio of the true detections of the rare outcome to the sum of the true detections of that outcome and the missed detections or false negatives for that outcome).

In general, measuring accuracy in the face of uncertainty is a challenge that must be given significant thought. The confidence level of your AI system will depend heavily on problems inherent in attempts to model a chaotic and changing reality. Concerns about accuracy must cope with issues of unavoidable noise present in the data sample, architectural uncertainties generated by the possibility that a given model is missing relevant features of the underlying distribution, and inevitable changes in input data over time.

- **Reliability:** The objective of reliability is that an AI system behaves exactly as its designers intended and anticipated. A reliable system adheres to the specifications it was programmed to carry out. Reliability is therefore a measure of **consistency** and can establish confidence in the safety of a system based upon the dependability with which it operationally conforms to its intended functionality.
- **Security:** The goal of security encompasses the protection of several operational dimensions of an AI system when confronted with possible adversarial attack. A secure system is capable of maintaining the **integrity** of the information that constitutes it. This includes protecting its architecture from the unauthorised modification or damage of any of its component parts. A secure system also remains continuously **functional** and **accessible** to its authorised users and keeps **confidential** and **private information** secure even under hostile or adversarial conditions.
- **Robustness:** The objective of robustness can be thought of as the goal that an AI system functions reliably and accurately under harsh conditions. These conditions may include adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications). The measure of robustness is therefore the strength of a system's integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, and undesirable reinforcement learning behaviour.

*Risks posed to accuracy and reliability:*

**Concept Drift:** Once trained, most machine learning systems operate on static models of the world that have been built from historical data which have become fixed in the systems' parameters. This freezing of the model before it is released 'into the wild' makes its accuracy and reliability especially vulnerable to changes in the underlying distribution of data. When the historical data that have crystallised into the trained model's architecture cease to reflect the population concerned, the model's mapping function will no longer be able to accurately and reliably transform its inputs into its target output values. These systems can quickly become prone to error in unexpected and harmful ways.

There has been much valuable research done on methods of detecting and mitigating concept and distribution drift, and you should consult with your technical team to ensure that its members have familiarised themselves with this research and have sufficient knowledge of the available ways to confront the issue. In all cases, you should remain vigilant to the potentially rapid concept drifts that may occur in the complex, dynamic, and evolving environments in which your AI project will intervene. Remaining aware of these transformations in the data is crucial for safe AI, and your team should actively formulate an action plan to anticipate and to mitigate their impacts on the performance of your system.

**Brittleness:** Another possible challenge to the accuracy and reliability of machine learning systems arises from the inherent limitations of the systems themselves. Many of the high-performing machine learning models such as deep neural nets (DNN) rely on massive amounts of data and brute force repetition of training examples to tune the thousands, millions, or even billions of parameters, which collectively generate their outputs.

However, when they are actually running in an unpredictable world, these systems may have difficulty processing unfamiliar events and scenarios. They may make unexpected and serious mistakes, because they have neither the capacity to contextualise the problems they are programmed to solve nor the common-sense ability to determine the relevance of new 'unknowns'. Moreover, these mistakes may remain unexplainable given the high-dimensionality and computational complexity of their mathematical structures. This fragility or brittleness can have especially significant consequences in safety-critical applications like fully automated transportation and medical decision support systems where undetectable changes in inputs may lead to significant failures. While progress is being made in finding ways to make these models more robust, it is crucial to consider safety first when weighing up their viability.

### *Risks posed to security and robustness*

**Adversarial Attack:** Adversarial attacks on machine learning models maliciously modify input data—often in imperceptible ways—to induce them into misclassification or incorrect prediction. For instance, by undetectably altering a few pixels on a picture, an adversarial attacker can mislead a model into generating an incorrect output (like identifying a panda as a gibbon or a 'stop' sign as a 'speed limit' sign) with an extremely high confidence. While a good amount of attention has been paid to the risks that adversarial attacks pose in deep learning applications like computer vision, these kinds of perturbations are also effective across a vast range of machine learning techniques and uses such as spam filtering and malware detection.

These vulnerabilities of AI systems to adversarial examples have serious consequences for AI safety. The existence of cases where subtle but targeted perturbations cause models to be misled into gross miscalculation and incorrect decisions have potentially serious safety implication for the adoption of critical systems like applications in autonomous transportation, medical imaging, and security and surveillance. In response to concerns about the threats posed to a safe and trusted environment for AI technologies by adversarial attacks a field called **adversarial machine learning** has emerged over the past several years. Work in this area focuses on securing systems from disruptive perturbations at all points of vulnerability across the AI pipeline.

One of the major safety strategies that has arisen from this research is an approach called **model hardening**, which has advanced techniques that combat adversarial attacks by strengthening the architectural components of the systems. Model hardening techniques may include adversarial training, where training data is methodically enlarged to include adversarial examples. Other model hardening methods involve architectural modification, regularisation, and data pre-processing manipulation. A second notable safety strategy is **run-time detection**, where the system is augmented with a discovery apparatus that can identify and trace in real-time the existence of adversarial examples. You should consult with members of your technical team to ensure that the risks of adversarial attack have been taken into account and mitigated throughout the AI lifecycle. A valuable collection of resources to combat adversarial attack can be found at <https://github.com/IBM/adversarial-robustness-toolbox>.

**Data Poisoning:** A different but related type of adversarial attack is called data poisoning. This threat to safe and reliable AI involves a malicious compromise of data sources at the point of collection and pre-processing. Data poisoning occurs when an adversary modifies or manipulates part of the dataset upon which a model will be trained, validated, and tested. By altering a selected subset of training inputs, a poisoning attack can induce a trained AI system into curated misclassification, systemic malfunction, and poor performance. An especially concerning dimension of targeted data poisoning is that an adversary may introduce a ‘backdoor’ into the infected model whereby the trained system functions normally until it processes maliciously selected inputs that trigger error or failure.

In order to combat data poisoning, your technical team should become familiar with the state of the art in filtering and detecting poisoned data. However, such technical solutions are not enough. Data poisoning is possible because data collection and procurement often involves potentially unreliable or questionable sources. When data originates in uncontrollable environments like the internet, social media, or the Internet of Things, many opportunities present themselves to ill-intentioned attackers, who aim to manipulate training examples. Likewise, in third- party data curation processes (such as ‘crowdsourced’ labelling, annotation, and content identification), attackers may simply handcraft malicious inputs. Your project team should focus on the best practices of responsible data management, so that they are able to tend to data quality as an end-to-end priority.

- **Misdirected Reinforcement Learning Behaviour:** A different set of safety risks emerges from the approach to machine learning called reinforcement learning (RL). In the more widely

applied methods of supervised learning that have largely been the focus of this guide, a model transforms inputs into outputs according to a fixed mapping function that has resulted from its passively received training. In RL, by contrast, the learner system actively solves problems by engaging with its environment through trial and error. This exploration and ‘problem-solving’ behaviour is determined by the objective of maximising a reward function that is defined by its designers.

This flexibility in the model, however, comes at the price of potential safety risks. An RL system, which is operating in the real-world without sufficient controls, may determine a reward-optimising course of action that is optimal for achieving its desired objective but harmful to people. Because these models lack context-awareness, common sense, empathy, and understanding, they are unable to identify, on their own, scenarios that may have damaging consequences but that were not anticipated and constrained by their programmers. This is a difficult problem, because the unbounded complexity of the world makes anticipating all of its pitfalls and detrimental variables veritably impossible.

Existing strategies to mitigate such risks of misdirected reinforcement learning behaviour include:

- Running extensive simulations during the testing stage, so that appropriate measures of constraint can be programmed into the system
- Continuous inspection and monitoring of the system, so that its behaviour can be better predicted and understood
- Finding ways to make the system more interpretable so that its decisions can be better assessed
- Hard-wiring mechanisms into the system that enable human override and system shut-down

## End-to-End AI Safety

The safety risks you face in your AI project will depend, among other factors, on the sort of algorithm(s) and machine learning techniques you are using, the type of applications in which those techniques are going to be deployed, the provenance of your data, the way you are specifying your objective, and the problem domain in which that specification applies. As a best practice, regardless of this variability of techniques and circumstances, safety considerations of accuracy, reliability, security, and robustness should be in operation at every stage of your AI project lifecycle.

This should involve both **rigorous protocols of testing, validating, verifying, and monitoring the safety of the system** and the performance of **AI safety self-assessments** by relevant members of your team at each stage of the workflow. Such self-assessments should evaluate how your team’s design and implementation practices line up with the safety objectives of accuracy, reliability, security, and robustness. Your AI safety self-assessments should be logged across the workflow on a single document in a running fashion that allows review and re-assessment.

## Transparency

## Defining transparent AI

It is important to remember that *transparency as a principle of AI ethics* differs a bit in meaning from the everyday use of the term. The common dictionary understanding of transparency defines it as either (1) the quality an object has when one can see clearly through it or (2) the quality of a situation or process that can be clearly justified and explained because it is open to inspection and free from secrets.

Transparency as a principle of AI ethics encompasses *both* of these meanings:

On the one hand, transparent AI involves the interpretability of a given AI system, i.e. **the ability to know how and why a model performed the way it did in a specific context and therefore to understand the rationale behind its decision or behaviour**. This sort of transparency is often referred to by way of the metaphor of ‘opening the black box’ of AI. It involves *content clarification and intelligibility* or *explicability*.

On the other hand, transparent AI involves **the justifiability both of the processes that go into its design and implementation and of its outcome**. It therefore involves the *soundness of the justification of its use*. In this more normative meaning, transparent AI is *practically justifiable* in an unrestricted way if one can demonstrate that both the design and implementation processes that have gone into the particular decision or behaviour of a system and the decision or behaviour itself are *ethically permissible, non-discriminatory/fair, and worthy of public trust/safety-securing*.

### *Three critical tasks for designing and implementing transparent AI*

This two-pronged definition of transparency as a principle of AI ethics asks that you to think about transparent AI both in terms of the *process* behind it (the design and implementation practices that lead to an algorithmically supported outcome) and in terms of its *product* (the content and justification of that outcome). Such a process/product distinction is crucial, because it clarifies the three tasks that your team will be responsible for in safeguarding the transparency of your AI project:

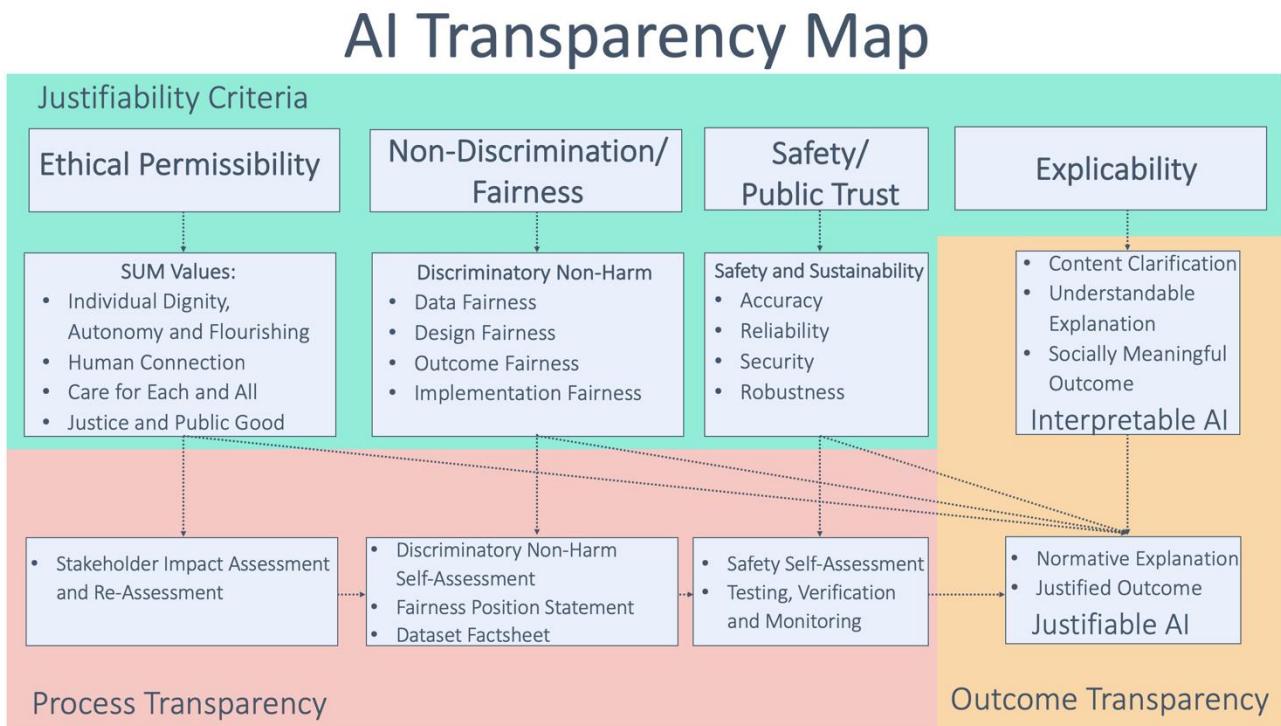
- **Process Transparency, Task 1: Justify Process.** In offering an explanation to affected stakeholders, you should be able to demonstrate that considerations of ethical permissibility, non-discrimination/fairness, and safety/public trustworthiness were operative end-to-end in the design and implementation processes that lead to an automated decision or behaviour. This task will be supported both by following the best practices outlined herein throughout the AI project lifecycle and by putting into place robust auditability measures through an accountability-by-design framework.
- **Outcome Transparency, Task 2: Clarify Content and Explain Outcome.** In offering an explanation to affected stakeholders, you should be able to show in plain language that is understandable to non-specialists how and why a model performed the way it did in a specific decision-making or behavioural context. You should therefore be able to clarify and communicate the rationale behind its decision or behaviour. This explanation should be *socially meaningful* in the sense that the terms and logic of the explanation should not simply

reproduce the formal characteristics or the technical meanings and rationale of the mathematical model but should rather be translated into the everyday language of human practices and therefore be understandable in terms of the societal factors and relationships that the decision or behaviour implicates.

- **Outcome Transparency, Task 3: Justify Outcome.** In offering an explanation to affected stakeholders, you should be able to demonstrate that a specific decision or behaviour of your system is ethically permissible, non-discriminatory/fair, and worthy of public trust/safety-securing. This outcome justification should take the content clarification/explicated outcome from task 2 as its starting point and weigh that explanation against the justifiability criteria adhered to throughout the design and use pipeline: ethical permissibility, non-discrimination/fairness, and safety/public trustworthiness. Undertaking an optimal approach to process transparency from the start should support and safeguard this demand for normative explanation and outcome justification.

### *Mapping AI transparency*

Before exploring each of the three tasks individually, it may be helpful to visualise the relationship between these connected components of transparent AI:



### Process Transparency: Establishing a Process-Based Governance Framework

The central importance of the end-to-end operability of good governance practices should guide your strategy to build out responsible AI project workflow processes. Three components are essential to creating a such a responsible workflow: (1) Maintaining strong regimes of professional and institutional transparency; (2) Having a clear and accessible Process-Based Governance

Framework (PBG Framework); (3) Establishing a well-defined auditability trail in your PBG Framework through robust activity logging protocols that are consolidated digitally in a process log.

1. **Professional and Institutional Transparency:** At every stage of the design and implementation of your AI project, team members should be held to rigorous standards of conduct that secure and maintain professionalism and institutional transparency. These standards should include the core values of **integrity, honesty, sincerity, neutrality, objectivity and impartiality**. All professionals involved in the research, development, production, and implementation of AI technologies are, first and foremost, acting as **fiduciaries of the public interest** and must, in keeping with these core values of the Civil Service, put the obligations to serve that interest above any other concerns.

Furthermore, from start to finish of the AI project lifecycle, the design and implementation process should be as transparent and as open to public scrutiny as possible with restrictions on accessibility to relevant information limited to the reasonable protection of justified public sector confidentiality and of analytics that may tip off bad actors to methods of gaming the system of service provision.

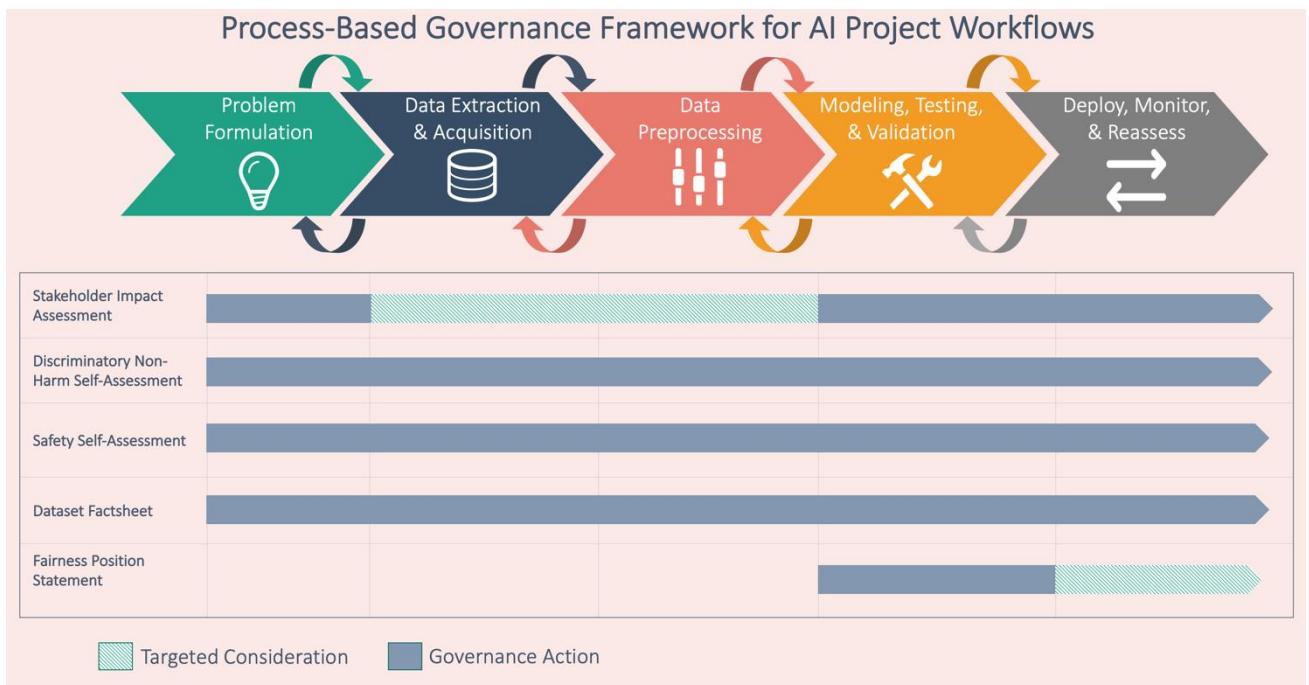
2. **Process-Based Governance Framework:** So far, this guide has presented some of the main steps that are necessary for establishing responsible innovation practices in your AI project. Perhaps the most vital of these measures is the effective operationalisation of the values and principles that underpin the development of ethical and safe AI. By organising all of your governance considerations and actions into a PBG Framework, you will be better able to accomplish this task.

The purpose of a PBG Framework is to provide a template for the integrations of the norms, values, and principles, which motivate and steer responsible innovation, with the actual processes that characterise the AI design and development pipeline. While the accompanying Guide has focused primarily on the Cross Industry Standard Process for Data Mining (CRISP-DM), keep in mind that such a structured integration of values and principles with innovation processes is just as applicable in other related workflow models like Knowledge Discovery in Databases (KDD) and Sample, Explore, Modify, Model, and Assess (SEMMA).

Your PBG Framework should give you a landscape view of the governance procedures and protocols that are organising the control structures of your project workflow. Constructing a good PBG Framework will provide you and your team with a big picture of:

- The relevant team members and roles involved in each governance action.
- The relevant stages of the workflow in which intervention and targeted consideration are necessary to meet governance goals
- Explicit timeframes for any necessary follow-up actions, re-assessments, and continual monitoring
- Clear and well-defined protocols for logging activity and for instituting mechanisms to assure end-to-end auditability

To help you get a summary picture of where the components of process transparency explored so far fit into a PBG Framework, here is a landscape view:



3. **Enabling Auditability with a Process Log:** With your controls in place and your governance framework organised, you will be better able to manage and consolidate the information necessary to assure end-to-end auditability. This information should include both the records and activity-monitoring results that are yielded by your PBG Framework and the model development data gathered across the modelling, training, testing, verifying, and implementation phases.

By centralising your information digitally in a process log, you are preparing the way for optimal process transparency. A process log will enable you to make available, in one place, information that may assist you in demonstrating to concerned parties and affected decision subjects both the responsibility of design and use practices and the justifiability of the outcomes of your system's processing behaviour.

Such a log will also allow you to differentially organise the accessibility and presentation of the information yielded by your project. Not only is this crucial to preserving and protecting data that legitimately should remain unavailable for public view, it will afford your team the capacity to cater the presentation of results to different tiers of stakeholders with different interests and levels of expertise. This ability to curate your explanations with the user-receiver in mind will be vital to achieving the goals of interpretable and justifiable AI.

## Outcome transparency: Explaining outcome and clarifying content

Beyond enabling process transparency through your PBG Framework, you must also put in place standards and protocols to ensure that clear and understandable explanations of the outcomes of your AI system's decisions, behaviours, and problem-solving tasks can:

1. Properly inform the evidence-based judgments of the implementers that they are designed to support;
2. Be offered to affected stakeholders and concerned parties in an accessible way.

This is a multifaceted undertaking that will demand careful forethought and participation across your entire project team.

There is no simple technological solution for how to effectively clarify and convey the rationale behind a model's output in a particular decision-making or behavioural context. Your team will have to use sound judgement and common sense in order to bring together the **technical aspects** of choosing, designing, using a sufficiently interpretable AI system and the **delivery aspects** of being able to clarify and communicate in plain, non-technical, and socially meaningful language how and why that system performed the way it did in a specific decision-making or behavioural context.

Having a good grasp of the rationale and criteria behind the decision-making and problem-solving behaviour of your system is essential for producing safe, fair, and ethical AI. If your AI model is not sufficiently interpretable—if you aren't able to draw from it humanly understandable explanations of the factors that played a significant role in determining its behaviours—then you may not be able to tell how and why things go wrong in your system when they do.

This is a crucial and unavoidable issue for reasons we have already explored. Ensuring the safety of high impact systems in transportation, medicine, infrastructure, and security requires human verification that these systems have properly learned the critical tasks they are charged to complete. It also requires confirmation that when confronted with unfamiliar circumstances, anomalies, and perturbations, these systems will not fail or make unintuitive errors. Moreover, ensuring that these systems operate without causing discriminatory harms requires effective ways to detect and to mitigate sources of bias and inequitable influence that may be buried deep within their feature spaces, inferences, and architectures. Without interpretability each one of these tasks necessary for delivering safe and morally justifiable AI will remain incomplete.

### **Defining Interpretable AI**

To gain a foothold in both the technical and delivery dimensions of AI interpretability, you will first need a solid working definition of what interpretable AI is. To this end, it may be useful to recall once again the definition of AI offered in the accompanying Guide: '*Artificial Intelligence is the science of making computers do things that require intelligence when done by humans.*'

This characterisation is important, because it brings out an essential feature of the explanatory demands of interpretable AI: to do things that require intelligence when done by humans means to do things that require *reasoning processes and cognitive functioning*. This cognitive dimension has a direct bearing on how you should think about offering suitable explanations about algorithmically generated outcomes:

**Explaining an algorithmic model's decision or behaviour should involve making explicit how the particular set of factors which determined that outcome can play the role of evidence in supporting**

the conclusion reached. It should involve making intelligible to affected individuals the rationale behind that decision or behaviour as if it had been produced by a reasoning, evidence-using, and inference-making person.

What makes this explanation-giving task so demanding when it comes to AI systems is that reasoning processes do not occur, for humans, at just one level. Rather, human-scale reasoning and interpreting includes:

1. Aspects of **logic** (applying the basic principles of validity that lie behind and give form to sound thinking): *This aspect aligns with the need for formal or logical explanations of AI systems.*
2. Aspects of **semantics** (gaining an understanding of how and why things work the way they do and what they mean): *This aspect aligns with the need for explanations of the technical rationale behind the outcomes AI systems.*
3. Aspects of the **social understanding of practices, beliefs, and intentions** (clarifying the content of interpersonal relations, societal norms, and individual objectives): *This aspect aligns with the need for the clarification of the socially meaningful content of the outcomes of AI systems.*
4. Aspects of **moral justification** (making sense of what should be considered right and wrong in our everyday activities and choices): *This aspect aligns with the justifiability of AI systems.*

There are good reasons why ***all four of these dimensions of human reasoning processes*** must factor in to explaining the decisions and behaviours of AI systems: First and most evidently, understanding the logic and technical innerworkings (i.e. semantic content) of these systems is a precondition for ensuring their safety and fairness. Secondly, because they are designed and used to achieve human objectives and to fulfil surrogate cognitive functions *in the everyday social world*, we need to make sense of these systems in terms of the consequential roles that their decisions and behaviours play in that human reality. The social context of these outcomes matters greatly. Finally, because they actually affect individuals and society in direct and morally consequential ways, we need to be able to understand and explain their outcomes not just in terms of their mathematical logic, technical rationale, and social context but also in terms of the justifiability of their impacts on people.

Delving more deeply into the technical and delivery aspects of interpretable AI will show how these four dimensions of human reasoning directly line up with the different levels of demand for explanations of the outcomes of AI systems. In particular, the logical and semantic dimensions will weigh heavily in technical considerations whereas the social and moral dimensions will be significant at the point of delivery.

Note here, though, that these different dimensions of human reasoning are not necessarily mutually exclusive but build off and depend upon each other in significant and cascading ways. Approaching explanations of interpretable AI should therefore be treated holistically and inclusively. Technical explanation of the logic and rationale of a given model, for instance, should be seen as a support for the context-based clarification of its socially meaningful content, just as that socially meaningful content should be viewed as forming the basis of explaining an outcome's moral justifiability. When

considering how to make the outcomes of decision-making and problem-solving AI systems maximally transparent to affected stakeholders, you should take this rounded view of human reasoning into account, because it will help you address more effectively the spectrum of concerns that these stakeholders may have.

### Technical aspects of choosing, designing, and using an interpretable AI system

Keep in mind that, while, on the face of it, the task of choosing between the numerous AI and machine learning algorithms may seem daunting, it need not be so. By sticking to the priority of outcome transparency, you and your team will be able to follow some straightforward and simple guidelines for selecting sufficiently interpretable but optimally performing algorithmic techniques.

Before exploring these guidelines, it is necessary to provide you with some background information to help you better understand what facets of explanation are actually involved in technically interpretable AI. A good grasp of what is actually needed from such an explanation will enable you to effectively target the interpretability needs of your AI project.

**Facets of explanation in technically interpretable AI:** A good starting point for understanding how the technical dimension of explanation works in interpretable AI systems is to remember that these systems are largely mathematical models that carry out step-by-step computations in transforming sets of statistically interacting or independent inputs into sets of target outputs. Machine learning is, at bottom, just applied statistics and probability theory fortified with several other mathematical techniques. As such, it is subject to same methodologically rigorous requirements of logical validation as other mathematical sciences.

Such a demand for rigour informs the facet of **formal and logical explanation of AI systems** that is sometimes called the **mathematical glass box**. This characterisation refers to the transparency of strictly formal explanation: No matter how complicated it is (even in the case of a deep neural net with a hundred million parameters), an algorithmic model is a closed system of effectively computable operations where rules and transformations are mechanically applied to inputs to determine outputs. In this restricted sense, all AI and machine learning models are fully intelligible and mathematically transparent if only **formally and logically** so.

This is an important characteristic of AI systems, because it makes it possible for supplemental and eminently interpretable computational approaches to model, approximate, and simplify even the most complex and high dimensional among them. In fact, such a possibility fuels some of the technical approaches to interpretable AI that will soon be explored.

This formal way of understanding the technical explanation of AI and machine learning systems, however, has immediate limitations. It can tell us that a model is mathematically intelligible because it operates according to a collection of fixed operations and parameters, but it cannot tell us much about how or why the components of the model transformed a specified group of inputs into their corresponding outputs. It cannot tell us anything about the *rationale behind the algorithmic generation of a given outcome*.

This second dimension of technical explanation has to do with the *semantic facet* of interpretable AI. A **semantic explanation** offers an interpretation of the functions of the individual parts of the

algorithmic system in the generation of its output. Whereas formal and logical explanation presents an account of the stepwise application of the procedures and rules that comprise the formal framework of the algorithmic system, semantic explanation helps us to understand the meaning of those procedures and rules in terms of their purpose in the input-output mapping operation of the system, i.e. what role they play in determining the outcome of the model's computation.

The difficulties surrounding the interpretability of algorithmic decisions and behaviours arise in this semantic dimension of technical explanation. It is easiest to illustrate this by starting from the simplest case.

When a machine learning model is very basic, the task of following the rationale of how it transforms a given set of inputs into a given set of outputs can be relatively unproblematic. For instance, in the simple linear regression,  $y = a + bx + \varepsilon$ , with a single predictor variable  $x$  and a response variable  $y$ , the predictive relationship of  $x$  to  $y$  is directly expressed in a regression coefficient  $b$ , representing the rate and direction at which  $y$  is predicted to change as  $x$  changes. This hypothetical model is completely interpretable from the technical perspective for the following reasons:

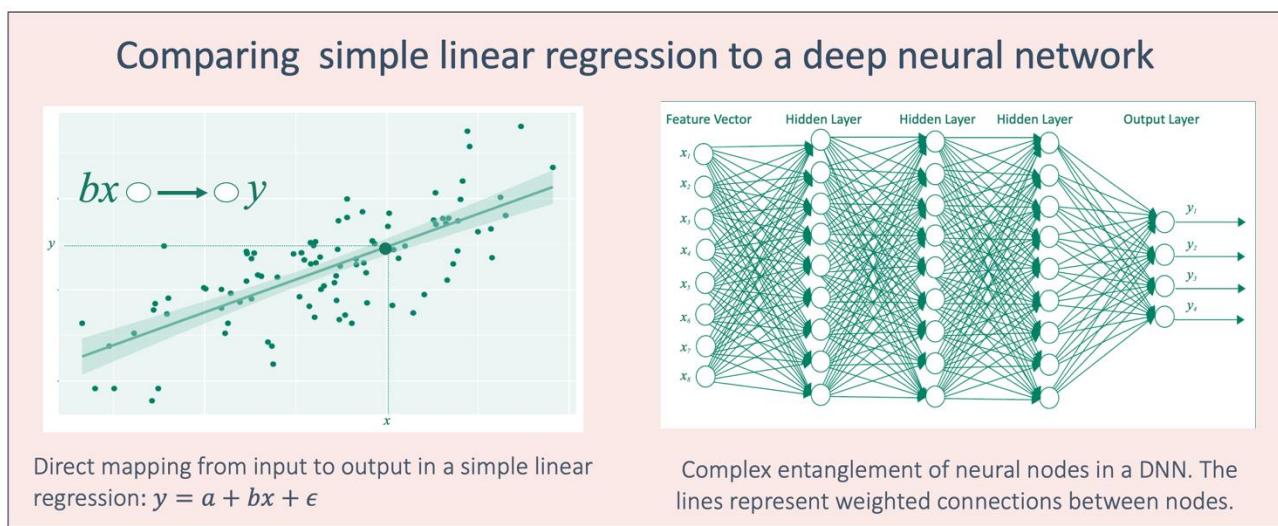
- **Linearity:** Any change in the value of the predictor variable is directly reflected in a change in the value of the response variable at a constant rate  $b$ . The interpretable prediction yielded by the model can therefore be directly inferred. This linearity dimension of predictive models has been an essential feature of the automated decision-making systems in many heavily regulated and high-impact sectors, because the predictions yielded have high inferential clarity and strength.
- **Monotonicity:** When the value of the predictor changes in a given direction, the value of the response variable changes consistently either in the same or opposite direction. The interpretable prediction yielded by the model can thus be directly inferred. This monotonicity dimension is also a highly desirable interpretability condition of predictive models in many heavily regulated sectors, because it incorporates reasonable expectations about the consistent application of sector specific selection constraints into automated decision-making systems. So, for example, if the selection criteria to gain employment at an agency or firm includes taking an exam, a reasonable expectation of outcomes would be that if candidate A scored better than candidate B, then candidate B, all other things being equal, would not be selected for employment when A is not. A monotonic predictive model that uses the exam score as the predictor variable and application success as the response variable would, in effect, guarantee this expectation is met by disallowing situations where A scores better than B but B gets selected and A does not.
- **Non-Complexity:** The number of features (dimensionality) and feature interactions is low enough and the mapping function is simple enough to enable a clear 'global' understanding of the function of each part of the model in relation to its outcome.

While, all three of these desirable interpretability characteristics of the imagined model allow for direct and intuitive reasoning about the relation of the predictor and response variables, the model itself is clearly too minimal to capture the density of relationships and interactions between attributes in complex real-world situations where some degree of noisiness is unavoidable and the task of apprehending the subtleties of underlying data distributions is tricky.

In fact, one of the great strides forward that has been enabled by the contemporary convergence of expanding computing power and big data availability with more advanced machine learning models has been exactly this capacity to better capture and model the intricate and complicated dynamics of real-world situations. Still, this incorporation of the complexity of scale into the models themselves has also meant significant challenges to the semantic dimension of the technical explanation of AI systems.

As machine learning systems have come to possess both ever greater access to big data and increasing computing power, their designers have correspondingly been able both to enlarge the feature spaces (the number of input variables) of these systems and to turn to gradually more complex mapping functions. In many cases, this has meant vast improvements in the predictive and classificatory performance of more accurate and expressive models, but this has also meant the growing prevalence of non-linearity, non-monotonicity, and high-dimensional complexity in an expanding array of so-called ‘black-box’ models.

Once high-dimensional feature spaces and complex functions are introduced into machine learning systems, the effects of changes in any given input become so entangled with the values and interactions of other inputs that understanding how individual components are transformed into outputs becomes extremely difficult. The complex and unintuitive curves of the decision functions of many of these models preclude linear and monotonic relations between their inputs and outputs. Likewise, the high-dimensionality of their optimisation techniques—frequently involving millions of parameters and complex correlations—ranges well beyond the limits of human-scale cognition and understanding. To illustrate the increasing complexity involved in comprehending input-output mappings, here is a visual representation that depicts the difference of between a linear regression function and a deep neural network:



These rising tides of computational complexity and algorithmic opacity consequently pose a key challenge for the responsible design and deployment of safe, fair, and ethical AI systems: how should the potential to advance the public interest through the implementation of high performing but increasingly uninterpretable machine learning models be weighed against the tangible risks posed by the lack of interpretability of such systems?

A careful answer to this question is, in fact, not so simple. While the trade-off between performance and interpretability may be real and important in *some domain-specific applications*, in others there exist increasingly sophisticated developments of standard interpretable techniques such as regression extensions, decision trees, and rule lists that may prove just as effective for use cases where the need for transparency is paramount. Furthermore, supplemental interpretability tools, which function to make ‘black box’ models more semantically and qualitatively explainable are rapidly advancing day by day.

These are all factors that you and your team should consider as you work together to decide on which models to use for your AI project. As a starting point for those considerations, let us now turn to some basic guidelines that may help you to steer that dialogue toward points of relevance and concern.

### Guidelines for designing and delivering a sufficiently interpretable AI system

You should use the table below to begin thinking about how to integrate interpretability into your AI project. While aspects of this topic can become extremely technical, it is important to make sure that dialogue about making your AI system interpretable remains multidisciplinary and inclusive. Moreover, it is crucial that key stakeholders be given adequate consideration when deciding upon the delivery mechanisms of your project. These should include policy or operational design leads, the technical personnel in charge of operating the trained models, the implementers of the models, and the decision subjects, who are affected by their outcomes.

Note that the first three guidelines focus on the big picture issues you will need to consider in order to incorporate interpretability needs into your project planning and workflow, whereas the last two guidelines shift focus to the user-centred requirements of designing and implementing a sufficiently interpretable AI system.

#### Guidelines for designing and delivering a sufficiently interpretable AI system

##### **Guideline 1: Look first to context, potential impact, and domain-specific need when determining the interpretability requirements of your project**

There are several related factors that should be taken into account as you formulate your project’s approach to interpretability:

- 1. Type of application:** Start by assessing both the kind of tool you are building and the environment in which it will apply. Clearly there is a big difference between a computer vision system that sorts handwritten employee feedback forms and one that sorts safety risks at a security checkpoint. Likewise, there is a big difference between a random forest model that triages applicants at a licencing agency and one that triages sick patients in an emergency department.

Understanding your AI system’s purpose and context of application will give you a better idea of the stakes involved in its use and hence also a good starting point to think about the scope of its interpretability needs. For instance, low-stakes AI models that are

not safety-critical, do not directly impact the lives of people, and do not process potentially sensitive social and demographic data will likely have a lower need for extensive resources to be dedicated to a comprehensive interpretability platform.

2. **Domain specificity:** By acquiring solid domain knowledge of the environment in which your AI system will operate, you will gain better insight into any potential sector-specific standards of explanation or benchmarks of justification which should inform your approach to interpretability. Through such knowledge, you may also obtain useful information about organisational and public expectations regarding the scope, content, and depth of explanations that have been previously offered in relevant use cases.
3. **Existing technology:** If one of the purposes of your AI project is to replace an existing algorithmic technology that may not offer the same sort of expressive power or performance level as the more advanced machine learning techniques that you are planning to deploy, you should carry out an assessment of the performance and interpretability levels of the existing technology. Acquiring this knowledge will provide you with an important reference point when you are considering possible trade-offs between performance and interpretability that may occur in your own prospective system. It will also allow you to weigh the costs and benefits of building a more complex system with higher interpretability-support needs in comparison to the costs and benefits of using a simpler model.

## Guideline 2: Draw on standard interpretable techniques when possible

In order to actively integrate the aim of sufficient interpretability into your AI project, your team should approach the model selection and development process with the goal of finding the right fit between **(1) domain-specific risks and needs, (2) available data resources and domain knowledge, and (3) task appropriate machine learning techniques**. Effectively assimilating these three aspects of your use case requires open-mindedness and practicality.

Often times, it may be the case that high-impact, safety-critical, or other potentially sensitive environments heighten demands for the thoroughgoing accountability and transparency of AI projects. In some of these instances, such demands may make choosing standard but sophisticated non-opaque techniques an overriding priority. These techniques may include **decisions trees, linear regression and its extensions like generalised additive models, decision/rule lists, case-based reasoning, or logistic regression**. In many cases, reaching for the ‘black box’ model first may not be appropriate and may even lead to inefficiencies in project development, because more interpretable models, which perform very well but do not require supplemental tools and techniques for facilitating interpretable outcomes, are also available.

Again, solid domain knowledge and context awareness are key components here. In use cases where data resources lend to well-structured, meaningful representations and domain expertise can be incorporated into model architectures, interpretable techniques may often be more desirable than opaque ones. Careful data pre-processing and iterative model development can, in these cases, hone the accuracy of such interpretable systems in ways that may make the advantages gained by the combination of their performance and transparency outweigh the benefits of more semantically intransparent approaches.

In other use cases, however, data processing needs may disqualify the deployment of these sorts of straightforward interpretable systems. For instance, when AI applications are sought for classifying images, recognising speech, or detecting anomalies in video footage, the most effective machine learning approaches will likely be opaque. The feature spaces of these kinds of AI systems grow exponentially to hundreds of thousands or even millions of dimensions. At this scale of complexity, conventional methods of interpretation no longer apply. Indeed, it is the unavoidability of hitting such an **interpretability wall** for certain important applications of supervised, unsupervised, and reinforcement learning that has given rise to an entire subfield of machine learning research which focuses on providing technical tools to facilitate interpretable and explainable AI.

When the use of ‘black box’ models best fits the purpose of your AI project, you should proceed diligently and follow the procedures recommended in Guideline 3. For clarity, let us define a ‘black box’ model as **any AI system whose innerworkings and rationale are opaque or inaccessible to human understanding**. These systems may include **neural networks** (including recurrent, convolutional, and deep neural nets), **ensemble methods** (an algorithmic technique such as the random forest method that strengthens an overall prediction by combining and aggregating the results of several or many different base models), and **support vector machines** (a classifier that uses a special type of mapping function to build a divider between two sets of features in a high dimensional feature space).

**Guideline 3: When considering the use of ‘black box’ AI systems, you should:**

1. Thoroughly weigh up impacts and risks;
2. Consider the options available for supplemental interpretability tools that will ensure a level of semantic explanation which is both *domain appropriate* and *consistent with the design and implementation of safe, fair, and ethical AI*;
3. Formulate an interpretability action plan, so that you and your team can put adequate forethought into how explanations of the outcomes of your system’s decisions, behaviours, or problem-solving tasks can be optimally provided to users, decision subjects, and other affected parties.

It may be helpful to explore each of these three suggested steps of assessing the viability of the responsible design and implementation of a ‘black box’ model in greater detail.

**(1) Thoroughly weigh up impacts and risks:** Your first step in evaluating the feasibility of using a complex AI system should be to focus on issues of ethics and safety. As a general policy, you and your team should utilise ‘black box’ models only:

- where their potential impacts and risks have been thoroughly considered in advance, and you and your team have determined that your use case and domain specific needs support the responsible design and implementations of these systems;

- where supplemental interpretability tools provide your system with a domain appropriate level of semantic explainability that is reasonably sufficient to mitigate its potential risks and that is therefore consistent with the design and implementation of safe, fair, and ethical AI.

**(2) Consider the options available for supplemental interpretability tools:** Next, you and your team should assess whether there are technical methods of explanation-support that **both** satisfy the specific interpretability needs of your use case as determined by the deliberations suggested in Guideline 1 **and** are appropriate for the algorithmic approach you intend to use. You should consult closely with your technical team at this stage of model selection. The exploratory processes of trial-and-error, which often guide this discovery phase in the innovation lifecycle, should be informed and constrained by a solid working knowledge of the technical art of the possible in the domain of available and useable interpretability approaches.

The task of lining up the model selection process with the demands of interpretable AI requires a few conceptual tools that will enable thoughtful evaluation of whether proposed supplemental interpretability approaches sufficiently meet your project's explanatory needs. First and most importantly, you should be prepared to ask the right questions when evaluating any given interpretability approach. This involves establishing with as much clarity as possible **how the explanatory results of that approach can contribute to the user's ability to offer solid, coherent, and reasonable accounts of the rationale behind any given algorithmically generated output**. Relevant questions to ask that can serve this end are:

- What sort of explanatory resources will the interpretability tool provide users and implementers in order (1) to enable them to exercise better-informed evidence-based judgments and (2) to assist them in offering plausible, sound, and reasonable accounts of the logic behind algorithmically generated output to affected individuals and concerned parties?
- Will the explanatory resources that the interpretability tool offers be useful for providing affected stakeholders with a sufficient understanding of a given outcome?
- How, if at all, might the explanatory resources offered by the tool be misleading or confusing?

You and your team should take these questions as a starting point for evaluating prospective interpretability tools. These tools should be assessed in terms of their capacities to render the reasoning behind the decisions and behaviours of the uninterpretable 'black box' systems sufficiently intelligible to users and affected stakeholders given use case and domain specific interpretability needs.

Keeping this in mind, there are two technical dimensions of supplemental interpretability approaches that should be systematically incorporated into evaluation processes at this stage of the innovation workflow.

The first involves the possible **explanatory strategies** you choose to pursue over the course of the design and implementation lifecycle. Such strategies will largely determine the paths to understanding you will be able to provide for its users and decision subjects. They will largely define *how you explain your model and its outcomes* and hence *what kinds of explanation you are able offer*.

The second involves the **coverage and scope** of the actual explanations themselves. The choices you make about explanatory coverage will determine the extent to which the kinds of explanations you are planning to pursue will address *single instances* of the model's outputs or range more broadly to cover the *underlying rationale of its behaviour in general and across instances*. Choices you make about explanatory coverage will largely govern the extent to which your AI system is locally and/or globally interpretable.

The very broad-brushed overview of these two dimensions that follows is just meant to orient you to some of the basic concepts in an expanding field of research, so that you are more prepared for working with your technical team to think through the strengths and weaknesses of various approaches. Note, additionally, that this is a rapidly developing area. Relevant members of your team should keep abreast of the latest developments in the field of interpretable AI or XAI (Explainable AI):

#### Two technical dimensions of supplemental interpretability approaches:

1. **Determining explanatory strategies:** To achieve the goal of securing a sufficiently interpretable AI system, you and your team will need to get clear on **how to explain** your model and its outcomes. The explanatory strategies you decide to pursue will shape the paths to understanding you are able to provide for the users of your model and for its decision subjects.

There are four such explanatory strategies to which you should pay special attention:

- a) **Internal explanation:** Pursuing the internal explanation of an opaque model involves making intelligible how the components and relationships within it function. There are two ways that such a goal of internal explanation can be interpreted. On the one hand, it can be seen as an endeavour to explain the operation of the model by considering it globally *as a comprehensible whole*. Here, the aspiration is to 'pry open the black box' by building an explanatory model that enables a full grasp of the opaque system's internal contents. The strengths and weaknesses of such an approach will be discussed in the next section on global interpretability.

On the other hand, the search for internal explanation can indicate the pursuit of a kind of **engineering insight**. In this sense, internal explanation can be seen as attempting to shed descriptive and inferential light on the parts and operation of the system as a whole in order to try to make it work better. Acquiring this sort of internal understanding of the more general relationships that the working parts of a trained model have with patterns of its responses can allow researchers to advance step-by-step in gaining a better data scientific grasp on

why it does what it does and how to improve it. Similarly, this type of internal explanation can be seen as attempting to shed light on an opaque model's operation by breaking it down into more understandable, analysable, and digestible parts (for instance, in the case of a DNN: into interpretable characteristics of its vectors, features, layers, parameters, etc.).

From a practical point of view, this kind of aspiration to *engineering insight* in the ends of data scientific advancement should inform the goals of your technical team throughout the model selection and design workflow.

Numerous methods exist to help provide informative representations of the innerworkings of various 'black box' systems. Gaining a clearer descriptive understanding of the internal composition of your system will contribute greatly to your project's ability to achieve a higher degree of outcome transparency and to its capacity to foster best practices in the pursuit of responsible data science in general.

- b) ***External or post-hoc explanation:*** External or post-hoc explanation attempts to capture essential attributes of the observable behaviour of a 'black box' system by subjecting it to a number of different techniques that reverse engineer explanatory insight. Some post-hoc approaches test the sensitivity of the outputs of an opaque model to perturbations in its inputs; others allow for the interactive probing of its behavioural characteristics; others, still, build proxy-based models that utilise simplified interpretable techniques to gain a better understanding of particular instances of its predictions and classifications.

This external or post-hoc approach has, at present, established itself in machine learning research as a go-to explanatory strategy and for good reason. It allows data scientists to pose mathematical questions to their opaque systems by testing them and by building supplemental models which enable greater insight through the inferences drawn from their experimental interventions. Such a post-hoc approach allows them, moreover, to seek out evidence for the reasoning behind a given opaque model's prediction or classification by utilising maximally interpretable techniques like linear regression, decision trees, rule lists, or case-based reasoning. Several examples of post-hoc explanation will be explored below in the section on local interpretability.

Take note initially though that, as some critics have rightly pointed out, because they are approximations or simplified supplemental models of the more complex originals, many post-hoc explanations can fail to accurately represent certain areas of the opaque model's feature space. This deterioration of accuracy in parts of the original model's domain can frequently produce misleading and uncertain results in the post-hoc explanations of concern.

- c) ***Supplemental explanatory infrastructure:*** A different kind of explanatory strategy involves actually incorporating secondary explanatory facilities into the system you are building. For instance, an image recognition system could have a primary component, like a convolutional neural net, that extracts features from

its inputs and classifies them while a secondary component, like a built-in recurrent neural net with an ‘attention-directing’ mechanism, translates the extracted features into a natural language representation that produces a sentence-long explanation of the result to the user. In other words, a system like this is designed to provide simple explanations of its own data processing results.

Research into integrating ‘attention-based’ interfaces like this in AI systems is continuing to advance toward making their implementations more sensitive to user needs, more explanation-forward, and more human-understandable. For instance, multimodal methods of combining visualisation tools and textual interface are being developed that may make the provision of explanations more interpretable for both implementers and decision subjects. Furthermore, the incorporation of domain knowledge and logic-based or convention-based structures into the architectures of complex models are increasingly allowing for better and more user-friendly representations and prototypes to be built into them. This is gradually enabling more sophisticated explanatory infrastructures to be integrated into opaque systems and makes it essential to think about building explanation-by-design into your AI projects.

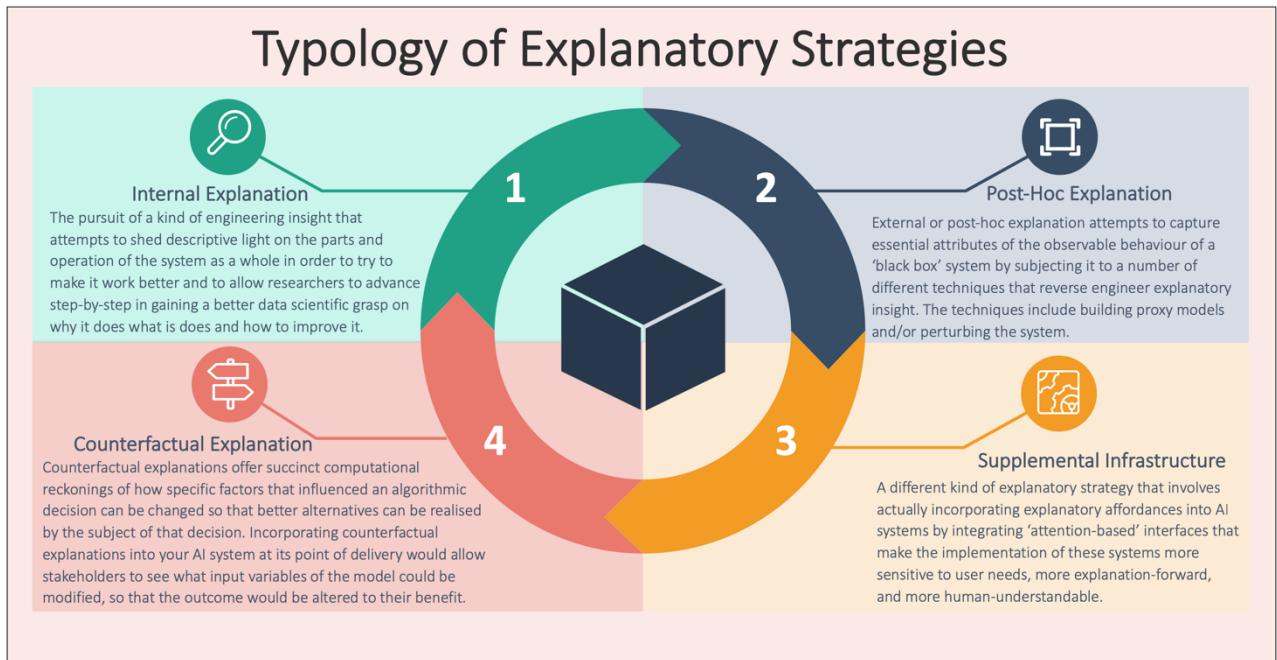
- d) ***Counterfactual explanation:*** While counterfactual explanation is a kind of post-hoc approach, it deserves special attention insofar as it moves beyond other post-hoc explanations to provide affected stakeholders with clear and precise options for actionable recourse and practical remedy.

Counterfactual explanations are contrastive explanations: They offer succinct computational reckonings of how specific factors that influenced an algorithmic decision can be changed so that better alternatives can be realised by the subject of that decision. Incorporating counterfactual explanations into your AI system at its point of delivery would allow stakeholders to see what input variables of the model can be modified, so that the outcome could be altered to their benefit. Additionally, from a responsible design perspective, incorporating counterfactual explanation into the development and testing phases of your system would allow your team to build a model that incorporates ***actionable variables***, i.e. input variables that will afford decision subjects with concise options for making practical changes that would improve their chances of obtaining the desired outcome. **Counterfactual explanatory strategies can be used as way to incorporate reasonableness and the encouragement of agency into the design and implementation of your AI project.**

All that said, it is important to recognise that, while counterfactual explanation does offer an innovative way to contrastively explore how feature importance may influence an outcome, it is not a complete solution to the problem of AI interpretability. In certain cases, for instance, the sheer number of potentially significant features that could be at play in counterfactual explanations of a given result can make a clear and direct explanation difficult to obtain and selected sets of explanations seem potentially arbitrary. Moreover, there are as

yet limitations on the types of datasets and functions to which these kinds of explanations are applicable. Finally, because this kind of explanation concedes the opacity of the algorithmic model outright, it is less able to address concerns about potentially harmful feature interactions and multivariate relationships that may be buried deep within the model's architecture.

Here is an at-a-glance view of a typology of these explanatory strategies:



2. **Coverage and Scope:** The main questions you will need to broach in the dimension of the coverage and scope of your supplemental interpretability approach are: To what extent does our interpretability approach cover the explanation of *singe predictions or classifications* of the model and to what extent does it cover the explanation of the *innerworkings and rationale of the model as a whole and across predictions*? To what extent does it cover both?

This distinction between single instance and total model explanation is often characterised as the difference between **local interpretability** and the **global interpretability**. Both types of explanation offer potentially helpful support for the provision of significant information about the rationale behind an algorithmic decision or behaviour, but both, in their own ways, also face difficulties.

**Local Interpretability:** A local semantic explanation aims to enable the interpretability of **individual cases**. The general idea behind attempts to explain a 'black box' system in terms of specific instances is that, regardless of how complex the architecture or decision function of that system may be, it is possible to gain interpretive insight into its innerworkings by focusing on single data points or neighbourhoods in its feature space. In other words, even if the high dimensionality and curviness of a model makes it opaque *as a whole*, there is an expectation that insight-generating

interpretable methods can be applied *locally* to smaller sections of the model, where changes in isolated or grouped variables are more manageable and understandable.

This general explanatory perspective has yielded several different interpretive strategies that have been successfully applied in significant areas of ‘black box’ machine learning. One family of such strategies has zeroed in on neural networks (DNNs, in particular) by identifying what features of an input vector’s data points make it representative of the target concept that a given model is trying to classify. So, for example, if we have a digital image of a dog that is converted into a vector of pixel values and then processed it through a dog-classifying deep neural net, this interpretive approach will endeavour to tell us why the system yielded a ‘dog-positive’ output by isolating the slices of this set of data points that are most relevant to its successful classification by the model.

This can be accomplished in several related ways. What is called **sensitivity analysis** identifies the most relevant features of an input vector by calculating local gradients to determine how a data point has to be moved to change the output label. Here, an output’s sensitivity to such changes in input values identifies the most relevant features. Another method to identify feature relevance that is downstream from sensitivity analysis is called **salience mapping**, where a strategy of moving backward through the layers of a neural net graph allows for the mapping of patterns of high activation in the nodes and ultimately generates interpretable groupings of salient input variables that can be visually represented in a heat or pixel attribution map.

A second local interpretive strategy also seeks to explain feature importance in a single prediction or classification by perturbing input variables. However, instead of using these nudges in the feature space to highlight areas of saliency, it uses them to prod the opaque model in the area around the relevant prediction, so that a supplemental interpretable model can be constructed which establishes the relative importance of features in the black box model’s output.

The most well-known example of this strategy is called **LIME (Local Interpretable Model-Agnostic Explanation)**. LIME works by fitting an interpretable model to a specific prediction or classification produced by the opaque system of concern. It does this by sampling data points at random around the target prediction or classification and then using them to build a local approximation of the decision boundary that can account for the features which figure prominently in the specific prediction or classification under scrutiny.

The way this works is relatively uncomplicated: LIME generates a simple linear regression model by weighting the values of the data points, which were produced by randomly perturbing the opaque model, according to their proximity to the original prediction or classification. The closest of these values to the instance being explained are weighted the heaviest, so that the supplemental model can produce an explanation of feature importance that is **locally faithful** to that instance. Note that the type of model that LIME uses most prominently is a sparse linear regression

function for reasons of semantic transparency that were discussed above. Other interpretable models such as decision trees can likewise be employed.

While LIME does indeed appear to be a step in the right direction for the future of interpretable AI, a host of issues that present challenges to the approach remains unresolved. For instance, the crucial aspect of how to properly define the proximity measure for the ‘neighbourhood’ or ‘local region’ where the explanation applies remains unclear, and small changes in the scale of the chosen measure can lead to greatly diverging explanations. Likewise, the explanation produced by the supplemental linear model can quickly become unreliable even with small and virtually unnoticeable perturbations of the system it is attempting to approximate. This challenges the basic assumption that there is always some simplified linear model that successfully approximates the underlying model reasonably well near any given data point.

LIME’s creators have largely acknowledged these shortcomings and have recently offered a new explanatory approach that they call ‘**anchors**’. These ‘high precision rules’ incorporate into their formal structures ‘reasonable patterns’ that are operating within the underlying model (such as the implicit linguistic conventions that are at work in a sentiment prediction model), so that they can establish suitable and faithful boundaries of their explanatory coverage of its predictions or classifications.

A related and equally significant local interpretive strategy is called **SHAP (Shapley Additive exPlanations)**. SHAP uses concepts from game theory to define a ‘Shapley value’ for a feature of concern that provides a measurement of its influence on the underlying model’s prediction. Broadly, this value is calculated for a feature by averaging its marginal contribution to *every possible prediction* for the instance under consideration.

This might seem impossible, but the strategy is straightforward. SHAP calculates the marginal contribution of the relevant feature for all possible combinations of inputs in the feature space of the instance. So, if the opaque model that it is explaining has 15 features, SHAP would calculate the marginal contribution of the feature under consideration 32,768 times (i.e. one calculation for each combination of all possible combinations of features:  $2^{15}$ , or  $2^k$  when  $k = 15$ ).

This method then allows SHAP to estimate the Shapley values for all input features in the set to produce the complete distribution of the prediction for the instance. In our example, this would entail 491,520 calculations. While such a procedure is computationally burdensome and becomes intractable beyond a certain threshold, this means that *locally*, that is, for the calculation of the specific instance, SHAP can axiomatically guarantee the consistency and accuracy of its reckoning of the marginal effect of the feature. (Note that the SHAP platform does offer methods of approximation to avoid this excessive computational expense.)

Despite this calculational robustness, SHAP also faces some of the same kinds of difficulties that LIME does. The way SHAP calculates marginal contributions is by

constructing two instances: the first instance includes the feature being measured while the second leaves it out. After calculating the prediction for each of these instances by plugging their values into the underlying model, the result of the second is subtracted from that of the first to determine the marginal contribution of the feature. This procedure is then repeated for all possible combinations of features so that the weighted average of all of the marginal contributions of the feature of concern can be computed.

The contestable part of this process comes with how SHAP defines the *absence* of variables under consideration. To leave out a feature—whether it's the one being directly measured or one of the others not included in the combination under consideration—SHAP replaces it with a *stand-in feature value* drawn from a selected donor sample (that is itself drawn from the existing dataset). This method of sampling values assumes feature independence (i.e. that values sampled are not correlated in ways that might significantly affect the output for a particular calculation). As a consequence, the interaction effects engendered by and between stand-in variables are necessarily unaccounted for when conditional contributions are approximated. The result is the introduction of uncertainty into the explanation that is produced because the complexity of multivariate interactions in the underlying model may not be sufficiently captured by the simplicity of this supplemental interpretability technique. This drawback in sampling (as well as a certain degree of arbitrariness in domain definition) can cause SHAP to become unreliable even with minimal perturbations of the model it is approximating.

Despite these limitations in the existing tools of local interpretability, it is important that you think ‘local-first’ when considering the issue of the coverage and scope of the explanatory approaches you plan to incorporate into your project. Being able to provide explanations of specific predictions and classifications is of paramount importance both to securing optimal outcome transparency and also to ensuring that your AI system will be implemented responsibly and reasonably.

**Global interpretability:** The motivation behind the creation of local interpretability tools like LIME or SHAP (as well as many others not mentioned here) has derived, at least in part, from a need to find a way of avoiding the kind of difficult *double bind* faced by the alternative approach to the coverage and scope of interpretable AI: global interpretability.

On the prevailing view, providing a global explanation of a ‘black box’ model entails offering an alternative interpretable model that captures the innerworkings and logic of a ‘black box’ model *in sum* and across predictions or classifications. The difficulty faced by global interpretability arises in the seemingly unavoidable trade-off between the need for the global explanatory model to be sufficiently simple so that it is understandable by humans and the need for that model to be sufficiently complex so that it can capture the intricacies of how the mapping function of a ‘black box’ model works as a whole.

While this is clearly a real problem that appears to be theoretically inevitable, it is important to keep in mind that, *from a practical standpoint*, a serviceable notion of global interpretability need not be limited to such a conceptual puzzle. There are at least two less ambitious but more constructive ways to view global interpretability as a potentially meaningful contributor to the responsible design and implementation of interpretable AI.

First, many useful attempts have already been made at building explanatory models that employ interpretable methods (like decision trees, rule lists, and case-based classification) to globally approximate neural nets, tree ensembles, and support vector machines. These results have enabled a deeper understanding of the way human interpretable logics and conventions (like if-then rules and representationally generated prototypes) can be measured against or mapped onto high dimensional computational structures and even allow for some degree of targeted comprehensibility of the logic of their parts.

This capacity to ‘peek into the black box’ is of great practical importance in domains where trust, user-confidence, and public acceptance are critical for the realisation optimal outcomes. Moreover, this ability to move back and forth between interpretable architectures and high-dimensional processing structures can enable knowledge discovery as well as insights into the kinds of dataset-level and population-level patterns, which are crucial for well-informed macroscale decision-making in areas ranging from public health and economics to the science of climate change.

Being able to uncover global effects and relationships between complex model behaviour and data distributions at the demographic and ecological level may prove vital for establishing valuable and practically useful knowledge about unobservable but significant biophysical and social configurations. Hence, although these models have not solved the understandability-complexity puzzle as such, they have opened up new pathways for innovative thinking in the applied data sciences that may be of immense public benefit in the future.

Secondly, as mentioned above, under the auspices of the aspiration to **engineering insight**, a *descriptive and analytical kind of global interpretability* can be seen as a driving force of data scientific advancement. When seen through a practitioner-centred lens, this sort of global interpretability allows data scientists to take a wide-angled and discovery-oriented view of a ‘black box’ model’s relationship to patterns that arise across the range of its predictions. Figuring out how an opaque system works and how to make it work better by more fully understanding these patterns is a continuous priority of good research. So too is understanding the relevance of features and of their complex interactions through dataset level measurement and analysis. These dimensions of incorporating the explanatory aspirations of global interpretability into best practices of research and innovation should be encouraged in your AI project.

(3) **Formulate an interpretability action plan:** The final step you will need to take to ensure a responsible approach to using ‘black box’ models is to formulate an interpretability action plan so that you and your team can put adequate forethought into how explanations of the outcomes of your system’s decisions, behaviours, or problem-solving tasks can be optimally provided to users, decision subjects, and other affected parties.

This action plan should include the following:

- A **clear articulation of the explanatory strategies** your team intends to use and a detailed plan that indicates the stages in the project workflow when the design and development of these strategies will need to take place.
- A succinct formulation of your **explanation delivery strategy**, which addresses the special provisions for clear, simple, and user-centred explication that are called for when supplemental interpretability tools for ‘black box’ models are utilised. See more about delivery and implementation in Guideline 5.
- A **detailed timeframe for evaluating your team’s progress** in executing its interpretability action plan and a **role responsibility list**, which maps in detail the various task-specific responsibilities that will need to be fulfilled to execute the plan.

#### Guideline 4: Think about interpretability in terms of the capacities of human understanding

When you begin to deliberate about the specific scope and content of your interpretability platform, it is important to reflect on what it is that you are exactly aiming to do in making your model sufficiently interpretable. A good initial step to take in this process is to think about what makes even the simplest explanations **clear and understandable**. In other words, you should begin by thinking about interpretability in terms of the capacities and limitations of human cognition.

From this perspective, it becomes apparent that even the most straightforward model like a linear regression function or a decision tree can become uninterpretable when its dimensionality presses beyond the cognitive limits of a thinking human. Recall our example of the simple linear regression:  $y = a + bx + \epsilon$ . In this instance, only one feature  $x$  relates to the response variable  $y$ , so understanding the predictive relationship is easy. The model is parsimonious.

However, if we started to add more features as covariates, even though the model would remain linear and hence intuitively predictable, being able to understand the relationship between the response variable and all the predictors and their coefficients (feature weights) would quickly become difficult. So, say we added ten thousand features and trained the model:  $y = a + b_0x_0 + b_1x_1 + \dots + b_{10000}x_{10000} + \epsilon$ . Understanding *how* this model’s prediction comes about—what role each of the individual parts play in producing the prediction—would become difficult because of a certain cognitive limit in the quantity of entities that human thinking can handle at any given time. This model would lose a significant degree of interpretability.

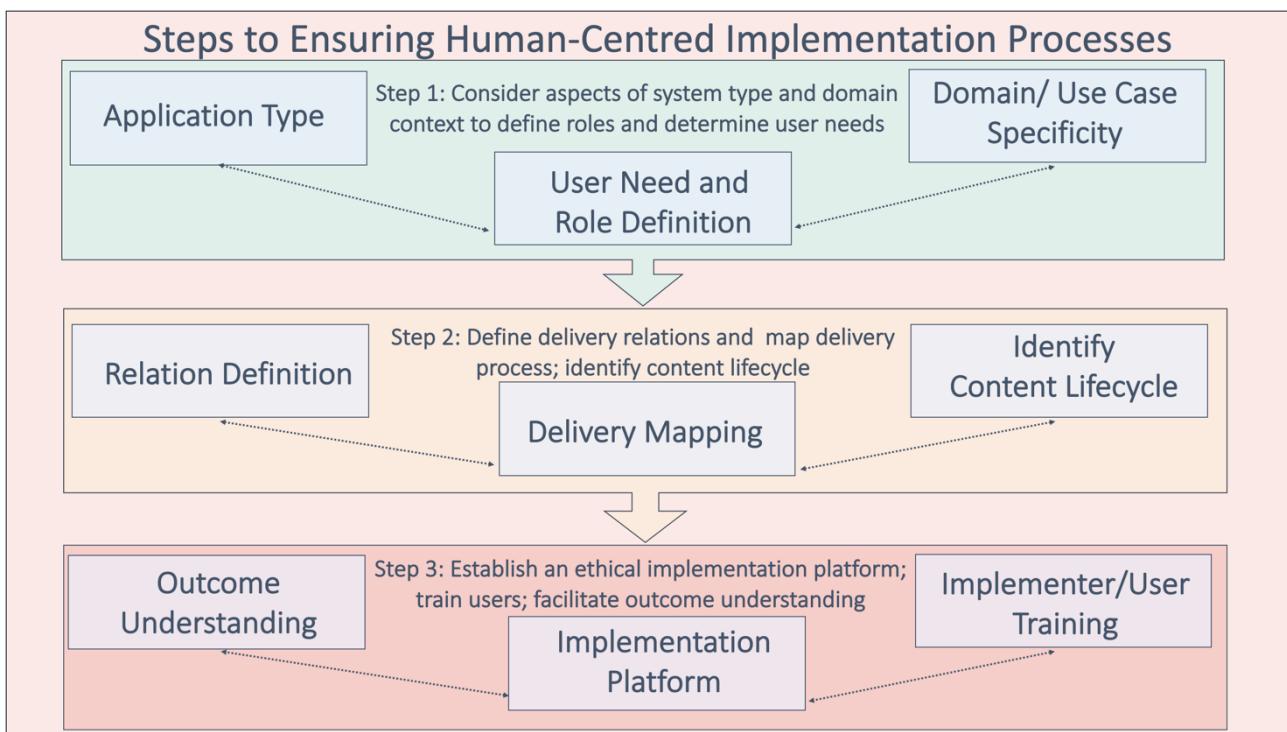
Seeing interpretability as a continuum of comprehensibility that is dependent on the capacities and limits of the individual human interpreter should key you in to what is needed in order to deliver an interpretable AI system. Such limits to consider should include not only cognitive

boundaries but also varying levels of access to relevant vocabularies of explanation; an explanation about the results of a trained model that uses a support vector machine to divide a 26-dimensional feature space with a planar separator, for instance, may be easy to understand for a technical operator or auditor but entirely inaccessible to a non-specialist. Offering good explanations should take expertise level into account. **Your interpretability platform should be cognitively equitable.**

## Securing responsible delivery through human-centred implementation protocols and practices

The demand for sensitivity to human factors should inform your approach to devising delivery and implementation processes from start to finish. To provide clear and effective explanations about the content and rationale of algorithmic outputs, you will have to begin by building ***from the human ground up***. You will have to pay close attention to the circumstances, needs, competences, and capacities of the people whom your AI project aims to assist and serve.

This means that ***context will be critical***. By understanding your use case well and by drawing upon solid domain knowledge, you will be better able to **define roles and relationships**. You will better be able to **train the users and implementers of your system**. And, you will be better able to **establish an effectual implementation platform, to clarify content, and to facilitate understanding of outcomes** for users and affected stakeholders alike. Here is a diagram of what securing human-centred implementation protocols and practices might look like:



Let us consider these steps in turn by building a checklist of essential actions that should be taken to help ensure the human-centred implementation of your AI project. Because the specifics of your approach will depend so heavily on the context and potential impacts of your project, we'll assume a

generic case and construct the checklist around a hypothetical algorithmic decision-making system that will be used for predictive risk assessment.

### Step 1: Consider aspects of application type and domain context to define roles and determine user needs

- (1) Assess which members of the communities you are serving will be most affected by the implementation of your AI system. Who are the most vulnerable among them? How will their socioeconomic, cultural, and education backgrounds affect their capacities to interpret and understand the explanations you intend to provide? How can you fine-tune your explanatory strategy to accommodate their requirements and provide them with clear and non-technical details about the rationale behind the algorithmically supported result?

When thinking about providing explanations to affected stakeholders, you should start with the needs of the most disadvantaged first. Only in this way, will you be able to establish an acceptable baseline for the equitable delivery of interpretable AI.

- (2) After reviewing [Guideline 1](#) above, make a list of and define all the roles that will potentially be involved at the delivery stage of your AI project. As you go through each role, specify levels of technical expertise and domain knowledge as well as possible goals and objectives for each role. For instance, in our predictive risk assessment case:

- **Decision Subject (DS)-**
  - **Role:** Subject of the predictive analytics.
  - **Possible Goals and Objectives:** To receive a fair, unbiased, and reasonable determination, which makes sense; to discover which factors might be changed to receive a different outcome.
  - **Technical and Domain Knowledge:** Most likely low to average technical expertise and average domain knowledge.
- **Advocate for the DS-**
  - **Role:** Support for the DS (for example, legal counsel or care worker) and concerned party to the automated decision.
  - **Possible Goals and Objectives:** To make sure the best interests of the DS are safeguarded throughout the process; to help make clear to the DS what is going on and how and why decisions are being made.
  - **Technical and Domain Knowledge:** Most likely average technical expertise and high level of domain knowledge.
- **Implementer-**
  - **Role:** User of the AI system as decision support.
  - **Possible Goals and Objectives:** To make an objective and fair decision that is sufficiently responsive to the particular circumstances of the DS and that is anchored in solid reasoning and evidence-based judgement.
  - **Technical and Domain Knowledge:** Most likely average technical expertise and high level of domain knowledge.
- **System Operator/Technician-**
  - **Role:** Provider of support and maintenance for the AI system and its use.

- **Possible Goals and Objectives:** To make sure the machine learning system is performing well and running in accordance with its intended design; to handle the technical dimension of information processing for the DS's particular case; to answer technical questions about the system and its results as they arise.
  - **Technical and Domain Knowledge:** Most likely high level of technical expertise and average domain knowledge.
- **Delivery Manager-**
  - **Role:** Member of the implementation team who oversees its operation and responds to problems as they arise.
  - **Possible Goals and Objectives:** To ensure that the quality of the automation-supported assessment process is high and that the needs of the decision subject are being served as intended by the project; to oversee the overall quality of the relationships within the implementation team and between the members of that team and the communities they serve.
  - **Technical and Domain Knowledge:** Most likely average technical expertise and good to high level domain knowledge

## Step 2: Define delivery relations and map delivery processes

- (1) Assess the possible relationships between the defined roles that will have significant bearing on your project's implementation and formulate a descriptive account of this relationship with an eye to the part it will play in the delivery process. For the predictive risk assessment example:
  - **Decision Subject/Advocate to Implementer:** This is the primary relationship of the implementation process. It should be information-driven and dialogue-driven with the implementer's exercise of unbiased judgment and the DS's comprehension of the outcome treated as the highest priorities. Implementers should be prepared to answer questions and to offer evidence-based clarifications and justifications for their determinations. The achievement of well-informed mutual understanding is a central aim.
  - **Implementer to System Operator:** This is the most critical operational relationship within the implementation team. Communication levels should be kept high from case to case, and the shared goal of the two parties should be to optimise the quality of the decisions by optimising the use of the algorithmic decision-support system in ways that are accessible both to the user and to the DS. The conversations between implementers and system operators should be problem-driven and should avoid, as much as possible, focus on the specialised vocabularies of each party's domain of expertise.
  - **Delivery Manager to Operator to Implementer:** The quality of this cross-disciplinary relationship within the implementation team will have direct bearing on the overall quality of the delivery of the algorithmically supported decisions. Safeguarding the latter will require that open and easily accessible lines of communication be maintained between delivery managers, operators, and implementers, so that unforeseen implementation problems can be tackled from multiple angles and in ways that anticipate and stem future difficulties. Additionally, different use cases may present different explanatory challenges that are best addressed by multidisciplinary

team input. Good communications within the implementation team will be essential to enable that such challenges are addressed in a timely and efficient manner.

- (2) Start building a map of the delivery process. This should involve incorporating your understanding of the needs, roles, and relationships of relevant actors involved in the implementation of your AI system into the wider objective of providing clear, informative, and understandable explanations of algorithmically supported decisions.

It is vital to recognise, at this implementation-planning stage of your project, that the principal goal of the delivery process is two-fold: *to translate statistically expressed results into humanly significant reasons and to translate algorithmic outputs into socially meaningful outcomes*.

These overlapping objectives should have a direct bearing on the way you build a map for your project's delivery process, because they organise the duties of implementation into two task-specific components:

1. A **technical component**, which involves determining the most effective way to convey and communicate to users and decision subjects the statistical results of your model's information processing so that the factors that figured into the logic and rationale of those results can be translated into understandable reasons that can be subjected to rational evaluation and critical assessment; and
2. A **social component**, which involves clarifying the socially meaningful content of the outcome of a given algorithmically assisted decision by translating that model's technical machinery—its input and output variables, parameters, and functional rationale—back into the everyday language of the humanly relevant categories and relationships that informed the formulation of its purpose, objective, and intended elements of design in the first place. Only through this re-translation will the effects of that model's output on the real human life it impacts be understandable in terms of the specific social and individual context of that life and be conveyable as such.

These two components of the delivery process will be fleshed out in turn.

**Technical component of responsible implementation:** As a general rule, we use the results of statistical analysis to guide our actions, because, when done properly, this kind of analysis offers a solid basis of empirically derived evidence that helps us to exercise sound and well-supported judgment about the matters it informs.

Having a good understanding of the factors that are at work in producing the result of a particular statistical analysis (such as in an algorithmic decision-support system) means that we are able to grasp these factors (for instance, input features that weigh heavily in determining a given algorithmically generated decision) as reasons that may warrant the rational acceptability of that result. After all, seen from the perspective of the interpretability of such an analysis, these factors are, in fact, nothing other than *reasons that are operating to support its conclusions*.

Clearly understood, these factors that lie behind the logic of the result or decision are not ‘causes’ of it. Rather, they form the evidentiary basis of its rational soundness and of the goodness of the inferences that support it. Whether or not we ultimately agree with the decision or the result of the analysis, the reasons that work together to comprise its conclusions make ***claims to validity*** and can *as such* be called before a tribunal of ***rational criticism***. These reasons, in other words, must bear the burden of continuous assessment, evaluation, and contestation.

This is an element especially crucial for the responsible implementation of AI systems: **Because they serve surrogate cognitive functions in society, their decisions and results are in no way immune from these demands for rational justification and thus must be delivered to be optimally responsive to such demands.**

The results of algorithmic decision support systems, in this sense, serve as stand-ins for acts of speech and representation and therefore bear the justificatory burdens of those cognitive functions. They must establish the validity of their conclusions and operate under the constraint of being surrogates of the dialogical goal to convince through good reasons.

This charge to be responsive to the demands of rational justification should be essential to the way you map out your delivery strategy. **When you devise how best to relay and explain the statistical results of your AI systems, you need to start from the role they play in supporting evidence-based reasoning.**

This, however, is no easy job. Interpreting the results of data scientific analysis is, more often than not, a highly technical activity and can depart widely from the conventional, everyday styles of reasoning that are familiar to most. Moreover, the various performance metrics deployed in AI systems can be confusing and, at times, seem to be at cross-purposes with each other, depending upon the metrics chosen. There is also an unavoidable dimension of uncertainty that must be accounted for and expressed in confidence intervals and error bars which may only bring further confusion to users and decision subjects.

Be that as it may, by taking a **deliberate and human-centred approach** to the delivery process, you should be able to find the most effective way to convey your model’s statistical results to users and decision subjects in non-technical and socially meaningful language that enables them to understand and evaluate the rational justifiability of those results. A good point of departure for this is to divide your map-building task into the *means of content delivery* and the *substance of the content to be delivered*.

***Means of content delivery:*** When you start mapping out serviceable ways of presenting and communicating your model’s results, you should consider **the users’ and decision subjects’ perspectives to be of primary importance**. Here are a few guiding questions to ask as you sketch out this dimension of your delivery process as well as some provisional answers to them:

- How can the delivery process of explaining the AI system’s results aid and augment the user’s and decision subject’s *mental models* (their ways of organising and filtering information), so that they can get a clear picture of the technical meaning of the

**assessment or explanation? What is the best way to frame the statistical inferences and meanings so that they can be effectively integrated into each user's own cognitive space of concepts and beliefs?**

While answering these questions will largely depend both on your use case and on the type of AI application you are building, it is just as important that you start responding to them by concentrating on the differing needs and capabilities of your explainees. To do this properly, you should first seek input from domain experts, users, and affected stakeholders, so that you can suitably scan the horizons of existing needs and capabilities. Likewise, you should take a human-centred approach to exploring the types of explanation delivery methods that would best be suited for each of your target groups. Much valuable research has been done on this in the field of human-computer interaction and in the study of human factors. This work should be consulted when mapping delivery means.

Once you have gathered enough background information, you should begin to plan out how you are going to **line up your means of delivery with the varying levels of technical literacy, expertise, and cognitive need possessed by the relevant stakeholder groups, who will be involved in the implementation of your project**. Such a **multi-tiered approach** minimally requires that individual attention be paid to the explanatory needs and capacities of implementers, system operators, and decision subjects and their advocates. This multi-tiered approach will pose different challenges at each different level.

For instance, the mental models of implementers—i.e. their ways of conceptualising the information they are receiving from the algorithmic decision-support system—may, in some cases, largely be shaped by their accumulation of domain know-how and by the filter of on-the-job expertise that they have developed over long periods of practice. These users may have a predisposition to automation distrust or aversion bias, and this should be taken into account when you are formulating appropriate means of explanation delivery.

In other contexts, the opposite may be the case. Where implementers tend to over-rely on or over-comply with automated systems, the means of explanation delivery must anticipate a different sort of mental model and adjust the presentation of information accordingly.

In any event, you will need to have a good empirical understanding of your implementer's decision-making context and maintain such knowledge through ongoing assessment. In both bias risk areas, the conveyance and communication of the assessments generated by algorithmic decision-support systems should attempt to bolster each user's practical judgment in ways that mitigate the possibility of either sort of bias. These assessments should present results as evidence-based reasons that support and better capacitate the objectivity of these implementers' reasoning processes.

The story is different with regard to the cognitive life of the technically inclined user. The mental models of system operators, who are natives in the technical vocabulary and epistemic representations of the statistical results, may be adept at the model-based problem-solving tasks that arise during implementation but less familiar with identifying and responding to the cognitive needs and limitations of non-technical stakeholders. Incorporating ongoing communication exercises and training into their roles in the delivery process may capacitate them to better facilitate implementers' and decision subjects' understanding of the technical details of the assessments generated by algorithmic decision-support systems. These ongoing development activities will not only helpfully enrich operators' mental models, they may also inspire them to develop deeper, more responsive, and more effective ways of communicating the technical yields of the analytics they oversee.

Finally, the mental models of decision subjects and their advocates will show the broadest range of conceptualisation capacities, so your delivery strategy should (1) prioritise the facilitation of optimal explanation at the baseline level of the needs of the most disadvantaged of them and (2) build the depth of your multi-tiered approach to providing effective explanations into the delivery options presented to decision subjects and their advocates. This latter suggestion entails that, beyond provision of the baseline explanation of the algorithmically generated result, options should be given to decision subjects and their advocates to view more detailed and technical presentations of the sort available to implementers and operators (with the proviso that reasonable limitations be placed on transparency in accordance with the need to protect the confidential personal and organisational information and to prevent gaming of the system).

- **How can non-technical stakeholders be adequately prepared to gain baseline knowledge of the kinds of statistical and probabilistic reasoning that have factored into the technical interpretation of the system's output, so that they are able to comprehend it on its own technical terms? How can the technical components be presented in a way that will enable explainees to easily translate the statistical inferences and meanings of the results into understandable and rationally assessable terms? What are the best available media for presenting the technical results in engaging and comprehensible ways?**

To meet these challenges, you should consider supplementing your implementation platform with knowledge-building and enrichment resources that will provide non-technical stakeholders with access to basic technical concepts and vocabulary. At a minimum, you should consider building a plain language glossary of basic terms and concepts that will include all of the technical ideas covered by the algorithmic component of a given explanation. If your explanation platforms are digital, you should also make them as user friendly as possible by hyperlinking the technical terms used in the explanations to their plain language glossary elaborations.

Where possible, explanatory demonstrations of technical concepts (like performance metrics, formal fairness criteria, confidence intervals, etc.) should be provided to users and decision subjects in an engaging and easy-to-comprehend way, and

graphical and visualisation techniques should be consistently used to make potentially difficult ideas more accessible. Moreover, the explanation interfaces themselves should be as simple, learnable, and usable as possible. They should be tested to measure the ease with which those with neither technical experience nor domain knowledge are able to gain proficiency in their use and in understanding their content.

**Substance of the technical content to be delivered:** The overall interpretability of your AI system will largely hinge on the effectiveness and even-handedness of your technical content delivery. You will have to strike a balance between (1) determining how best to convey and communicate the rationale of the statistical results so that they may be treated appropriately as decision supporting and clarifying reasons and (2) being clear about the limitations of and potential uncertainties in the statistical results themselves so that the explanations you offer will not mislead implementers and decision subjects. These are not easy tasks and will require substantial forethought as you map out the content clarification aspect of your delivery process.

To assist you in this, here is a non-exhaustive list of recommendations that you should consider as you map out the execution of the technical content delivery component of the responsible implementation of your AI project (This list will, for the sake of specificity, assume the predictive risk assessment example):

- Each explanation should be presented in plain, non-technical language and in an optimally understandable way so that the results provided can enable the affordance of better judgment on the part of implementers and optimal understanding on the part of decision subjects. On the implementer's side, the primary goal of the explanation should be to support the user's ability to offer solid, coherent, and reasonable justifications of their determinations of decision outcomes. On the decision subject's side, the primary goal of the explanation should be to make maximally comprehensible the rationale behind the algorithmic component of the decision process, so that the decision subject can undertake a properly informed critical evaluation of the decision outcome as a whole.
- Each explanation should present its results as facts or evidence in as sparse but complete and sound a manner as possible with a clear indication of what components in the explanation are operating as premises, what components are operating as conclusions, and what the inferential rationale is that is connecting the premises to the conclusions. Each explanation should therefore make explicit the rational criteria for its determination whether this be, for example, global inferences drawn from the population-based reasoning of a demographic analysis or more locally or instance-based inferences drawn from the indication of feature significance by a proxy model. In all cases, the optimisation criteria of the operative algorithmic system should be specified, made explicit, and connected to the logic and rationale of the decision.
- Each explanation should make available the records and activity-monitoring results that the design and development processes of your AI project yielded. Building this link between the process transparency dimension of your project and its outcome transparency will help to make its result, as a whole, more sufficiently interpretable. This

can be done by simply linking or including the public-facing component of the process log of your PBG Framework.

- Each explanation provided to an implementer should come with a standard **Implementation Disclaimer** that may read as follows:

**Implementation Disclaimer:**

These results are intended to assist you in making an evidence-based judgment. They are meant neither to replace your reasoned deliberations nor to constitute the sole evidentiary basis of your judgement. These results are also derived from statistical analysis. This means (1) that there are unavoidable possibilities of error and uncertainty in their results, which are specified in the performance measures and confidence intervals provided and (2) that these results are based on population-level data that do not refer specifically to the actual circumstances and abilities of the individual subject of their prediction. The inferences you draw directly from them will therefore be based on statistical generalisation not on an understanding of the life context or concrete potential of the individual person, who will be impacted by your decision.

- Each explanation should specify and make explicit its governing performance metrics together with the acceptability criteria used to select those metrics and any standard benchmarks followed in establishing that criteria. Where appropriate and possible, fuller information about model validation measurement (including confusion matrix and ROC curve results) and any external validation results should be made available.
- Each explanation should provide confirmatory information that the formal fairness criteria specified in your project's Fairness Policy Statement has been met.
- Each explanation should include clear representations of confidence intervals and error bars. These certainty estimates should make as quantitatively explicit as possible the confidence range of specific predictions, so that users and decision subjects can more fully understand their reliability and the levels of uncertainty surrounding them.
- When an explanation offers categorically ordered scores (for instance, risk scores on a scale of 1 to 10), that explanation must also explicitly indicate the actual raw numerical probabilities for the labels (predicted outcomes) that have been placed into those categories. This will help your delivery process avoid producing confusion about the relative magnitudes of the categorical groupings under which the various scores fall. Information should also be provided about the relative distances between the risk scores of specific cases if the risk categories under which they are placed are unevenly distributed. It may be possible, for example, for two cases, which fall under the same high risk category (say, 9) to be farther apart in terms of the actual values of their risk probabilities than two other cases in two different categories (say 1 and 4). This may be misleading to the user.

- Each explanation should, where possible, include a counterfactual explanatory tool, so that implementers and affected individuals have the opportunity to gain a better contrastive understanding of the logic of the outcome and its alternative possibilities.

**Social component of responsible implementation:** We have now established the first step in the delivery of a responsible implementation process: making clear the rationale behind the technical content of an algorithmic model’s statistical results and determining how best to convey and communicate it so that these results may be appropriately treated as decision supporting and clarifying reasons. This leaves us with a second related task of content clarification, which is only implicit in the first step but must be made explicit and treated reflectively in a second.

Beyond translating statistically expressed results into humanly significant reasons, you will have to make sure that their ***socially meaningful content*** is clarified by implementers, so that they are able to thoughtfully apply these results to the real human lives they impact in terms of the specific societal and individual contexts in which those lives are situated.

This will involve explicitly translating that model’s technical machinery—its input and output variables, parameters, and functional rationale—*back* into the everyday language of the humanly relevant meanings, categories, and relationships that informed the formulation of its purpose, objectives, and intended elements of design in the first place. It will also involve training and preparing implementers to intentionally assist in carrying out this translation in each particular case, so that due regard for the dignity of decision subjects can be supported by the interpretive charity, reasonableness, empathy, and context-specificity of the determination of the outcomes that affect them.

Only through this re-translation will the internals, mechanisms, and output of the model become ***useably interpretable*** by implementers: Only then will they be able to apply input features of relevance to the specific situations and attributes of decision subjects. Only then will they be able to critically assess the manner of inference-making that led to its conclusion. And only then will they be able to adequately weigh the normative considerations (such as prioritising public interest or safeguarding individual well-being) that factored into the system’s original objectives.

Having clarified the socially meaningful content of the model’s results, the implementer will be able to more readily apply its evidentiary contribution to a more holistic and wide-angled consideration of the particular circumstances of the decision subject while, at the same time, weighing these circumstances against the greater purpose of the algorithmically assisted assessment. It is important to note here that the understanding enabled by the clarification of the social context and stakes of an algorithmically supported decision-making process goes hand-in-glove with fuller considerations of the moral justifiability of the outcome of that process.

A good starting point for considering how to integrate this clarification of the socially meaningful content of an algorithmic model’s output into your map of the delivery process is to consider what you might think of as your AI project’s **content lifecycle**.

**The content lifecycle:** The output of an algorithmic system does not begin and end with the computation. Rather, it begins with the very human purposes, ideas, and initiatives that lay behind the conceptualisation and design of that system. Creating technology is a shared public activity, and it is animated by human objectives and beliefs. An algorithmic system is brought into the world as the result of this collective enterprise of ingenuity, intention, action, and collaboration.

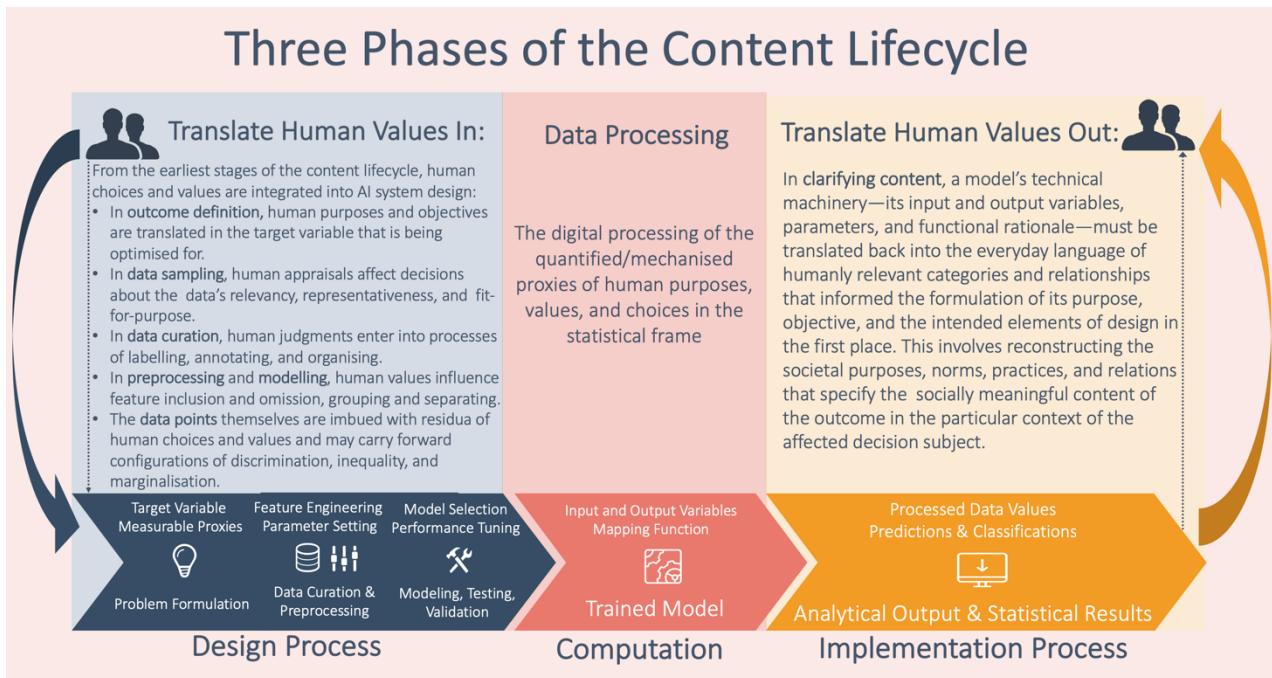
Human choices and values therefore punctuate the design and implementation of AI systems. These choices and values are inscribed in algorithmic models:

- At the very inception of an AI project, human choices and values come into play when we formulate the goals and objectives to be achieved by our algorithmic technologies. They come into play when we define the optimal outcome of our use of such technologies and when we translate these goals and objectives into target variables and their measurable proxies.
- Human choices and values come into play when decisions are made about the sufficiency, fit-for-purpose, representativeness, relevance, and appropriateness of the data sampled. They come into play in how we curate our data—in how we label, organise, and annotate them.
- Such choices and values operate as well when we make decisions about how we craft a feature space—how we select or omit and aggregate or segregate attributes. Determinations of what is relevant, reasonable, desirable, or undesirable will factor into what kinds of inputs we are going to include in the processing and how we are going to group and separate them.
- Moreover, the data points themselves are imbued with residua of human choices and values. They carry forward historical patterns of social and cultural activity that may contain configurations of discrimination, inequality, and marginalisation—configurations that must be thoughtfully and reflectively considered by implementers as they incorporate the analytics into their reasoned determinations.

Whereas all of these human choices and values are translated into the algorithmic systems we build, the responsible implementation of these systems requires that they be translated out. The rationale and logic behind an algorithmic model's output can be properly understood as it affects the real existence of a decision subject only when we transform its variables, parameters, and analytical structures back into the human currency of values, choices, and norms that shaped the construction of its purpose, its intended design, and its optimisation logic from the start.

It is only in virtue of this **re-translation** that an algorithmically supported outcome can afford stakeholders the degree of deliberation, dialogue, assessment, and mutual understanding that is necessary to make it fully comprehensible and justifiable to them. And, it is likewise only in virtue of this re-translation that the implementation process itself can, at once, secure end-to-end accountability and give due regard to the SUM values.

The content lifecycle of algorithmic systems therefore has three phases: (1) The **translation in** of human purposes, values, and choices during the design process; (2) The digital processing of the quantified/mechanised proxies of these purposes, values, and choices in the statistical frame; (3) The **translation out** of the purposes, values, and choices in clarifying the socially meaningful content of the result as it affects the life of the decision subject through the implementation process. Here is a visualisation of these three phases of the content lifecycle:



**The translation rule:** A beneficial result of framing the implementation process in terms of the content lifecycle is that it gives us a clear and context-sensitive measure by which to identify the explanatory needs of any given AI application. We can think of this measurement as the **translation rule**. It states that:

What is *translated in* to an algorithmic system with regard to the human choices and societal values that determine its content and purpose is directly proportional to what, in terms of the explanatory needs of clarification and justification, must be *translated out*.

The translation rule organically makes two distinctions that have great bearing on the delivery process for responsible implementation. First, it divides the question of what needs explaining into two parts: (1) issues of socially meaningful content in need of clarification (i.e., the explanatory need that comes from the **translation in** to the AI model of the categories, meanings, and relations that originate in social practices, beliefs, and intentions) (2) issues of normative rightness in need of justification (i.e. the explanatory need that comes from **translation in** to the AI model of choices and considerations that have bearing on its ethical permissibility, discriminatory non-harm, and public trustworthiness). These two parts line up with what we have [above](#) called **interpretable AI** and **justifiable AI** respectively, and what we have also identified as [tasks 2 and 3](#) of delivering transparent AI.

Secondly, the translation rule divides the two dimensions of translation (translation in and translation out) into aspects of **intention-in-design** and **intention-in-application**. *Translating in*

has to do with *intention-in-design*. It involves an active awareness of the human purposes, objectives, and intentions that factor into the construction of AI systems. *Translating out*, on the other hand, has directly to do with *intention-in-application*, or put differently, the intentional dimension of the implementation of an AI system by a user in a specific context and with direct consequences for a subject affected by its outcome.

In human beings, intention-in-design and intention-in-application are *united in intelligent action*, and it is precisely this unity that enables people to reciprocally hold each other accountable for the consequences of what they say and what they do. By contrast, in artificial intelligence systems, which fulfil surrogate cognitive functions in society but are themselves neither intentional nor accountable, design and application are divided. In these systems, intention-in-design and intention-in-application are and must remain *punctuation points of human involvement and responsibility* that manifest on either side of the vacant mechanisms of data processing. This is why translation is so important, and this is why enabling the implementer's capacity to *intentionally translate out the social and normative content* of the model's results is such a critical element of the responsible delivery of your AI project.

It might be helpful to think more concretely about the translation rule by considering it in action. Let's compare two hypothetical examples: (1) a use case about an early cancer detection system in radiomics (a machine learning application that uses high throughput computing to identify features of pathology that are undetectable to the trained radiological eye); and (2) a use case about a predictive risk assessment application that supports decision-making in child social care.

In the radiomics case, the *translating in* dimension involves minimal social content: the clinical goal inscribed in the model's objective is that of lesion detection and the features of relevance are largely voxels extracted from PET and CT scanner images. However, the normative aspect of *translating in* is, in this case, significant. Ethical considerations about looking after patient wellbeing and clinical safety are paramount and wider justice concerns about improving healthcare for all and health equity factor in as well.

The explanatory needs of the physician/implementer receiving clinical decision support and of the clinical decision subject will thus lean less heavily on the dimension of the clarification of socially meaningful content than it will on the normative dimension of justifying the safety of the system, the priority of the patient's wellbeing, and the issues of improved delivery and equitable access. The technical content of the decision support may be crucial here (Issues surrounding the reproducibility of the results and the robustness of the system may, in fact, be of great concern in the assessment of the validity of the outcome.), but the *translating out* component of the implementation remains directly proportional to the minimal social content and to the substantial ethical concerns and objectives that were *translated in* and that thus inform the explanatory and justificatory needs of the result in general.

The explanatory demands in the child social care risk assessment use case are entirely different. The social content of the *translating in* dimension is intricate, multi-layered, and extensive. The chosen target variable may be child safety or the prevention of severe mistreatment and the measurable proxy, home removal of at-risk children within a certain timeframe. Selected features that are deemed relevant may include the age of the at-risk

children, public health records, previous referrals, family history of violent crime, welfare records, juvenile criminal records, demographic information, and mental health records. Complex socioeconomic and cultural formations may additionally influence the representativeness and quality of the dataset as well as the substance of the data itself.

The normative aspect of *translating in* here is also subtle and complicated. Ethical considerations about protecting the welfare of children at risk are combined with concerns that parents and guardians be treated fairly and without discrimination. Objectives of providing evidence-based decision support are also driven by hopes that accurate results and well-reasoned determinations will preserve the integrity and sanctity of familial relations where just, safe, and appropriate. Other goals and purposes may be at play as well such as making an overburdened system of service provision more efficient or accelerating real-time decision-making without harming the quality of the decisions themselves.

In this case of predictive risk assessment, the *translating out* burdens of the frontline social worker are immense both in terms of clarifying content and in terms of moral justification. If, for example, analytical results yielding a high risk score were based on the relative feature importance of demographic information, welfare records, mental health records, and criminal history, the implementer would have to scrutinise the particular decision subject's situation, so that the socially meaningful content of these factors could be clarified in terms of the living context, relevant relationships, and behavioural patterns of the stakeholders directly affected. Only then could the features of relevance be thoroughly and deliberatively assessed.

The effective interpretability of the model's result would, in this case, heavily depend on the implementer's ability to apply domain-knowledge in order to reconstruct the meaningful social formations, intentions, and relationships that constituted the concrete form of life in which the predictive risk modelling applies. The implementer's well-reasoned decision here would involve a careful weighing of this socially clarified content against the wider predictive patterns in the data distribution yielded by the model's results—patterns that may have otherwise gone unnoticed.

Such a weighing process would, in turn, be informed by the normative-explanatory need to *translate out* the morally implicating choices, concerns, and objectives that influenced and informed the predictive risk assessment model's development in the first place. Again, the interpretive burden of the frontline social worker would be immense here. First, this implementer would have to deliberate with a critically informed awareness of the legacies of discrimination and inequity that tend to feed forward in the kinds of evidentiary sources drawn upon by the analytics. Such an active reflexivity is crucial for retaining the punctuating role of human involvement and responsibility in these sensitive and high-stakes environments.

Just as importantly, the frontline social worker would have to evaluate the real impact of ethical objectives at the point of delivery. Not only would the results of the analytics have to be aligned with the ethical concerns and purposes that fostered the construction of the model, this implementer would have to reflectively align their own potentially diverging ethical point of view both with those results and with those objectives. This *normative*

***triangulation*** between the original intention-in-design, the implementer's intention-in-application, and the content clarification of the AI system's results is, in fact, a crucial safeguard to the delivery of justifiable AI. It again enables a reanimation of moral involvement and responsibility at the most critical juncture of the content lifecycle.

### Step 3: Build an ethical implementation platform:

- (1) **Train ethical implementation.** The continuous challenges of translation, content clarification, and normative explanation should inform how you set up your implementation training to achieve optimal outcome transparency. In addition to the necessary [training to prevent implementation biases in the users of your AI system](#) (discussed above), you should prepare and train the implementers to be stewards of interpretable and justifiable AI. This entails that they be able to:
  - Rationally evaluate and critically assess the logic and rationale behind the outputs of the AI systems;
  - Convey and communicate their algorithmically assisted decisions to the individuals affected by them in plain language. This includes explaining to them in an everyday, non-technical, and accessible way how and why the decision-supporting model performed the way it did in a specific context and how that result factored into the final outcome of the implementation;
  - Apply the conclusions reached by the AI model to a more focused consideration of the particular social circumstances and life context of the decision subject and other affected parties;
  - Treat the inferences drawn from the results of the model's computation as evidentiary contributions to a broader, more rounded, and coherent understanding of the individual situations of the decision subject and other affected parties;
  - Weigh the interpretive understanding gained by integrating the model's insights into this rounded picture of the life context of the decision subject against the greater purpose and societal objective of the algorithmically assisted assessment;
  - Justify the ethical permissibility, the discriminatory non-harm, and the public trustworthiness both of the AI system's outcome and of the processes behind its design and use
- (2) **Make your implementation platform a relevant part and capstone of the sustainability track of your project.** An important element of gauging the impacts of your AI technology on the individuals and communities it touches is having access to the frontlines of its potentially transformative and long-term effects. Your implementation platform should assist you in gaining this access by being a *two-way medium of application and communication*. It should both enable you to sustainably achieve the objectives and goals you set for your project through responsible implementation, but it should also be a sounding board as well as a site for feedback and cooperative sense-checking about the real-life effects of your system's use.

Your implementation platform should be dialogically and collaboratively connected to the stakeholders it effects. It should be bound to the communities it serves as part of a shared project to advance their immediate and long-run wellbeing.

- (3) **Provide a model sheet to implementers and establish protocols for implementation reporting.** As part of the roll-out of your AI project, you should prepare a summary/model sheet for implementers, which includes summation information about the system's technical specifications and all of the relevant details indicated above in the section on [substance of the technical content to be delivered](#). This should include relevant information about performance metrics, formal fairness criteria and validation, the implementation disclaimer, links or summaries to the relevant information from the process logs of your PBG Framework, and links or summary information from the Stakeholder Impact Assessment.

You should also set up protocols for implementation reporting that are proportional to the potential impacts and risks of the system's use.

- (4) **Foster outcome understanding through dialogue.** Perhaps the single most important aspect of building a platform for ethical implementation is the awareness that the realisation of interpretable and justifiable AI is a dialogical and collaborative effort. Because all types of explanation are mediated by language, each and every explanatory effort is a participatory enterprise where understanding can be reached only through acts of communication. The interpretability and justifiability of AI systems depend on this shared human capacity to give and ask for reasons in the ends of reaching mutual understanding. Implementers and decision subjects are, in this respect, first and foremost participants in an explanatory dialogue, and the success of their exchange will hinge both on a reciprocal readiness take the other's perspective and on a willingness to enlarge their respective mental models in accordance with new, communicatively achieved, insights and understandings.

For these reasons, your implementation platform should encourage open, mutually respectful, sincere, and well-informed dialogue. Reasons from all affected voices must be heard and considered as demands for explanation arise, and manners of response and expression should remain clear, straightforward, and optimally accessible. Deliberations that have been inclusive, unfettered, and impartial tend to generate new ideas and insights as well as better and more inferentially sound conclusions, so approaching the interpretability and justifiability of your AI project in this manner will not only advance its responsible implementation, it will likely encourage further improvements in its design, delivery, and performance.

## Conclusion:

In 1936, a 23-year-old mathematician from Maida Vale named Alan Turing sat down with pencil and paper. Using just the image of a linear tape divided evenly into squares, a list of symbols, and a few basic rules, he drew a sketch to show the step-by-step process of how a human being can carry out any calculation, from the simplest operation of arithmetic to the most complex nonlinear differential equation.

Turing's remarkable invention (now known simply as the Turing machine) solved the perplexing and age-old mathematical question of *what an effective calculation is*—the question of *how to define an algorithm*. Not only did Turing show what it means to compute a number by showing *how humans do it*, he created, in the process, the idea behind the modern general purpose computer. Turing's astonishingly humble innovation ushered in the digital age.

Just over eight decades later, as we step forward together into the open horizons of a rapidly evolving digital future, it is difficult to image that what started as a thought experiment in a small room at Kings College, Cambridge has now become such a humanly defining force. We live in an increasingly dynamic and integrated computational reality where connected devices containing countless sensors and actuators intermingle with omnipresent algorithmic systems and cloud computing platforms.

With the rise of the Internet of Things, edge computing, and the expanding smart automation of infrastructure, industry, and the workplace, AI systems are progressively more coming to comprise the cyber-physical frame and fabric of our networked society. For better or worse, artificial intelligence is not simply becoming a general purpose technology (like steam power or electricity). It is, more essentially, becoming a gatekeeper technology that uniquely holds the key both to the potential for the exponential advancement of human wellbeing and to possibilities for the emergence of significant risks for society's future. It is, as yet, humankind that must ultimately choose which direction the key will turn.

This choice leaves difficult questions in the lap of the moral agency of the present: What shape will the data-driven society of tomorrow take? How will the values and motivations that are currently driving the gathering energies of technological advancement in artificial intelligence come both to influence that future society's ways of life and to transform the identities of its warm-blooded subjects?

This guide on understanding AI ethics and safety has offered you one way to move forward in answering these questions. In a significant sense, it has attempted to prepare you to take Turing's lead: to see the design and implementation of algorithmic models as an eminently *human activity*—an activity guided by our purposes and values, an activity for which, each of us, who is involved in the development and deployment of AI systems, is morally and socially responsible.

This starting point in human action and intention is a crucial underpinning of responsible innovation. For, it is only when we prioritise considerations of the ethical purposes and values behind the trajectories of our technological advancement, that we, as vested societal stakeholders, will be able to take the reins of innovation and to steer the course of our algorithmic creations in accordance with a shared vision of what a better human future should look like.

## Acknowledgments:

Writing this guide would simply not have been possible without the hard work, dedication, and insight of so many interlocutors both within The Alan Turing Institute and through the meaningful partnerships that the Turing’s Public Policy Programme has formed with stakeholders from across the UK Government.

To take the latter group first, the Office for Artificial Intelligence (OAI) and the Government Digital Service (GDS)’s keen vision and their commitment to responsible AI innovation have been an enabling condition of the development of this work. In particular, the patience and incisiveness of OAI’s Sébastien Krier and Jacob Beswick, and GDS’ Bethan Charnley have been instrumental in bringing the project to its completion.

I am also incredibly grateful for the impact that our interactions with the Ministry of Justice (MoJ)’s Data Science Hub has had on developing the framing for this guide. Input from the MoJ’s Megan Whewell, Philip Howard, Jonathan Roberts, Olivia Lewis, Ross Wyatt, and from its Data Science Innovation Board have left a significant mark on the research.

Last, but not least, our ongoing partnership with the Information Commissioner’s Office on Project ExplA/n—and, in particular, with ICO colleagues Carl Wiper and Alex Hubbard—has been a key contributor to this guide’s focus on fairness, transparency, and accountability. Project ExplA/n aims to provide practical guidance for organisations on explaining AI supported decisions to the subjects of those decisions. Taking inspiration from our work on Project ExplA/n and from the input gathered over the course of the two citizens’ juries we held in Manchester and Coventry, the current guide emphasises the importance of communication and attempts to build out a vision of human-centred and context-sensitive implementation.

As the Ethics Fellow within the Public Policy Programme at the Turing, I have benefited tremendously from being surrounded by an immensely talented group of thinkers and doers, whose commitment to making the connected world a better place through interdisciplinary research and advisory intervention is an inspiration every day. Programme Director, Helen Margetts, and Deputy Director, Cosmina Dorobantu, have been crucial and inimitable supports of this project from its inception as has my small but brilliant team of researchers, Josh Cowls and Christina Hitrova. My involvement with the Turing’s Data Ethics Group has also been a tremendous source of insight and inspiration for this project. Given the ambitious deadlines that accompanied this guide’s final stages of production, heroic efforts to review its contents as a whole or in parts were made by Florian Ostmann, Michael Veale, David Watson, Mark Briers, Evelina Gabsova, Alexander Harris, and Anna FitzMaurice. Their perceptive feedback notwithstanding, any unclarities that appear in *Understanding Artificial Intelligence Ethics and Safety* reflect the faults of its author alone.

## Bibliography and Further Readings

Included here is a bibliography organised into the main themes covered in this guide. Please use this as a starting point for further exploration of these complex topics. Many thanks to the tireless efforts of Jess Morley and Corianna Moffatt without whom this bibliography could not have been compiled.

[The SUM Values](#)

[General fairness](#)

[Data fairness](#)

[Design fairness](#)

[Outcome fairness](#)

[Implementation fairness](#)

[Accountability](#)

[Stakeholder Impact Assessment](#)

[Safety: Accuracy, reliability, security, and robustness](#)

[Transparency](#)

[Process-Based Governance](#)

[Interpretable AI](#)

[Responsible delivery through human-centred implementation protocols and practices](#)

[Individual and societal impacts of machine learning and algorithmic systems](#)

### The SUM Values

Access Now. (2018). *The Toronto declaration: Protecting the rights to equality and non-discrimination in machine learning systems*. Retrieved from [https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration\\_ENG\\_08-2018.pdf](https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf)

Adamson, G., Havens, J. C., & Chatila, R. (2019). Designing a value-driven future for ethical autonomous and intelligent systems. Proceedings of the IEEE, 107(3), 518–525. <https://doi.org/10.1109/JPROC.2018.2884923>

American Medical Association. (2001). AMA code of medical ethics. Retrieved from <https://www.ama-assn.org/sites/ama-assn.org/files/corp/media-browser/principles-of-medical-ethics.pdf>

American Psychological Association. (2016). *Ethical principles of psychologists and code of conduct*. Retrieved from <https://www.apa.org/ethics/code/>

Article 19. (2019). *Governance with teeth: How human rights can strengthen FAT and ethics initiatives on artificial intelligence*. Retrieved from <https://www.article19.org/resources/governance-with-teeth-how-human-rights-can-strengthen-fat-and-ethics-initiatives-on-artificial-intelligence/>

Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics*. 6th edition. Oxford University Press, USA.

- Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), 20180080. <https://doi.org/10.1098/rsta.2018.0080>
- Cowls, J., & Floridi, L. (2018). Prolegomena to a White Paper on an Ethical Framework for a Good AI Society. <http://dx.doi.org/10.2139/ssrn.3198732>
- European Commission. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- European Group on Ethics in Science and New Technologies. (2018). *Artificial intelligence, robotics, and 'autonomous' systems*. Retrieved from [https://ec.europa.eu/research/ege/pdf/ege\\_ai\\_statement\\_2018.pdf](https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf)
- Felten, E. (2016). Preparing for the future of artificial intelligence. *Washington DC: The White House*.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707. Retrieved from <https://link.springer.com/article/10.1007/s11023-018-9482-5>
- Floridi, L., & Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160360. <https://doi.org/10.1098/rsta.2016.0360>
- Future of Life Institute. (2017). *Asilomar AI principles*. Retrieved from <https://futureoflife.org/ai-principles/>
- Global Future Council on Human Rights 2016-2018. (2018). How to prevent discriminatory outcomes in machine learning. *World Economic Forum*. Retrieved from [http://www3.weforum.org/docs/WEF\\_40065\\_White\\_Paper\\_How\\_to\\_Prevent\\_Discriminatory\\_Outcomes\\_in\\_Machine\\_Learning.pdf](http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf)
- House of Lords Select Committee on Artificial Intelligence. (2018). *AI in the UK: Ready, willing and able?*. Retrieved from <https://publications.parliament.uk/pa/ld201719/ldselect/l dai/100/100.pdf>
- IEEE. (2018). *The IEEE Global Initiative on ethics of autonomous and intelligent systems*. Retrieved from [https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead\\_v2.pdf](https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf)
- Latonero, M. (2018). Governing artificial intelligence: Upholding human rights & dignity. *Data & Society*. Retrieved from [https://datasociety.net/wp-content/uploads/2018/10/DataSociety\\_Governing\\_Artificial\\_Intelligence\\_Upholding\\_Human\\_Rights.pdf](https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf)
- The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, D.C.: United States Government Printing Office.
- Nuffield Council on Bioethics. (2015). *The collection, linking and use of data in biomedical research and health care: ethical issues*. Retrieved from <http://nuffieldbioethics.org/wp-content/uploads/Biodata-a-guide-to-the-report-PDF.pdf>
- Nuffield Council on Bioethics. (2018). *Artificial intelligence (AI) in healthcare and research*. Retrieved from <http://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf>
- Pielemeier, J. (2018). The advantages and limitations of applying the international human rights framework to artificial intelligence. *Data & Society: Points*. Retrieved from <https://points.datasociety.net/the-advantages-and-limitations-of-applying-the-international-human-rights-frame-work-to-artificial-291a2dfe1d8a>
- Ramesh, S. (2017). A checklist to protect human rights in artificial-intelligence research. *Nature*, 552(7685), 334–334. <https://doi.org/10.1038/d41586-017-08875-1>
- Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). Artificial intelligence & human rights: Opportunities & risks. *Berkman Klein Center Research Publication*, (2018-6). Retrieved from [https://cyber.harvard.edu/sites/default/files/2018-09/2018-09\\_AIHumanRightsSmall.pdf](https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf)
- Reform. (2018). *Thinking on its own: AI in the NHS*. Retrieved from [https://reform.uk/sites/default/files/2018-11/AI%20in%20Healthcare%20report\\_WEB.pdf](https://reform.uk/sites/default/files/2018-11/AI%20in%20Healthcare%20report_WEB.pdf)

- Royal Society. (2017). *Machine learning: The power and promise of computers that learn by example*. Retrieved from <https://royalsociety.org/~media/policy/projects/machine-learning/publications/machine-learning-report.pdf>
- Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, 361(6404), 751–752. <https://doi.org/10.1126/science.aat5991>
- UK Statistics Authority. (2017). *Code of practice for statistics: Ensuring public confidence in statistics*. Retrieved from <https://www.statisticsauthority.gov.uk/wp-content/uploads/2017/07/DRAFT-Code-2.pdf>
- UNESCO. (2017). *Report of COMEST on robotics ethics*. Retrieved from <http://unesdoc.unesco.org/images/0025/002539/253952E.pdf>
- Université de Montréal. (2017). *Montreal declaration for responsible AI*. Retrieved from <https://www.montrealdeclaration-responsibleai.com/the-declaration>
- US Department of Homeland Security. (2012). *The Menlo report: Ethical principles guiding information and communication technology research*. Retrieved from [https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803\\_1.pdf](https://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803_1.pdf)
- US National Science and Technology Council. (2016). *Preparing for the future of artificial intelligence*. Retrieved from [https://obamawhitehouse.archives.gov/sites/default/files/whitehouse\\_files/microsites/ostp/NSTC/preparing\\_for\\_the\\_future\\_of\\_ai.pdf](https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf)
- Villani, C. (2018). For a meaningful artificial intelligence: Towards a French and European strategy. *AI For Humanity*. Retrieved from [https://www.aiforhumanity.fr/pdfs/MissionVillani\\_Report\\_ENG-VF.pdf](https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf).
- Yuste, R., Goering, S., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., ... & Kellmeyer, P. (2017). Four ethical priorities for neurotechnologies and AI. *Nature News*, 551(7679), 159. Retrieved from <https://www.nature.com/news/four-ethical-priorities-for-neurotechnologies-and-ai-1.22960>

## General fairness

- Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. *arXiv:1712.03586*. Retrieved from <https://arxiv.org/abs/1712.03586>
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 377). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3173951>
- Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M., & Wallach, H. (2018). Improving fairness in machine learning systems: What do industry practitioners need?. *ArXiv:1812.05239*. <https://doi.org/10.1145/3290605.3300830>
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31(4), 611–627. <https://doi.org/10.1007/s13347-017-0279-x>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 205395171667967. <https://doi.org/10.1177/2053951716679679>
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59-68). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287598>
- Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv:1901.10002*. Retrieved from <https://arxiv.org/abs/1901.10002>
- Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 440). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3174014>

## Data fairness

- Abadi, D., Agrawal, R., Ailamaki, A., Balazinska, M., Bernstein, P. A., Carey, M. J., ... & Gehrke, J. (2016). The Beckman report on database research. *Communications of the ACM*, 59(2), 92-99. Retrieved from <https://dl.acm.org/citation.cfm?id=2845915>
- Abiteboul, S., & Stoyanovich, J., & Weikum, G. (2015). Data, Responsibly. *ACM Sigmod Blog*. Retrieved from <http://wp.sigmod.org/?p=1900>
- Alper, P., Becker, R., Satagopam, V., Grouès, V., Lebioda, J., Jarosz, Y., ... & Schneider, R. (2018). *Provenance-enabled stewardship of human data in the GDPR era*. <https://doi.org/10.7490/f1000research.1115768.1>
- Ambacher, B., Ashley, K., Berry, J., Brooks, C., Dale, R. L., & Flecker, D. (2007). Trustworthy repositories audit & certification: Criteria and checklist. *Center for Research Libraries, Chicago/Illinois*. Retrieved from [https://www.crl.edu/sites/default/files/d6/attachments/pages/trac\\_0.pdf](https://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf)
- Antignac, T., Sands, D., & Schneider, G. (2016). Data minimisation: A language-based approach (long version). *ArXiv:1611.05642*. Retrieved from <http://arxiv.org/abs/1611.05642>
- Bell, D., L'Hours, H., Lungley, D., Cunningham, & N., Corti, L. (n.d.). Scaling up: digital data services for the social sciences. *UK Data Service*. Retrieved from <https://www.ukdataservice.ac.uk/media/604995/ukds-case-studies-scaling-up.pdf>
- Bower, A., Niss, L., Sun, Y., & Vargo, A. (2018). Debiasing representations by removing unwanted variation due to protected attributes. *arXiv:1807.00461*. Retrieved from <https://arxiv.org/abs/1807.00461>
- Custers, B. (2013). Data dilemmas in the information society: Introduction and overview. In *Discrimination and Privacy in the Information Society* (pp. 3-26). Springer, Berlin, Heidelberg. Retrieved from [https://link.springer.com/chapter/10.1007/978-3-642-30487-3\\_1](https://link.springer.com/chapter/10.1007/978-3-642-30487-3_1)
- Custers, B. H., & Schermer, B. W. (2014). Responsibly innovating data mining and profiling tools: A new approach to discrimination sensitive and privacy sensitive attributes. In *Responsible Innovation 1* (pp. 335-350). Springer, Dordrecht. Retrieved from [https://link.springer.com/chapter/10.1007/978-94-017-8956-1\\_19](https://link.springer.com/chapter/10.1007/978-94-017-8956-1_19)
- Dai, W., Yoshigoe, K., & Parsley, W. (2018). Improving data quality through deep learning and statistical models. *ArXiv:1810.07132*, 558, 515–522. [https://doi.org/10.1007/978-3-319-54978-1\\_66](https://doi.org/10.1007/978-3-319-54978-1_66)
- Davidson, S. B., & Freire, J. (2008). Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 1345-1350). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=1376772>
- European Commission Expert Group on FAIR Data. (2018). Turning FAIR into reality. *European Union*. Retrieved from [https://ec.europa.eu/info/sites/info/files/turning\\_fair\\_into\\_reality\\_1.pdf](https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf)
- Faundeen, J. (2017). Developing criteria to establish trusted digital repositories. *Data Science Journal*, 16. Retrieved from <https://datascience.codata.org/article/10.5334/dsj-2017-022/>
- Joshi, C., Kaloskampis, I., & Nolan, L. (2019). Generative adversarial networks (GANs) for synthetic dataset generation with binary classes. *Data Science Campus*. Retrieved from <https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset-generation-with-binary-classes/>
- L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with Big Data: Challenges and approaches. *IEEE Access*, 5, 7776-7797. Retrieved from <https://ieeexplore.ieee.org/abstract/document/7906512/>
- Ruggieri, S., Pedreschi, D., & Turini, F. (2010). DCUBE: Discrimination discovery in databases. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 1127-1130). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=1807298>
- Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. *LREC*, 859–866.
- Stoyanovich, J., Howe, B., Abiteboul, S., Miklau, G., Sahuguet, A., & Weikum, G. (2017). Fides: Towards a platform for responsible data science. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (p. 26). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3085530>

- Swingler, K. (2011). *The perils of ignoring data suitability: The suitability of data used to train neural networks deserves more attention*. Presented at the NCTA 2011 - Proceedings of the International Conference on Neural Computation Theory and Applications. Retrieved from <http://hdl.handle.net/1893/3950>
- Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big Data*, 5(3), 246-255. Retrieved from <https://www.liebertpub.com/doi/abs/10.1089/big.2016.0051>
- Vidgen, B., Nguyen, D., Tromble, R., Hale, S., Margetts, H., Harris, A. (2019) 'Challenges and frontiers in abusive content detection', *Forthcoming ACL 2019*.
- Zheng, X., Wang, M., & Ordieres-Meré, J. (2018). Comparison of data preprocessing approaches for applying deep learning to human activity recognition in the context of industry 4.0. *Sensors*, 18(7), 2146. Retrieved from <https://www.mdpi.com/1424-8220/18/7/2146>

## Design fairness

- Barocas, S., & Selbst, A. D. (2016). Big Data's disparate impact. *Calif. L. Rev.*, 104, 671. Retrieved from [https://heinonline.org/hol-cgi-bin/get\\_pdf.cgi?handle=hein.journals/calr104&section=25](https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/calr104&section=25)
- Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2), 277–292. <https://doi.org/10.1007/s10618-010-0190-x>
- Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems* (pp. 3992-4001). Retrieved from <http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention>
- d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2), 120-134. <https://doi.org/10.1089/big.2016.0048>
- Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125-2126). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2945386>
- Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1-33. Retrieved from <https://link.springer.com/article/10.1007/s10115-011-0463-8>
- Lehr, D., & Ohm, P. (2017). Playing with the data: What legal scholars should learn about machine learning. *UCDL Rev.*, 51, 653. Retrieved from [https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2\\_Lehr\\_Ohm.pdf](https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf)
- Passi, S., & Barocas, S. (2019). Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 39-48). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287567>
- Singh, J., & Sane, S. S. (2014). Preprocessing technique for discrimination prevention in data mining. *The IJES*, 3(6), 12-16. Retrieved from [https://www.academia.edu/6994180/Pre-Processing\\_Approach\\_for\\_Discrimination\\_Prevention\\_in\\_Data\\_Mining](https://www.academia.edu/6994180/Pre-Processing_Approach_for_Discrimination_Prevention_in_Data_Mining)
- Singhal, S., & Jena, M. (2013). A study on WEKA tool for data preprocessing, classification and clustering. *International Journal of Innovative technology and exploring engineering (IJItee)*, 2(6), 250-253. Retrieved from [https://pdfs.semanticscholar.org/095c/fd6f1a9dc6eaac7cc3100\\_a16cca9750ff9d8.pdf](https://pdfs.semanticscholar.org/095c/fd6f1a9dc6eaac7cc3100_a16cca9750ff9d8.pdf)
- van der Aalst, W. M., Bichler, M., & Heinzl, A. (2017). Responsible data science. *Springer Fachmedien Wiesbaden*. <https://doi.org/10.1007/s12599-017-0487-z>

## Outcome fairness

- Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *ArXiv:1803.02453*. Retrieved from <http://arxiv.org/abs/1803.02453>
- Albarghouthi, A., & Vinitsky, S. (2019). Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 211-219). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287588>
- Chiappa, S., & Gillam, T. P. (2018). Path-specific counterfactual fairness. *arXiv:1802.08139*. Retrieved from <https://arxiv.org/abs/1802.08139>
- Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv:1610.07524*. Retrieved from <http://arxiv.org/abs/1610.07524>
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *ArXiv:1701.08230*. <https://doi.org/10.1145/3097983.309809>
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2090255>
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259-268). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2783311>
- Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 329-338). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287589>
- Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law* (Vol. 1, p. 2). Retrieved from <http://www.mlandthelaw.org/papers/grgic.pdf>
- Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2017). On Fairness, Diversity and Randomness in Algorithmic Decision Making. *arXiv:1706.10208*. Retrieved from <https://arxiv.org/abs/1706.10208>
- Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*. Retrieved from <http://mlg.eng.cam.ac.uk/adrian/AAAI18-BeyondDistributiveFairness.pdf>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323). Retrieved from <http://papers.nips.cc/paper/6373-equality-of-opportunity-in-supervised-learning>
- Johansson, F. D., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. *ArXiv:1605.03661*. Retrieved from <http://arxiv.org/abs/1605.03661>
- Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In P. A. Flach, T. De Bie, & N. Cristianini (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 7524, pp. 35–50). [https://doi.org/10.1007/978-3-642-33486-3\\_3](https://doi.org/10.1007/978-3-642-33486-3_3)
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *ArXiv:1609.05807*. Retrieved from <http://arxiv.org/abs/1609.05807>
- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* (pp. 4066-4076). Retrieved from <http://papers.nips.cc/paper/6995-counterfactual-fairness>
- Russell, C., Kusner, M. J., Loftus, J., & Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 6414–6423). Retrieved from <http://papers.nips.cc/paper/7220-when-worlds-collide-integrating-different-counterfactual-assumptions-in-fairness.pdf>

- Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 10-19). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287566>
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1-7). IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8452913>
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *ArXiv:1711.00399*. Retrieved from <http://arxiv.org/abs/1711.00399>
- Wexler, J. (2018). The what-if tool: Code-free probing of machine. *Google AI Blog*. Retrieved from <https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html>
- Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. *arXiv:1507.05259*. Retrieved from <https://arxiv.org/abs/1507.05259>
- Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1171-1180). International World Wide Web Conferences Steering Committee. Retrieved from <https://dl.acm.org/citation.cfm?id=3052660>
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning* (pp. 325-333). Retrieved from <http://proceedings.mlr.press/v28/zemel13.pdf>
- Zhang, J., & Bareinboim, E. (2018). *Fairness in decision-making the causal explanation formula*. Presented at the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018. Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16949>
- Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4), 1060-1089. Retrieved from <https://link.springer.com/article/10.1007/s10618-017-0506-1>

## Implementation fairness

- Alexander, V., Blinder, C., & Zak, P. J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Computers in Human Behavior*, 89, 279-288. <https://doi.org/10.1016/j.chb.2018.07.026>
- Bahner, J. E., Hüper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, 66(9), 688-699. <https://doi.org/10.1016/j.ijhcs.2008.06.001>
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211-233. [https://doi.org/10.1016/0001-6918\(80\)90046-3](https://doi.org/10.1016/0001-6918(80)90046-3)
- Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions, *Cognition*, Vol. 181, 21-34. <https://doi.org/10.1016/j.cognition.2018.08.003>
- Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency* (No. ARL-TR-6905). Aberdeen Proving Ground, MD: U.S. Army Research Laboratory. Retrieved from <https://apps.dtic.mil/docs/citations/ADA600351>
- Crocill, W. M., & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 34, No. 19, pp. 1524-1528). Los Angeles, CA: SAGE Publications. <https://doi.org/10.1177/154193129003401922>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114. Retrieved from [https://repository.upenn.edu/cgi/viewcontent.cgi?article=1392&context=fnce\\_papers](https://repository.upenn.edu/cgi/viewcontent.cgi?article=1392&context=fnce_papers)
- Domeinski, J., Wagner, R., Schoebel, M., & Manzey, D. (2007). Human redundancy in automation monitoring: Effects of social loafing and social compensation. In *Proceedings of the Human Factors and Ergonomics*

- Society 51st Annual Meeting (pp. 587–591). Santa Monica, CA: Human Factors and Ergonomics Society. <https://doi.org/10.1177/154193120705101004>
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1), 79–94. <https://doi.org/10.1518/0018720024494856>
- Gigerenzer, G., & Todd, P. A. (1999). *Simple heuristics that make us smart*. London, England: Oxford University Press.
- Gilovich, Thomas (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press.
- Kahneman, D. (2000). Evaluation by moments: Past and future. *Choices, values, and frames*, 693-708. Retrieved from <http://www.vwl.tuwien.ac.at/hanappi/TEI/momentfull.pdf>
- Kahneman, D. (2011). Thinking, fast and slow. London, England: Allen Lane.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgement under uncertainty: Heuristics and biases. New York, NY: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, 80(4), 237. Retrieved from <https://web.archive.org/web/20160518202232/https://faculty.washington.edu/jmiyamot/p466/kahneman%20psych%20o%20prediction.pdf>
- Karau, S. J., & Williams, K. D. (1993). Social-loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4), 681–706. Retrieved from <https://psycnet.apa.org/buy/1994-33384-001>
- Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological review*, 107(4), 852. <http://dx.doi.org/10.1037/0033-295X.107.4.852>
- Lee, J. D., & See, J. (2004). Trust in automation and technology: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1). <https://doi.org/10.1177/2053951718756684>
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90-103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11), 2098. <http://dx.doi.org/10.1037/0022-3514.37.11.2098>
- McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, 48, 656–665. <https://doi.org/10.1518/001872006779166334>
- Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors*, 58(3), 401-415. <https://doi.org/10.1177/0018720815621206>
- Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics*, 31, 175–178. Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science*, 1, 354–365. [https://doi.org/10.1016/S0169-8141\(02\)00194-4](https://doi.org/10.1016/S0169-8141(02)00194-4)
- Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other?. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and application* (pp. 201–220). Mahwah, NJ: Erlbaum.
- Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision-making and performance in hightech cockpits. *International Journal of Aviation Psychology*, 8, 47–63. [https://doi.org/10.1207/s15327108ijap0801\\_3](https://doi.org/10.1207/s15327108ijap0801_3)
- Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22, 390 – 409. <http://dx.doi.org/10.1002/bdm.637>

- Packin, N. G. (2019). Algorithmic Decision-Making: The Death of Second Opinions?. *New York University Journal of Legislation and Public Policy, Forthcoming*. Retrieved from [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3361639](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3361639)
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3), 381-410. <https://doi.org/10.1177/0018720810376055>
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, 3(1), 1-23. [https://doi.org/10.1207/s15327108ijap0301\\_1](https://doi.org/10.1207/s15327108ijap0301_1)
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, 49(1), 76-87. <https://doi.org/10.1518/001872007779598082>
- Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., & Keim, D. A. (2015). The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1), 240-249. [https://bib.dbvis.de/uploadedFiles/uncertainty\\_trust.pdf](https://bib.dbvis.de/uploadedFiles/uncertainty_trust.pdf)
- Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, 43, 573–583. <https://doi.org/10.1518/001872001775870403>
- Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3), 377-400. <https://pdfs.semanticscholar.org/629b/f1f076f8d5bc203c573d4ba1dad5bb6743cf.pdf>
- Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & cognition*, 21(4), 546-556. <https://doi.org/10.3758/BF03197186>
- Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755-769. <https://doi.org/10.1111/j.1540-5907.2006.00214.x>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157), 1124-1131. Retrieved from <https://science.scienmag.org/content/185/4157/1124>
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453-458. Retrieved from <https://science.scienmag.org/content/211/4481/453>

## Accountability

- AI Now Institute. (2018). *Algorithmic Accountability Policy Toolkit*. Retrieved from <https://ainowinstitute.org/aap-toolkit.pdf>
- Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & Technology*, 31(4), 543-556. Retrieved from <https://link.springer.com/article/10.1007/s13347-017-0263-5>
- Cavoukian, A., Taylor, S., & Abrams, M. E. (2010). Privacy by Design: essential for organizational accountability and strong business practices. *Identity in the Information Society*, 3(2), 405–413. <https://doi.org/10.1007/s12394-010-0053-z>
- Center for Democracy & Technology. (n.d.). *Digital decisions*. Retrieved from <https://cdt.org/issue/privacy-data/digital-decisions/>
- Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism*, 3(3), 398-415. <https://doi.org/10.1080/21670811.2014.976411>
- Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., ... & Wilson, C. (2017). Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML*. Retrieved from <https://www.fatml.org/resources/principles-for-accountable-algorithms>
- Donovan, J., Caplan, R., Hanson, L., & Matthews, J. (2018). Algorithmic accountability: A primer. *Data & Society Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality*. Retrieved from <https://datasociety.net/output/algorithmic-accountability-a-primer/>

- ICO. (2017). *Big Data, artificial intelligence, machine learning and data protection*. Retrieved from <https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf>
- Janssen, M., & Kuk, G. (2016). The challenges and limits of Big Data algorithms in technocratic governance. *Government Information Quarterly*, 33(3), 371–377. <https://doi.org/10.1016/j.giq.2016.08.011>
- Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *U. Pa. L. Rev.*, 165, 633. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/pnlr165&div=20&id=&page=&t=1559932490>
- Malgieri, G., & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law*. Retrieved from <https://academic.oup.com/idpl/article-abstract/7/4/243/4626991?redirectedFrom=fulltext>
- O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., ... & Ashrafi, H. (2019). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 15(1), e1968. <https://doi.org/10.1002/rcs.1968>
- Reed, C. (2018). How should we regulate artificial intelligence?. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128), 20170360. Retrieved from <https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2017.0360>
- Stahl, B. C., & Wright, D. (2018). Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security & Privacy*, 16(3), 26–33. <https://doi.org/10.1109/MSP.2018.2701164>
- Veale, M., Binns, R., & Edwards, L. (2018). Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133). <https://doi.org/10.1098/rsta.2018.0083>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017a). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, 2(6). <https://doi.org/10.1126/scirobotics.aan6080>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017b). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. <https://doi.org/10.1093/idpl/ixp005>
- Zook, M., Barocas, S., boyd, danah, Crawford, K., Keller, E., Gangadharan, S. P., ... Pasquale, F. (2017). Ten simple rules for responsible Big Data research. *PLOS Computational Biology*, 13(3). <https://doi.org/10.1371/journal.pcbi.1005399>

## Stakeholder Impact Assessment

- AI Now Institute. (2018). Algorithmic Impact Assessments: Toward Accountable Automation in Public Agencies. Retrieved from <https://medium.com/@AINowInstitute/algorithmic-impact-assessments-toward-accountable-automation-in-public-agencies-bd9856e6fdde>
- Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H. V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C., & Zevenbergen, B. (n.d.). Principles for accountable algorithms and a social impact statement for algorithms. Fairness, Accountability, and Transparency in Machine Learning. Retrieved from: <http://www.fatml.org/resources/principles-for-accountable-algorithms>
- Karlin, M. (2018). A Canadian algorithmic impact assessment. Retrieved from <https://medium.com/@supergovernance/a-canadian-algorithmic-impact-assessment-128a2b2e7f85>
- Karlin, M., & Corriveau, N. (2018). The government of Canada's algorithmic impact assessment: Take two. Retrieved from <https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f>
- Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now institute*. Retrieved from: <https://ainowinstitute.org/aiareport2018.pdf>

Vallor, S. (2018) An ethical toolkit for engineering/design practice. Retrieved from: <https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/>

## Hong Kong

The Information Accountability Foundation. (2018a). *Ethical accountability framework for Hong Kong, China: A report prepared for the Office of the Privacy Commission for Personal Data*. Retrieved from [https://www.pcpd.org.hk/mis/files/Ethical\\_Accountability\\_Framework.pdf](https://www.pcpd.org.hk/mis/files/Ethical_Accountability_Framework.pdf)

The Information Accountability Foundation. (2018b). *Data stewardship accountability, data impact assessments and oversight models: Detailed support for an ethical accountability framework*. Retrieved from [https://www.pcpd.org.hk/mis/files/Ethical\\_Accountability\\_Framework\\_Detailed\\_Support.pdf](https://www.pcpd.org.hk/mis/files/Ethical_Accountability_Framework_Detailed_Support.pdf)

## Canada

Treasury Board of Canada Secretariat. (2019). *Algorithmic impact assessment*. Retrieved from <https://open.canada.ca/data/en/dataset/748a97fb-6714-41ef-9fb8-637a0b8e0da1>

## Safety: Accuracy, reliability, security, and robustness

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv:1606.06565*. Retrieved from <https://arxiv.org/abs/1606.06565>

Auerhammer, K., Kolagari, R. T., & Zoppelt, M. (2019). Attacks on Machine Learning: Lurking Danger for Accountability [PowerPoint Slides]. Retrieved from <https://safeai.webs.upv.es/wp-content/uploads/2019/02/3.SafeAI.pdf>

Demšar, J., & Bosnić, Z. (2018). Detecting concept drift in data streams using model explanation. *Expert Systems with Applications*, 92, 546–559. <https://doi.org/10.1016/j.eswa.2017.10.003>

Google. (2019). *Perspectives on issues in AI governance*. Retrieved from <https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf>

Göpfert, J. P., Hammer, B., & Wersing, H. (2018). Mitigating concept drift via rejection. In *International Conference on Artificial Neural Networks* (pp. 456-467). Springer, Cham. [https://doi.org/10.1007/978-3-030-01418-6\\_45](https://doi.org/10.1007/978-3-030-01418-6_45)

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill*, 4(2). <https://doi.org/10.23915/distill.00014>

Kohli, P., Dvijotham, K., Uesato, J., & Gowal, S. (2019). Towards a robust and verified AI: Specification testing, robust training, and formal verification. *DeepMind Blog*. Retrieved from <https://deepmind.com/blog/robust-and-verified-ai/>

Kolter, Z., & Madry, A. (n.d.). Materials for tutorial adversarial robustness: Theory and practice. Retrieved from <https://adversarial-ml-tutorial.org/>

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv:1801.00631*. Retrieved from <https://arxiv.org/abs/1801.00631>

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C., & Roli, F. (2017, November). Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 27-38). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3140451>

Nicolae, M. I., Sinn, M., Tran, M. N., Rawat, A., Wistuba, M., Zantedeschi, V., ... & Edwards, B. (2018). Adversarial Robustness Toolbox v0.4.0. *arXiv:1807.01069*. Retrieved from <https://arxiv.org/abs/1807.01069>

- Ortega, P. A., & Maini, V. (2018). Building safe artificial intelligence: specification, robustness, and assurance. *DeepMind Safety Research Blog, Medium*. Retrieved from <https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1>
- Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. (2017). Improving network robustness against adversarial attacks with compact convolution. *arXiv:1712.00699*. Retrieved from <https://arxiv.org/abs/1712.00699>
- Ratasich, D., Khalid, F., Geissler, F., Grosu, R., Shafique, M., & Bartocci, E. (2019). A roadmap toward the resilient internet of things for cyber-physical systems. *IEEE Access*, 7, 13260-13283. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8606923>
- Salay, R., & Czarnecki, K. (2018). Using machine learning safely in automotive software: An assessment and adaption of software process requirements in iso 26262. *arXiv:1808.01614*. Retrieved from <https://arxiv.org/abs/1808.01614>
- Shi, Y., Erpek, T., Sagduyu, Y. E., & Li, J. H. (2018). Spectrum data poisoning with adversarial deep learning. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)* (pp. 407-412). IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8599832/>
- Song, Q., Jin, H., Huang, X., & Hu, X. (2018). Multi-Label Adversarial Perturbations. In *2018 IEEE International Conference on Data Mining (ICDM)* (pp. 1242-1247). IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8594975>
- Warde-Farley, D., & Goodfellow, I. (2016). Adversarial perturbations of deep neural networks. In T. Hazan, G. Papandreou, & D. Tarlow (Eds.), *Perturbations, Optimization, and Statistics*, 311. Cambridge, MA: The MIT Press.
- Webb, G. I., Lee, L. K., Goethals, B., & Petitjean, F. (2018). Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32(5), 1179-1199. Retrieved from <https://link.springer.com/article/10.1007/s10618-018-0554-1>
- Zantedeschi, V., Nicolae, M. I., & Rawat, A. (2017). Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 39-49). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3140449>
- Zhao, M., An, B., Yu, Y., Liu, S., & Pan, S. J. (2018). Data poisoning attacks on multi-task relationship learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*. Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16073>
- Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2019). Adversarial attacks on deep learning models in natural language processing: A survey. 1(1). *arXiv:1901.06796*. <https://arxiv.org/abs/1901.06796>

## Transparency

- ACM US Public Policy Council. (2017). *Statement on algorithmic transparency and accountability*. Retrieved from [https://www.acm.org/binaries/content/assets/public-policy/2017\\_usacm\\_statement\\_algorithms.pdf](https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf)
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973-989. Retrieved from <https://journals.sagepub.com/doi/abs/10.1177/1461444816676645>
- Antunes, N., Balby, L., Figueiredo, F., Lourenco, N., Meira, W., & Santos, W. (2018). Fairness and transparency of machine learning for trustworthy cloud services. *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 188–193. <https://doi.org/10.1109/DSN-W.2018.00063>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 205395171562251. <https://doi.org/10.1177/2053951715622512>
- Citron, D. K. (2008). Technological due process. *Washington University Law Review*, 85(6). Retrieved from [https://heinonline.org/hol-cgi-bin/get\\_pdf.cgi?handle=hein.journals/walq85&section=38](https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/walq85&section=38)

- Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89, 1. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/washlr89&div=4&id=&page=&t=1560014586>
- Crawford, K., & Schultz, J. (2014). Big Data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, 55, 93. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/bclr55&div=5&id=&page=&t=1560014537>
- Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16, 18. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/dltr16&div=3&id=&page=&t=1560014649>
- Kemper, J., & Kolkman, D. (2018). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, 1-16. Retrieved from <https://www.tandfonline.com/doi/full/10.1080/1369118X.2018.1477967>
- Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, 11(2), 105–112. <https://doi.org/10.1007/s10676-009-9187-9>
- Weller, A. (2017). Challenges for transparency. arXiv preprint arXiv:1708.01870. Retrieved from <https://arxiv.org/abs/1708.01870>

## Process-Based Governance

- Andrews, L., Benbouzid, B., Brice, J., Bygrave, L. A., Demortain, D., Griffiths, A., ... & Yeung, K. (2017). Algorithmic Regulation. *The London School of Economics and Political Science*. Retrieved from <https://www.kcl.ac.uk/law/research/centres/telos/assets/DP85-Algorithmic-Regulation-Sep-2017.pdf>
- Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Reimer, D., Olteanu, A., Tsay, J., & Varshney, K. R & Piorkowski, D. (2018). FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *arXiv:1808.07261*. Retrieved from <https://arxiv.org/abs/1808.07261>
- Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587-604. Retrieved from [https://www.mitpressjournals.org/doi/abs/10.1162/tacl\\_a\\_00041](https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00041)
- Calo, R. (2017). Artificial Intelligence policy: a primer and roadmap. *UCDL Rev.*, 51, 399. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/davlr51&div=18&id=&page=&t=1560015127>
- D'Agostino, M., & Durante, M. (2018). Introduction: The governance of algorithms. *Philosophy & Technology*, 31(4), 499–505. <https://doi.org/10.1007/s13347-018-0337-z>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv:1803.09010*. Retrieved from <https://arxiv.org/abs/1803.09010>
- Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *arXiv:1805.03677*. Retrieved from <https://arxiv.org/abs/1805.03677>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287596>
- Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., ... & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*, 162(1), W1-W73. Retrieved from <https://annals.org/aim/fullarticle/2088542>
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. *arXiv:1905.06876*. Retrieved from <https://arxiv.org/abs/1905.06876>

- Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: a practical framework for public agency accountability. *AI Now*. Retrieved from <https://ainowinstitute.org/aiareport2018.pdf>
- Saurwein, F., Just, N., & Latzer, M. (2015). Governance of algorithms: options and limitations. *info*, 17(6), 35–49. Retrieved from <https://www.emeraldinsight.com/doi/abs/10.1108/info-05-2015-0025>
- Tutt, A., (2016). An FDA for algorithms. 69 *Admin. L. Rev.* 83 (2017). <http://dx.doi.org/10.2139/ssrn.2747994>
- Wachter, S., & Mittelstadt, B. D. (2018). A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI. *Columbia Business Law Review*. Retrieved from [https://ora.ox.ac.uk/objects/uuid:d53f7b6a-981c-4f87-91bc-743067d10167/download\\_file?file\\_format=pdf&safe\\_filename=Wachter%2Band%2BMittelstadt%2B2018%2B-%2BA%2Bright%2Bto%2Breasonable%2Binferences%2B-%2BVersion%2B6%2Bssrn%2Bversion.pdf&type\\_of\\_work=Journal+article](https://ora.ox.ac.uk/objects/uuid:d53f7b6a-981c-4f87-91bc-743067d10167/download_file?file_format=pdf&safe_filename=Wachter%2Band%2BMittelstadt%2B2018%2B-%2BA%2Bright%2Bto%2Breasonable%2Binferences%2B-%2BVersion%2B6%2Bssrn%2Bversion.pdf&type_of_work=Journal+article)

## Interpretable AI

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8466590>
- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1), 8753-8830. Retrieved from <http://www.jmlr.org/papers/volume18/17-716/17-716.pdf>
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7), e0130140. <https://doi.org/10.1371/journal.pone.0130140>
- Bathaei, Y. (2018). The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, 31(2), 889. Retrieved from <https://www.questia.com/library/journal/1G1-547758123/the-artificial-intelligence-black-box-and-the-failure>
- Bibal, A., & Frénay, B. (2016). *Interpretability of Machine Learning Models and Representations: an Introduction*. Retrieved from [https://www.researchgate.net/profile/Adrien\\_Bibal/publication/326839249\\_Interpretability\\_of\\_Machine\\_Learning\\_Models\\_and\\_Representations\\_an\\_Introduction/links/5b6861caa6fdcc87df6d58e4/Interpretability-of-Machine-Learning-Models-and-Representations-an-Introduction.pdf](https://www.researchgate.net/profile/Adrien_Bibal/publication/326839249_Interpretability_of_Machine_Learning_Models_and_Representations_an_Introduction/links/5b6861caa6fdcc87df6d58e4/Interpretability-of-Machine-Learning-Models-and-Representations-an-Introduction.pdf)
- Bracamonte, V. (2019). *Challenges for transparent and trustworthy machine learning* [Power Point]. KDDI Research, Inc. Retrieved from [https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20190121/Documents/Vanessa\\_Bracamonte\\_Presentation.pdf](https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20190121/Documents/Vanessa_Bracamonte_Presentation.pdf)
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>
- Card, D. (2017). The “black box” metaphor in machine learning. *Towards Data Science*. Retrieved from <https://towardsdatascience.com/the-black-box-metaphor-in-machine-learning-4e57a3a1d2b0>
- Caruana, R., Kangaroo, H., Dionisio, J. D., Sinha, U., & Johnson, D. (1999). Case-based explanation of non-case-based learning methods. *Proceedings. AMIA Symposium*, 212–215. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232607/>
- Chen, C., Li, O., Tao, C., Barnett, A., Su, J., & Rudin, C. (2018). This looks like that: deep learning for interpretable image recognition. *arXiv:1806.10574*. Retrieved from <https://arxiv.org/abs/1806.10574>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*. Retrieved from <https://arxiv.org/abs/1702.08608>
- Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., ... & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv:1711.01134*. Retrieved from <https://arxiv.org/abs/1711.01134>

- Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. <https://doi.org/10.23919/MIPRO.2018.8400040>
- Eisenstadt, V., & Althoff, K. (2018). A Preliminary Survey of Explanation Facilities of AI-Based Design Support Approaches and Tools. *LWDA*. Presented at the LWDA. [https://www.researchgate.net/profile/Viktor\\_Eisenstadt/publication/327339350\\_A\\_Preliminary\\_Survey\\_of\\_Explanation\\_Facilities\\_of\\_AI-Based\\_Design\\_Support\\_Approaches\\_and\\_Tools/links/5b891ecd299bf1d5a7338b1a/A-Preliminary-Survey-of-Explanation-Facilities-of-AI-Based-Design-Support-Approaches-and-Tools.pdf](https://www.researchgate.net/profile/Viktor_Eisenstadt/publication/327339350_A_Preliminary_Survey_of_Explanation_Facilities_of_AI-Based_Design_Support_Approaches_and_Tools/links/5b891ecd299bf1d5a7338b1a/A-Preliminary-Survey-of-Explanation-Facilities-of-AI-Based-Design-Support-Approaches-and-Tools.pdf)
- Feldmann, F. (2018). *Measuring machine learning model interpretability*. Retrieved from [https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/860270201/felix\\_feldmann\\_eml2018\\_report.pdf](https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/860270201/felix_feldmann_eml2018_report.pdf)
- Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3429-3437). Retrieved from [http://openaccess.thecvf.com/content\\_iccv\\_2017/html/Fong\\_Interpretable\\_Explanations\\_of\\_ICCV\\_2017\\_paper.html](http://openaccess.thecvf.com/content_iccv_2017/html/Fong_Interpretable_Explanations_of_ICCV_2017_paper.html)
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. Explaining explanations: An approach to evaluating interpretability of machine. *arXiv:1806.00069*. Retrieved from <https://arxiv.org/abs/1806.00069>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 93. Retrieved from <https://dl.acm.org/citation.cfm?id=3236009>
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics*. <https://doi.org/10.1093/qje/qjx032>
- Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), 20180084. <https://doi.org/10.1098/rsta.2018.0084>
- Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1675-1684). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2939874>
- Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 275–284. <https://doi.org/10.1145/3097983.3098066>
- Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4), 611-627. <https://doi.org/10.1007/s13347-017-0279-x>
- Li, O., Liu, H., Chen, C., & Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*. Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/17082>
- Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv:1606.03490*. Retrieved from <https://arxiv.org/abs/1606.03490>
- Lipton, Z. C., & Steinhardt, J. (2018). Troubling trends in machine learning scholarship. *arXiv:1807.03341*. Retrieved from <https://arxiv.org/abs/1807.03341>
- Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13*, 623. <https://doi.org/10.1145/2487575.2487579>
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *ArXiv:1705.07874*. Retrieved from <http://arxiv.org/abs/1705.07874>
- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3287574>

- Molnar, C. (2018). Interpretable machine learning. A guide for making black box models explainable. *Leanpub*. Retrieved from <https://christophm.github.io/interpretable-ml-book/>
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. *arXiv:1901.04592*. Retrieved from <https://arxiv.org/abs/1901.04592>
- Olhede, S. C., & Wolfe, P. J. (2018). The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128). <https://doi.org/10.1098/rsta.2017.0364>
- Park, D. H., Hendricks, L. A., Akata, Z., Schiele, B., Darrell, T., & Rohrbach, M. (2016). Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv:1612.04757*. Retrieved from <https://arxiv.org/abs/1612.04757>
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S., & Turini, F. (2018). Open the black box data-driven explanation of black box decision systems. *arXiv:1806.09936*. Retrieved from <https://arxiv.org/abs/1806.09936>
- Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box AI decision systems. *AAAI Press*.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability. *ArXiv:1802.07810*. Retrieved from <http://arxiv.org/abs/1802.07810>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv:1606.05386*. Retrieved from <https://arxiv.org/abs/1606.05386>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM. Retrieved from <https://dl.acm.org/citation.cfm?Id=2939778>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Thirty-second AAAI Conference on Artificial Intelligence*. Retrieved from <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16982>
- Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *arXiv:1811.10154*. Retrieved from <https://arxiv.org/abs/1811.10154>
- Rudin, C., & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5), 449-466. <https://doi.org/10.1287/inte.2018.0957>
- Shmueli, G. (2010). To explain or to predict?. *Statistical science*, 25(3), 289-310. Retrieved from <https://projecteuclid.org/euclid.ss/1294167961>
- Shaywitz, D. (2018). AI doesn't ask why – But physicians and drug developers want to know. *Forbes*. Retrieved from <https://www.forbes.com/sites/davidshaywitz/2018/11/09/ai-doesnt-ask-why-but-physicians-and-drug-developers-want-to-know/>
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *ArXiv:1704.02685*. Retrieved from <http://arxiv.org/abs/1704.02685>
- Simonite, T. (2017). AI experts want to end "black box" algorithms in government. *Wired Business*, 10, 17. Retrieved from <https://www.wired.com/story/ai-experts-want-to-end-black-box-algorithms-in-government/>
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *ArXiv:1312.6034*. Retrieved from <http://arxiv.org/abs/1312.6034>
- Sokol, K., & Flach, P. (2018). Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 5868–5870. <https://doi.org/10.24963/ijcai.2018/865>
- Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3), 349-391. Retrieved from: <https://link.springer.com/article/10.1007/s10994-015-5528-6>
- Zhang, Q., & Zhu, S. (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1), 27–39. <https://doi.org/10.1631/FITEE.1700808>

## Responsible delivery through human-centred implementation protocols and practices

- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 582). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=3174156>
- Antaki, C., & Leudar, I. (1992). Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2), 181-194. <https://doi.org/10.1002/ejsp.2420220206>
- Arioua, A., & Croitoru, M. (2015). Formalizing explanatory dialogues. In *International Conference on Scalable Uncertainty Management* (pp. 282-297). Springer, Cham. [https://doi.org/10.1007/978-3-319-23540-0\\_19](https://doi.org/10.1007/978-3-319-23540-0_19)
- Bex, F., & Walton, D. (2016). Combining explanation and argumentation in dialogue. *Argument & Computation*, 7(1), 55-68. Retrieved from <https://content.iospress.com/articles/argument-and-computation/aac001>
- Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (Vol. 8). Retrieved from [http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17\\_XAI\\_WS\\_Proceedings.pdf#page=8](http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17_XAI_WS_Proceedings.pdf#page=8)
- Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. *arXiv:1901.03729*. Retrieved from <https://arxiv.org/abs/1901.03729>
- Ginet, C. (2008). In defense of a non-causal account of reasons explanations. *The Journal of Ethics*, 12(3-4), 229-237. <https://doi.org/10.1007/s10892-008-9033-z>
- Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., ... & Holzinger, A. (2018). Explainable AI: the new 42?. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 295-303). Springer, Cham. [https://doi.org/10.1007/978-3-319-99740-7\\_21](https://doi.org/10.1007/978-3-319-99740-7_21)
- Habermas, J. (1993). Remarks on discourse ethics. *Justification and application: Remarks on discourse ethics*, 44, 313-314. Cambridge, UK: Polity Press.
- Habermas, J. (2003). Rightness versus truth: on the sense of normative validity in moral judgments and norms. *Truth and justification*, 248. Cambridge, UK: Polity Press.
- Hoffman, R. R., Mueller, S. T., & Klein, G. (2017). Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems*, 32(4), 78-86. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8012316>
- Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). A Grounded Interaction Protocol for Explainable Artificial Intelligence. *arXiv:1903.02409*. Retrieved from <https://arxiv.org/abs/1903.02409>
- McCarthy, T. (1974). The operation called Verstehen: Towards a redefinition of the problem. In *PSA 1972* (pp. 167-193). Springer, Dordrecht. [https://doi.org/10.1007/978-94-010-2140-1\\_12](https://doi.org/10.1007/978-94-010-2140-1_12)
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Rapanta, C., & Walton, D. (2016). The use of argument maps as an assessment tool in higher education. *International Journal of Educational Research*, 79, 211-221. <https://doi.org/10.1016/j.ijer.2016.03.002>
- Springer, A., & Whittaker, S. (2018). Progressive disclosure: Designing for effective transparency. *arXiv:1811.02164*. Retrieved from <https://arxiv.org/abs/1811.02164>
- Taylor, C. (1973). Interpretation and the sciences of man. In *Explorations in Phenomenology* (pp. 47-101). Springer, Dordrecht. [https://doi.org/10.1007/978-94-010-1999-6\\_3](https://doi.org/10.1007/978-94-010-1999-6_3)
- Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv:1806.07552*. Retrieved from <https://arxiv.org/abs/1806.07552>

- Tsai, C. H., & Brusilovsky, P. (2019). Designing explanation interfaces for transparency and beyond. In *Joint Proceedings of the ACM IUI 2019 Workshops*. Retrieved from <http://ceur-ws.org/Vol-2327/IUI19WS-IUIATEC-4.pdf>
- Von Wright, G. H. (2004). *Explanation and understanding*. Ithaca, NY: Cornell University Press.
- Walton, D. (2004). A new dialectical theory of explanation. *Philosophical Explorations*, 7(1), 71-89. <https://doi.org/10.1080/1386979032000186863>
- Walton, D. (2005). Dialectical Explanation in AI. *Argumentation Methods for Artificial Intelligence in Law*, 173-212. [https://doi.org/10.1007/3-540-27881-8\\_6](https://doi.org/10.1007/3-540-27881-8_6)
- Walton, D. (2007). Dialogical Models of Explanation. *ExaCt*, 2007, 1-9. Retrieved from <https://www.aaai.org/Papers/Workshops/2007/WS-07-06/WS07-06-001.pdf>
- Walton, D. (2011). A dialogue system specification for explanation. *Synthese*, 182(3), 349-374. <https://doi.org/10.1007/s11229-010-9745-z>
- Walton, D. (2016). Some artificial intelligence tools for argument evaluation: An introduction. *Argumentation*, 30(3), 317-340. <https://doi.org/10.1007/s10503-015-9387-x>
- Weld, D. S., & Bansal, G. (2018). The challenge of crafting intelligible intelligence. *arXiv:1803.04263*. Retrieved from <https://arxiv.org/abs/1803.04263>
- Walton, D., Toniolo, A., & Norman, T. (2016). Speech acts and burden of proof in computational models of deliberation dialogue. In *Proceedings of the First European Conference on Argumentation*, ed. D. Mohammed and M. Lewinski, London, College Publications (Vol. 1, pp. 757-776). Retrieved from [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2852054](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852054)
- Wendt, A. (1998). On constitution and causation in international relations. *Review of international studies*, 24(5), 101-118. <https://doi.org/10.1017/S0260210598001028>
- Winikoff, M. (2017). Debugging agent programs with why?: Questions. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (pp. 251-259). International Foundation for Autonomous Agents and Multiagent Systems. Retrieved from <https://dl.acm.org/citation.cfm?id=3091166>
- Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G. M. (2018). Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)* (pp. 1-8). IEEE. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8490433>

### Individual and societal impacts of machine learning and algorithmic systems

- Amoore, L. (2018a). Cloud geographies: Computing, data, sovereignty. *Progress in Human Geography*, 42(1), 4-24. <https://doi.org/10.1177/0309132516662147>
- Amoore, L. (2018b). Doubtful algorithms: of machine learning truths and partial accounts. *Theory, culture & society*. Retrieved from <http://dro.dur.ac.uk/26913/1/26913.pdf>
- Amoore, L., & Raley, R. (2017). Securing with algorithms: Knowledge, decision, sovereignty. *Security Dialogue*, 48(1), 3-10. <https://doi.org/10.1177/0967010616680753>
- Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, 41(1), 93-117. <https://doi.org/10.1177/0162243915606523>
- Anderson, B. (2010). Preemption, precaution, preparedness: Anticipatory action and future geographies. *Progress in Human Geography*, 34(6), 777-798. <https://doi.org/10.1177/0309132510362600>
- Anderson, B. (2010). Security and the future: Anticipating the event of terror. *Geoforum*, 41(2), 227-235. <https://doi.org/10.1016/j.geoforum.2009.11.002>
- Anderson, S. F. (2017). *Technologies of vision: The war between data and images*. MIT Press.

- Arnoldi, J. (2016). Computer algorithms, market manipulation and the institutionalization of high frequency trading. *Theory, Culture & Society*, 33(1), 29-52. <https://doi.org/10.1177/0263276414566642>
- Beer, D. (2013). Algorithms: Shaping tastes and manipulating the circulations of popular culture. In *Popular Culture and New Media* (pp. 63-100). Palgrave Macmillan, London. [https://doi.org/10.1057/9781137270061\\_4](https://doi.org/10.1057/9781137270061_4)
- Beer, D. (2017). The social power of algorithms. In *Information, Communication & Society*, (20), 1-13. <https://doi.org/10.1080/1369118X.2016.1216147>
- Bodó, B., Helberger, N., Irion, K., Zuiderveen Borgesius, F., Moller, J., van de Velde, B., ... & de Vreese, C. (2017). Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents. *Yale JL & Tech.*, 19, 133. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/yjolt19&div=4&id=&page=&t=1560029464>
- Bogost, I. (2015). The cathedral of computation. *The Atlantic*, 15. Retrieved from <https://www.theatlantic.com/technology/archive/2015/01/the-cathedral-of-computation/384300/>
- Bolin, G., & Andersson Schwarz, J. (2015). Heuristics of the algorithm: Big Data, user interpretation and institutional translation. *Big Data & Society*, 2(2). <https://doi.org/10.1177/2053951715608406>
- Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *NIPS*. Retrieved from <http://papers.nips.cc/paper/6227-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings>
- Browne, S. (2015). *Dark matters: On the surveillance of blackness*. Duke University Press.
- Bucher, T. (2017). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1), 30-44. <https://doi.org/10.1080/1369118X.2016.1154086>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186. Retrieved from <https://science.sciencemag.org/content/356/6334/183>
- Cheney-Lippold, J. (2011). A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society*, 28(6), 164-181. <https://doi.org/10.1177/0263276411424420>
- Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance & Society*, 15(5), 609-625. <https://doi.org/10.24908/ss.v15i5.6433>
- Crandall, J. (2006). Precision + guided + seeing. *CTheory*, 1-10. Retrieved from <https://journals.uvic.ca/index.php/ctheory/article/view/14468/5310>
- Crandall, J. (2010). The geospatialization of calculative operations: Tracking, sensing and megacities. *Theory, Culture & Society*, 27(6), 68-90. <https://doi.org/10.1177/0263276410382027>
- Crawford, K. (2014). The anxieties of Big Data. *The New Inquiry*, 30, 2014. Retrieved from <https://thenewinquiry.com/the-anxieties-of-big-data/>
- Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature News*, 538(7625), 311. Retrieved from <https://www.nature.com/news/there-is-a-blind-spot-in-ai-research-1.20805>
- Eckhouse, L., Lum, K., Conti-Cook, C., & Ciccolini, J. (2019). Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, 46(2), 185-209. Retrieved from <https://journals.sagepub.com/doi/abs/10.1177/0093854818811379>
- Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., ... & Sandvig, C. (2015). I always assumed that I wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 153-162). ACM. Retrieved from <https://dl.acm.org/citation.cfm?id=2702556>
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
- Ferguson, A. G. (2017). Policing Predictive Policing. *Washington University Law Review*, 94(5). Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/walq94&div=35&id=&page=&t=1559934122>

- Geiger, R. S. (2014). Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society*, 17(3), 342-356. <https://doi.org/10.1080/1369118X.2013.873069>
- Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society*. Cambridge, MA: The MIT Press.
- Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, 3(2). <https://doi.org/10.1177/2053951716674238>
- Jasanoff, S. (2015). Future imperfect: Science, technology, and the imaginations of modernity. In S. Jasanoff & S. Kim (Eds.), *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. Chicago, IL: The University of Chicago Press.
- Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14-29. <https://doi.org/10.1080/1369118X.2016.1154087>
- Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *ArXiv:1805.04508*. Retrieved from <http://arxiv.org/abs/1805.04508>
- Kushner, S. (2013). The freelance translation machine: Algorithmic culture and the invisible industry. *New Media & Society*, 15(8), 1241-1258. <https://doi.org/10.1177/1461444812469597>
- Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2016). The tyranny of data? The bright and dark sides of data-driven decision-making for social good. *ArXiv:1612.00323*. Retrieved from <http://arxiv.org/abs/1612.00323>
- Mackenzie, A. (2015a). Machine learning and genomic dimensionality: From features to landscapes. In S. Richardson & H. Stevens (Eds.), *Postgenomics: Perspectives on Biology after the Genome*. Durham, NC: Duke University Press.
- Mackenzie, A. (2015b). The production of prediction: What does machine learning want?. *European Journal of Cultural Studies*, 18(4-5), 429-445. <https://doi.org/10.1177/1367549415577384>
- Mackenzie, A., & McNally, R. (2013). Living multiples: How large-scale scientific data-mining pursues identity and differences. *Theory, Culture & Society*, 30(4), 72-91. <https://doi.org/10.1177/0263276413476558>
- Mackenzie, A., & Vurdubakis, T. (2011). Codes and codings in crisis: signification, performativity and excess. *Theory, Culture & Society*, 28(6), 3-23. <https://doi.org/10.1177/0263276411424761>
- Mager, A. (2012). Algorithmic ideology: How capitalist society shapes search engines. *Information, Communication & Society*, 15(5), 769-787. <https://doi.org/10.1080/1369118X.2012.676056>
- Manokha, I. (2018). Surveillance, panopticism, and self-discipline in the digital age. *Surveillance & Society*, 16(2), 219-237. <https://doi.org/10.24908/ss.v16i2.8346>
- Matzner, T. (2014). Why privacy is not enough privacy in the context of “ubiquitous computing” and “Big Data.” *Journal of Information, Communication and Ethics in Society*, 12(2), 93–106. <https://doi.org/10.1108/JICES-08-2013-0030>
- Mendoza, I., & Bygrave, L. A. (2017). The right not to be subject to automated decisions based on profiling. In *EU Internet Law* (pp. 77-98). Springer, Cham. Retrieved from [https://link.springer.com/chapter/10.1007/978-3-319-64955-9\\_4](https://link.springer.com/chapter/10.1007/978-3-319-64955-9_4)
- Mollicchi, S. (2017). Flatness versus depth: A study of algorithmically generated camouflage. *Security Dialogue*, 48(1), 78-94. <https://doi.org/10.1177/0967010616650227>
- Molnar, P., & Gill, L. (2018). Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada’s Immigration and Refugee System. *Citizen Lab and International Human Rights Program (Faculty of Law, University of Toronto)*. Retrieved from <https://tspace.library.utoronto.ca/handle/1807/94802>
- Monahan, T. (2018). Algorithmic fetishism. *Surveillance & Society*, 16(1), 1-5. <https://doi.org/10.24908/ss.v16i1.10827>
- Murphy, M. H. (2017). Algorithmic surveillance: The collection conundrum. *International Review of Law, Computers & Technology*, 31(2), 225–242. <https://doi.org/10.1080/13600869.2017.1298497>
- Napoli, P. M. (2014). Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication Theory*, 24(3), 340-360. <https://doi.org/10.1111/comt.12039>

- Neyland, D. (2015). On organizing algorithms. *Theory, Culture & Society*, 32(1), 119-132.  
<https://doi.org/10.1177/0263276414530477>
- Neyland, D. (2016). Bearing account-able witness to the ethical algorithmic system. *Science, Technology, & Human Values*, 41(1), 50-76. <https://doi.org/10.1177/0162243915598056>
- Neyland, D., & Möllers, N. (2017). Algorithmic IF... THEN rules and the conditions and consequences of power. *Information, Communication & Society*, 20(1), 45-62. <https://doi.org/10.1080/1369118X.2016.1156141>
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. nyu Press.
- O'Grady, N. (2015). A politics of redeployment: malleable technologies and the localisation of anticipatory calculation. In *Algorithmic Life* (pp. 86-100). Routledge. Retrieved from <http://eprints.uwe.ac.uk/id/eprint/33134>
- Plantin, J. C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, 20(1), 293-310.  
<https://doi.org/10.1177/1461444816661553>
- Redden, J., & Brand, J. (2017). Data Harm Record. *Data Justice Lab*. Retrieved from <http://orca-mwe.cf.ac.uk/107924/1/data-harm-record-djl2.pdf>
- Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online, Forthcoming*. Retrieved from [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3333423](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423)
- Roberge, J., & Seyfert, R. (2016). What are algorithmic cultures. *Algorithmic cultures: Essays on meaning, performance and new technologies*, 1-25. Retrieved from <https://www.taylorfrancis.com/books/e/9781315658698/chapters/10.4324/9781315658698-7>
- Schüll, N. D. (2018). Self in the Loop: Bits, Patterns, and Pathways in the Quantified Self. In *A Networked Self and Human Augmentics, Artificial Intelligence, Sentience* (pp. 41-54). New York, NY: Routledge.
- Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, 87, 1085. Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/flr87&div=44&id=&page=&t=1560020999>
- Smith, G. (2018). High-tech redlining: AI is quietly upgrading institutional racism. *Fast Company*. Retrieved from <https://www.fastcompany.com/90269688/high-tech-redlining-ai-is-quietly-upgrading-institutional-racism>
- Striphas, T. (2015). Algorithmic culture. *European Journal of Cultural Studies*, 18(4-5), 395-412.  
<https://doi.org/10.1177/1367549415577392>
- Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), 197-208. <https://doi.org/10.24908/ss.v12i2.4776>
- Wilf, E., Cheney-Lippold, J., Duranti, A., Eisenlohr, P., Gershon, I., Mackenzie, A., ... & Wilf, E. (2013). Toward an anthropology of computer-mediated, algorithmic forms of sociality. *Current Anthropology*, 54(6), 000-000. Retrieved from <https://www.journals.uchicago.edu/doi/abs/10.1086/673321>
- Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society*, 20(1), 137-150.  
<https://doi.org/10.1080/1369118X.2016.1200645>
- Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1), 118-132. <https://doi.org/10.1177/0162243915605575>
- Ziewitz, M. (2016). Governing algorithms: Myth, mess, and methods. *Science, Technology, & Human Values*, 41(1), 3-16. <https://doi.org/10.1177/0162243915608948>
- Zuboff, S. (2019). *The age of surveillance capitalism: the fight for the future at the new frontier of power*. Profile Books.



**turing.ac.uk**  
**@turinginst**



European Parliament

---

# The ethics of artificial intelligence: Issues and initiatives

---

STUDY

Panel for the Future of Science and Technology

---

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)  
PE 634.452 – March 2020

EN



# The ethics of artificial intelligence: Issues and initiatives

---

This study deals with the ethical implications and moral questions that arise from the development and implementation of artificial intelligence (AI) technologies. It also reviews the guidelines and frameworks which countries and regions around the world have created to address them. It presents a comparison between the current main frameworks and the main ethical issues, and highlights gaps around the mechanisms of fair benefit-sharing; assigning of responsibility; exploitation of workers; energy demands in the context of environmental and climate changes; and more complex and less certain implications of AI, such as those regarding human relationships.

## AUTHORS

This study has been drafted by Eleanor Bird, Jasmin Fox-Skelly, Nicola Jenner, Ruth Larbey, Emma Weitkamp and Alan Winfield from the Science Communication Unit at the University of the West of England, at the request of the Panel for the Future of Science and Technology (STOA), and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

### *Acknowledgements*

The authors would like to thank the following interviewees: John C. Havens (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS)) and Jack Stilgoe (Department of Science & Technology Studies, University College London).

## ADMINISTRATOR RESPONSIBLE

Mihalis Kritikos, Scientific Foresight Unit (STOA)

To contact the publisher, please e-mail [stoa@ep.europa.eu](mailto:stoa@ep.europa.eu)

## LINGUISTIC VERSION

Original: EN

Manuscript completed in March 2020.

## DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.

Brussels © European Union, 2020.

PE 634.452  
ISBN: 978-92-846-5799-5  
doi: 10.2861/6644  
QA-01-19-779-EN-N

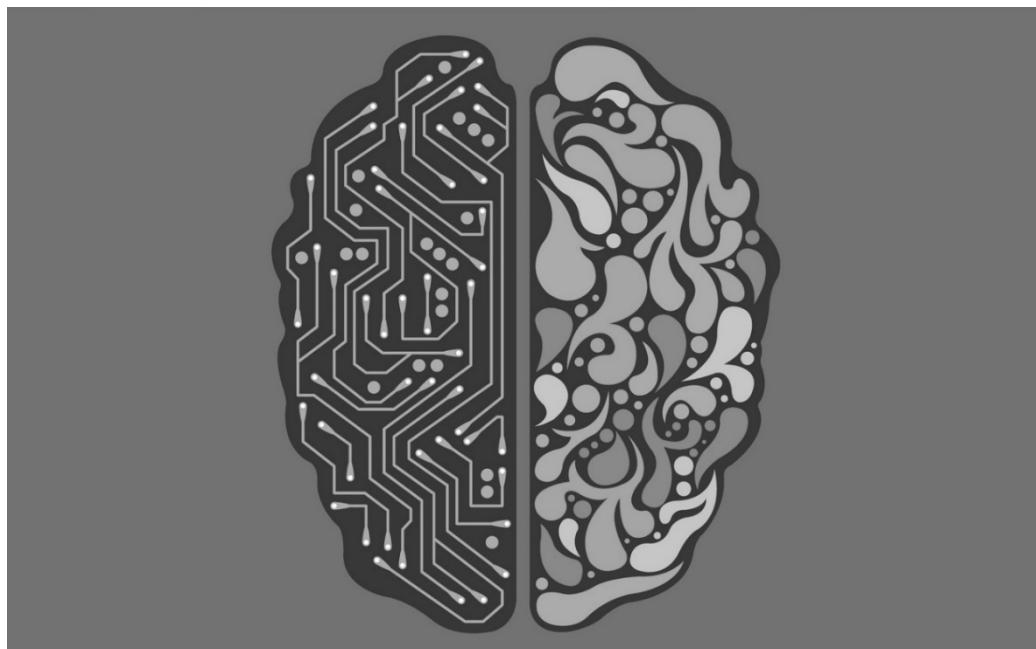
<http://www.europarl.europa.eu/stoa> (STOA website)

<http://www.eprs.ep.parl.union.eu> (intranet)

<http://www.europarl.europa.eu/thinktank> (internet)

<http://epthinktank.eu> (blog)

## Executive summary



© Seanbatty / Pixabay

This report deals with the ethical implications and moral questions that arise from the development and implementation of artificial intelligence (AI) technologies. It also reviews the guidelines and frameworks that countries and regions around the world have created to address them. It presents a comparison between the current main frameworks and the main ethical issues, and highlights gaps around mechanisms of fair benefit sharing; assigning of responsibility; exploitation of workers; energy demands in the context of environmental and climate changes; and more complex and less certain implications of AI, such as those regarding human relationships.

Chapter 1 introduces the scope of the report and defines key terms. The report draws on the European Commission's definition of AI as 'systems that display intelligent behaviour'. Other key terms defined in this chapter include intelligence and how this is used in the context of AI and intelligent robots (i.e. robots with an embedded AI), as well as defining machine learning, artificial neural networks and deep learning, before moving on to consider definitions of morality and ethics and how these relate to AI.

In Chapter 2 the report **maps the main ethical dilemmas and moral questions associated with the deployment of AI**. The report begins by outlining a number of potential benefits that could arise from AI as a context in which to situate ethical, social and legal considerations. Within the context of issues for society, the report considers the potential impacts of AI on the labour market, focusing on the likely impact on economic growth and productivity, the impact on the workforce, potential impacts on different demographics, including a worsening of the digital divide, and the consequences of deployment of AI on the workplace. The report considers the potential impact of AI on inequality and how the benefits of AI could be shared within society, as well as issues concerning the concentration of AI technology within large internet companies and political stability. Other societal issues addressed in this chapter include privacy, human rights and dignity, bias, and issues for democracy.

Chapter 2 moves on to consider the impact of AI on human psychology, raising questions about the impact of AI on relationships, as in the case of intelligent robots taking on human social roles, such as nursing. Human-robot relationships may also affect human-human relationships in as yet unanticipated ways. This section also considers the question of personhood, and whether AI systems should have moral agency.

Impacts on the financial system are already being felt, with AI responsible for high trading volumes of equities. The report argues that, although markets are suited to automation, there are risks including the use of AI for intentional market manipulation and collusion.

AI technology also poses questions for both civil and criminal law, particularly whether existing legal frameworks apply to decisions taken by AIs. Pressing legal issues include liability for tortious, criminal and contractual misconduct involving AI. While it may seem unlikely that AIs will be deemed to have sufficient autonomy and moral sense to be held liable themselves, they do raise questions about who is liable for which crime (or indeed if human agents can avoid liability by claiming they did not know the AI could or would do such a thing). In addition to challenging questions around liability, AI could abet criminal activities, such as smuggling (e.g. by using unmanned vehicles), as well as harassment, torture, sexual offences, theft and fraud. Self-driving autonomous cars are likely to raise issues in relation to product liability that could lead to more complex cases (currently insurers typically avoid lawsuits by determining which driver is at fault, unless a car defect is involved).

Large-scale deployment of AI could also have both positive and negative impacts on the environment. Negative impacts include increased use of natural resources, such as rare earth metals, pollution and waste, as well as energy consumption. However, AI could help with waste management and conservation offering environmental benefits.

The potential impacts of AI are far-reaching, but they also require trust from society. AI will need to be introduced in ways that build trust and understanding, and respect human and civil rights. This requires transparency, accountability, fairness and regulation.

Chapter 3 explores **ethical initiatives in the field of AI**. The chapter first outlines the ethical initiatives identified for this report, summarising their focus and where possible identifying funding sources. The harms and concerns tackled by these initiatives is then discussed in detail. The issues raised can be broadly aligned with issues identified in Chapter 2 and can be split into questions around: human rights and well-being; emotional harm; accountability and responsibility; security, privacy, accessibility and transparency; safety and trust; social harm and social justice; lawfulness and justice; control and the ethical use (or misuse) of AI; environmental harm and sustainability; informed use; existential risk.

All initiatives focus on human rights and well-being, arguing that AI must not affect basic and fundamental human rights. The IEEE initiative further recommends governance frameworks, standards and regulatory bodies to oversee use of AI and ensure that human well-being is prioritised throughout the design phase. The Montreal Protocol argues that AI should encourage and support the growth and flourishing of human well-being.

Another prominent issue identified in these initiatives is concern about the impact of AI on the human emotional experience, including the ways in which AIs address cultural sensitivities (or fail to do so). Emotional harm is considered a particular risk in the case of intelligent robots with whom humans might form an intimate relationship. Emotional harm may also arise should AI be designed to emotionally manipulate users (though it is also recognised that such nudging can also have

positive impacts, e.g. on healthy eating). Several initiatives recognise that nudging requires particular ethical consideration.

The need for accountability is recognised by initiatives, the majority of which focus on the need for AI to be auditable as a means of ensuring that manufacturers, designers and owners/operators of AI can be held responsible for harm caused. This also raises the question of autonomy and what that means in the context of AI.

Within the initiatives there is a recognition that new standards are required that would detail measurable and testable levels of transparency so that systems can be objectively assessed for compliance. Particularly in situations where AI replaces human decision-making initiatives, we argue that AI must be safe, trustworthy, reliable and act with integrity. The IEEE focus on the need for researchers to operate with a 'safety mindset' to pre-empt unintended or unanticipated behaviours.

With regard to societal harms, the IEEE suggests that social and moral norms should be considered in design, while the Japanese Society for AI, suggests that AI should be designed with social responsibility in mind. Several initiatives focus on the need to consider social inclusion and diversity, and the risk that AI could widen gaps between developed and developing economies. There is concern that AI-related degree programmes fail to equip designers with appropriate knowledge of ethics.

Legal issues are also addressed in the initiatives, with the IEEE arguing that AI should not be granted the status of 'personhood' and that existing laws should be scrutinised to ensure that they do not practically give AI legal autonomy.

Concerns around environmental harms are evident across initiatives, including concerns about resource use but also acknowledgement that AI could play a role in conservation and sustainable stewardship. The UNI Global Union states that AI should put people and plants first, striving to protect and enhance biodiversity and ecosystems.

Throughout the initiatives, there is a recognition of the need for greater public engagement and education with regard to the potential harms of AI. The initiatives suggest a range of ways in which this could be achieved, as a way of raising a number of topics that should be addressed through such initiatives.

Autonomous weapons systems attract particular attention from initiatives, given their potential to seriously harm society.

Case studies in Chapter 3 cover the particular risks associated with healthcare robots, which may be involved in diagnosis, surgery and monitoring health and well-being as well as providing caring services. The first case study highlights particular risks associated with embodied AI, which have moving parts that can cause injury. Healthcare AI applications also have implications for training of healthcare professionals and present data protection, legal and equality challenges. The case study raises a number of ethical concerns in relation to the deployment of robots for the care of the elderly in particular. The use of AI in healthcare also raises questions about trust, for example, how trust in professionals might change if they are seen as 'users' of technology.

A second case study explores ethical issues associated with the development of autonomous vehicles (AVs). In the context of driving, six levels of automation are recognised by SAE International: no automation, hands on (e.g. Cruise Control), hands off (driver still monitors driving), eyes off (driver can turn attention elsewhere, but must be prepared to intervene), minds off (no driver attention required) and steering wheel optional (human intervention is not required). Public safety is a key

concern regarding the deployment of autonomous vehicles, particularly following high-profile deaths associated with the use such vehicles. Liability is also a key concern with this emerging technology and the lack of standards, processes and regulatory frameworks for accident investigation hampers efforts to investigate accidents. Furthermore, with the exception of the US state of California, manufacturers are not required to log near misses.

Manufacturers of autonomous vehicles also collect significant amounts of data from AVs, which raises questions about the privacy and data protection rights of drivers and passengers. AVs could change urban environments, with, for example, additional infrastructure needed (AV-only lanes), but also affecting traffic congestion and requiring the extension of 5G network coverage.

A final case study explores the use of AI in warfare and the potential for AI applications to be used as weapons. AI is already used in military contexts. However, there are particular aspects of developing AI technologies that warrant consideration. These include: lethal autonomous weapons; drone technologies; robotic assassination and mobile-robotic-improvised explosive devices.

Key ethical issues arising from greater military use of AI include questions about the involvement of human judgement (if human judgement is removed, could this violate International Humanitarian Law). Would increasing use of AI reduce the threshold for going to war (affecting global stability)?

Chapter 4 discusses emerging **AI ethics standards and regulations**. There are a number of emerging standards that address emerging ethical, legal and social impacts of robotics and AI. Perhaps the earliest of these is the BS 8611 Guide to the Ethical Design and Application of Robots and Robotic Systems. It is based on a set of 20 distinct ethical hazards and risks, grouped under four categories: societal, application, commercial & financial, and environmental. The standard recognises physical hazards as implying ethical hazards and recognises that both physical and emotional hazards should be balanced against expected benefits to the user.

National and International policy initiatives are addressed in Chapter 5: **National and International Strategies on AI**. Canada launched the first national strategy on AI in March 2017, followed soon after by Japan, with many initiatives published since (see Figure 5. 1), including national strategies for Denmark, Finland, France, Germany, Sweden and the UK. The EU Strategy was the first international initiative on AI and supports the strategies of individual Member States. Strategies vary however in the extent to which they address ethical issues. At the European level, public concerns feature prominently in AI initiatives. Other international AI initiatives that cover ethical principles include: G7 Common Vision for the Future of AI, Nordic-Baltic Region Declaration on AI, OECD Principles on AI and the World Economic Form's Global AI Council. The United Nations has several initiatives relating to AI, including the AI for Good Global Summit; UNICRI Centre for AI and Robotics; UNESCO Report on Robotics Ethics.

Finally, Chapter 6 draws together the **themes emerging** from the literature, ethical initiatives and national and international strategies in relation to AI, highlighting gaps. It questions whether the two current international frameworks (EU High Level Expert Group, 2018<sup>2</sup> and OECD principles for AI, 2019) for the governance of AI are sufficient to meet the challenges it poses. The analysis highlights gaps in relation to environmental concerns; human psychology; workforce, particularly in relation to inequality and bias; democracy and finance.

## Table of contents

Executive summary.....	1
1. Introduction .....	1
2. Mapping the main ethical dilemmas and moral questions associated with the deployment of AI	5
2.1. Impact on society.....	6
2.1.1. The labour market.....	6
2.1.2. Inequality.....	8
2.1.3. Privacy, human rights and dignity.....	12
2.1.4. Bias.....	15
2.1.5 Democracy .....	16
2.2 Impact on human psychology .....	18
2.2.1 Relationships.....	18
2.2.4 Personhood .....	20
2.3 Impact on the financial system .....	21
2.4 Impact on the legal system .....	22
2.4.1 Criminal law .....	22
2.4.2 Tort law.....	27
2.5 Impact on the environment and the planet .....	28
2.5.1 Use of natural resources.....	28
2.5.2 Pollution and waste .....	28
2.5.3 Energy concerns.....	28
2.5.4 Ways AI could help the planet .....	29
2.6 Impact on trust .....	29
2.6.1 Why trust is important .....	30
2.6.2 Fairness.....	30
2.6.3 Transparency.....	31
2.6.4 Accountability.....	34
2.6.5 Control.....	35
3. Ethical initiatives in the field of artificial intelligence.....	37
3.1. International ethical initiatives .....	37
3.2. Ethical harms and concerns tackled by these initiatives .....	42

---

3.2.1 Harms in detail.....	45
3.3. Case studies .....	53
3.3.1. Case study: healthcare robots.....	53
3.3.2 Case study: Autonomous Vehicles .....	59
3.3.3 Case study: Warfare and weaponisation.....	63
4. AI standards and regulation.....	66
5. National and International Strategies on AI .....	71
5.1. Europe.....	73
5.2. North America.....	76
5.3. Asia.....	77
5.4. Africa.....	78
5.5. South America.....	79
5.6. Australasia .....	79
5.7. International AI Initiatives, in addition to the EU .....	80
5.8. Government Readiness for AI.....	82
6. Emerging Themes .....	84
6.1. Addressing ethical issues through national and international strategies.....	84
6.2. Addressing the governance challenges posed by AI.....	85
7. Summary .....	88
8. Appendix.....	90
Building ethical robots.....	90

## Table of figures

Figure 1: Main ethical and moral issues associated with the development and implementation of AI	5
Figure 2: General principles for the ethical and values-based design, development, and implementation of autonomous and intelligent systems (as defined by the IEEE's <i>Ethically Aligned Design</i> First Edition March 2019)	44
Figure 3: National and International Strategies on AI published as of May 2019.	72

## Table of tables

Table 1: Ethical initiatives and harms addressed	38
Table 2: IEEE 'human standards' with implications for AI	68
Table 3: Top 10 rankings for Government AI Readiness 2018/19. Source: Oxford Insights, 2019.	83



## 1. Introduction

Rapid developments in artificial intelligence (AI) and machine learning carry huge potential benefits. However it is necessary to explore the full ethical, social and legal aspects of AI systems if we are to avoid unintended, negative consequences and risks arising from the implementation of AI in society.

This chapter introduces AI broadly, including current uses and definitions of intelligence. It also defines robots and their position within the broader AI field.

### 1.1. What is AI – and what is intelligence?

The European Commission's Communication on Artificial Intelligence (European Commission, 2018a) defines artificial intelligence as follows:

*'Artificial Intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.'*

*'AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).'*

Within this report, we consider both software-based AI and intelligent robots (i.e. robots with an embedded AI) when exploring ethical issues. Intelligent robots are therefore a subset of AI (whether or not they make use of machine learning).

**How do we define intelligence?** A straightforward definition is that intelligent behaviour is 'doing the right thing at the right time'. Legg and Hunt (2007) survey a wide range of informal definitions of intelligence, identifying three common features: that intelligence is (1) 'a property that an individual agent has as it interacts with its environment or environments', (2) 'related to the agent's ability to succeed or profit with respect to some goal or objective', and (3) 'depends on how able that agent is to adapt to different objectives and environments'. They point out that intelligence involves adaptation, learning and understanding. At its simplest, then, intelligence is 'the ability to acquire and apply knowledge and skills and to manipulate one's environment'.

In interpreting these definitions of intelligence, we need to understand that for a physical **robot** its environment is the real world, which can be a human environment (for social robots), a city street (for an autonomous vehicle), a care home or hospital (for a care or assisted living robot), or a workplace (for a workmate robot). The 'environment' of a software AI is its context, which might be clinical (for a medical diagnosis AI), or a public space – for face recognition in airports, for instance, or virtual for face recognition in social media. But, like physical robots, software AIs almost always interact with humans, whether via question and answer interfaces: via text for chatbots, or via speech for digital assistants on mobile phones (i.e. Siri) or in the home (i.e. Alexa).

It is this interaction with humans that gives rise to almost all of the ethical issues surveyed in this report.

All present-day AIs and robots are examples of what we refer to as '**narrow**' AI: a term that reflects the fact that current AIs and robots are typically only capable of undertaking one specialised task. A long-term goal of AI and robotics research is so-called **artificial general intelligence (AGI)** which

---

would be comparable to human intelligence.<sup>1</sup> It is important to understand that present-day narrow AI is often better than most humans at one particular task; examples are chess- or Go-playing AIs, search engines or natural language translation systems. But a general-purpose care robot capable of, for instance, preparing meals for an elderly person (and washing the dishes afterwards), helping them dress or undress, get into and out of bed or the bath etc., remains a distant research goal.

**Machine learning** is the term used for AIs which are capable of learning or, in the case of robots, adapting to their environment. There are a broad range of approaches to machine learning, but these typically fall into two categories: supervised and unsupervised learning. Supervised learning systems generally make use of **Artificial Neural Networks (ANNs)**, which are trained by presenting the ANN with inputs (for instance, images of animals) each of which is tagged (by humans) with an output (i.e. giraffe, lion, gorilla). This set of inputs and matched outputs is called a training data set. After training, an ANN should be able to identify which animal is in an image it is presented with (i.e. a lion), even though that particular image with a lion wasn't present in the training data set. In contrast, unsupervised learning has no training data; instead, the AI (or robot) must figure out on its own how to solve a particular task (i.e. how to navigate successfully out of a maze), generally by trial and error.

Both supervised and unsupervised learning have their limitations. With supervised learning, the training data set must be truly representative of the task required; if not, the AI will exhibit bias. Another limitation is that ANNs learn by picking out features of the images in the training data unanticipated by the human designers. So, for instance, they might wrongly identify a car against a snowy background as a wolf, because all examples of wolves in the images of the training data set had snowy backgrounds, and the ANN has learned to identify snowy backgrounds as wolves, rather than the wolf itself. Unsupervised learning is generally more robust than supervised learning but suffers the limitation that it is generally very slow (compared with humans who can often learn from as few as one trial).

The term **deep learning** simply refers to (typically) supervised machine learning systems with large (i.e. many-layered) ANNs and large training data sets.

It is important to note the terms AI and machine learning are not synonymous. Many highly capable AIs and robots do not make use of machine learning.

## 1.2. Definition of morality and ethics, and how that relates to AI

Ethics are moral principles that govern a person's behaviour or the conduct of an activity. As a practical example, one ethical principle is *to treat everyone with respect*. Philosophers have debated ethics for many centuries, and there are various well-known principles, perhaps one of the most famous being Kant's categorical imperative 'act as you would want all other people to act towards all other people'.<sup>2</sup>

AI ethics is concerned with the important question of how human developers, manufacturers and operators should behave in order to minimise the ethical harms that can arise from AI in society, either arising from poor (unethical) design, inappropriate application or misuse. The scope of AI ethics spans immediate, here-and-now concerns about, for instance, data privacy and bias in current AI systems; near- and medium-term concerns about, for instance, the impact of AI and robotics on

---

<sup>1</sup> AGI could be defined as technologies that are explicitly developed as systems that can learn incrementally, reason abstractly and act effectively over a wide range of domains—just like humans can.

<sup>2</sup> From Kant's 1785 book *Groundwork of the Metaphysics of Morals*, with a variety of translations from the original German.

---

jobs and the workplace; and longer-term concerns about the possibility of AI systems reaching or exceeding human-equivalent capabilities (so-called superintelligence).

Within the last 5 years AI ethics has shifted from an academic concern to a matter for political as well as public debate. The increasing ubiquity of smart phones and the AI-driven applications that many of us now rely on every day, the fact that AI is increasingly impacting all sectors (including industry, healthcare, policing & the judiciary, transport, finance and leisure), as well as the seeming prospect of an AI 'arms race', has prompted an extraordinary number of national and international initiatives, from NGOs, academic and industrial groupings, professional bodies and governments. These initiatives have led to the publication of a large number of sets of ethical principles for robotics and AI (at least 22 different sets of ethical principles have been published since January 2017), new ethical standards are emerging (notably from the British Standards Institute and the IEEE Standards Association), and a growing number of countries (and groups of countries) have announced AI strategies (with large-scale investments) and set up national advisory or policy bodies.

In this report we survey these initiatives in order to draw out the main ethical issues in AI and robotics.

### 1.3. Report structure

Robots and artificial intelligence (AI) come in various forms, as outlined above, each of which raises a different **range of ethical concerns**. These are outlined in Chapter 2: Mapping the main ethical dilemmas and moral questions associated with the deployment of AI. This chapter explores in particular:

**Social impacts:** this section considers the potential impact of AI on the labour market and economy and how different demographic groups might be affected. It addresses questions of inequality and the risk that AI will further concentrate power and wealth in the hands of the few. Issues related to privacy, human rights and dignity are addressed as are risks that AI will perpetuate the biases, intended or otherwise, of existing social systems or their creators. This section also raises questions about the impact of AI technologies on democracy, suggesting that these technologies may operate for the benefit of state-controlled economies.

**Psychological impacts:** what impacts might arise from human-robot relationships? How might we address dependency and deception? Should we consider whether robots deserve to be given the status of 'personhood' and what are the legal and moral implications of doing so?

**Financial system impacts:** potential impacts of AI on financial systems are considered, including risks of manipulation and collusion and the need to build in accountability.

**Legal system impacts:** there are a number of ways in which AI could affect the legal system, including: questions relating to crime, such as liability if an AI is used for criminal activities, and the extent to which AI might support criminal activities such as drug trafficking. In situations where an AI is involved in personal injury, such as in a collision involving an autonomous vehicle, then questions arise around the legal approach to claims (whether it is a case of negligence, which is usually the basis for claims involving vehicular accidents, or product liability).

**Environmental impacts:** increasing use of AIs comes with increased use of natural resources, increased energy demands and waste disposal issues. However, AIs could improve the way we manage waste and resources, leading to environmental benefits.

**Impacts on trust:** society relies on trust. For AI to take on tasks, such as surgery, the public will need to trust the technology. Trust includes aspects such as fairness (that AI will be impartial), transparency (that we will be able to understand how an AI arrived at a particular decision),

accountability (someone can be held accountable for mistakes made by AI) and control (how we might 'shut down' an AI that becomes too powerful).

In Chapter 3, **Ethical initiatives in the field of artificial intelligence**, the report reviews a wide range of ethical initiatives that have sprung up in response to the ethical concerns and issues emerging in relation to AI. **Section 3.1** discusses the issues each initiative is exploring and identifies reports available (as of May 2019).

**Ethical harms and concerns tackled by the initiatives** outlined above, are discussed in Section 3.2. These are broadly split into 12 categories: human rights and well-being; emotional harm; accountability and responsibility; security, privacy, accessibility, and transparency; safety and trust; social harm and social justice; financial harm; lawfulness and justice; control and the ethical use (or misuse) of AI; environmental harm and sustainability; informed use and existential risks. The chapter explores each of these topics and the ways in which they are being addressed by the initiatives.

Chapter 4 presents the current status of **AI Ethical standards and regulation**. At present only one standard (British Standard BS8611, *Guide to the ethical design of robots and robotic systems*) specifically addresses AI. However, the IEEE is developing a number of standards that affect AI in a range of contexts. While these are in development, they are presented here as an indication of where standards and regulation is progressing.

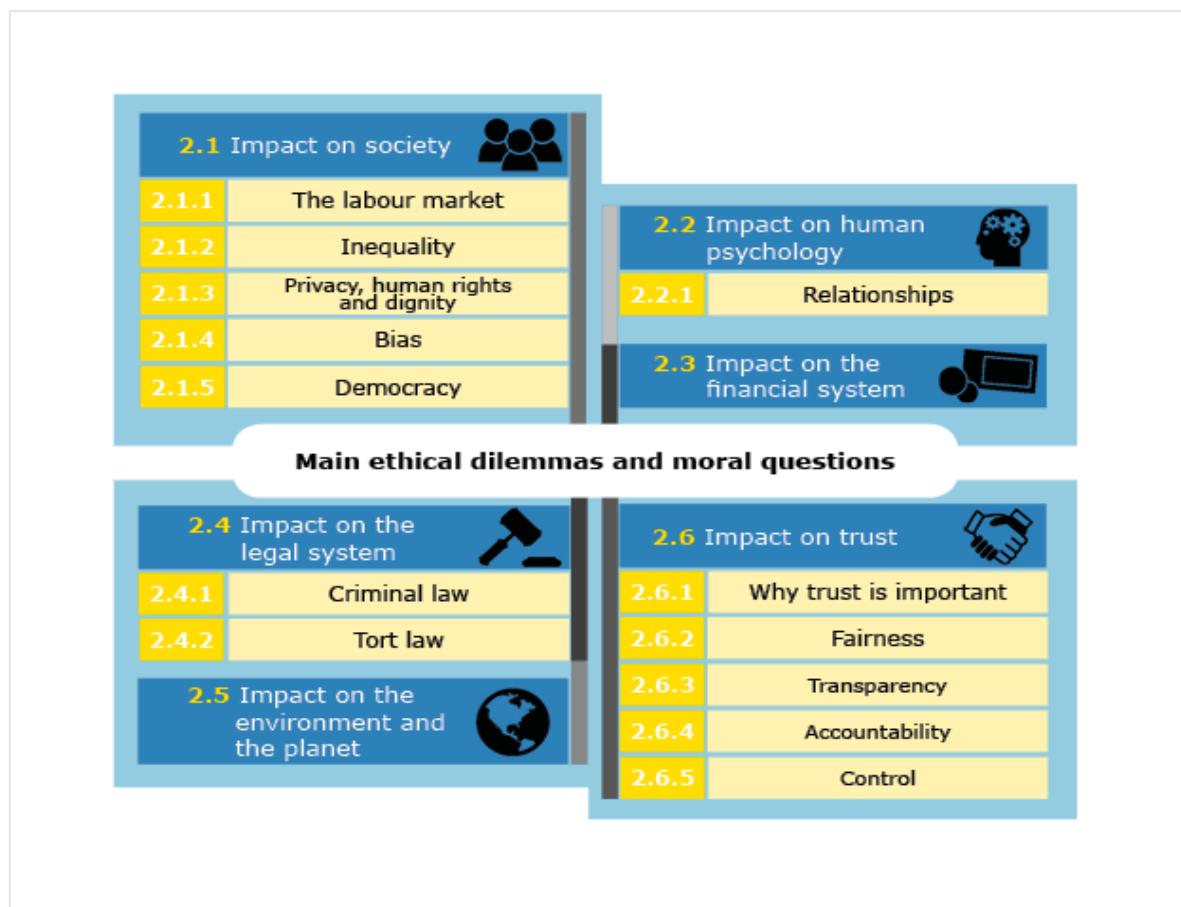
Finally, Chapter 5 explores **National and international strategies on AI**. The chapter considers what is required for a trustworthy AI and visions for the future of AI as they are articulated in national and international strategies.

## 2. Mapping the main ethical dilemmas and moral questions associated with the deployment of AI

According to the Future of Life Institute (n.d.), AI 'holds great economic, social, medical, security, and environmental promise', with potential benefits including:

- Helping people to acquire new skills and training;
- Democratising services;
- Designing and delivering faster production times and quicker iteration cycles;
- Reducing energy usage;
- Providing real-time environmental monitoring for air pollution and quality;
- Enhancing cybersecurity defences;
- Boosting national output;
- Reducing healthcare inefficiencies;
- Creating new kinds of enjoyable experiences and interactions for people; and
- Improving real-time translation services to connect people across the globe.

Figure 1: Main ethical and moral issues associated with the development and implementation of AI



In the long term, AI may lead to 'breakthroughs' in numerous fields, says the Institute, from basic and applied science to medicine and advanced systems. However, as well as great promise, increasingly capable intelligent systems create significant ethical challenges (Winfield, 2019a). This section of the report summarises the main ethical, social and legal considerations in the deployment

of AI, drawing insights from relevant academic literature. The issues discussed deal with impacts on: human society; human psychology; the financial system; the legal system; the environment and the planet; and impacts on trust.

## 2.1. Impact on society

### 2.1.1. The labour market

People have been concerned about the displacement of workers by technology for centuries. Automation, and then mechanisation, computing, and more recently AI and robotics have been predicted to destroy jobs and create irreversible damage to the labour market. Leontief (1983), observing the dramatic improvements in the processing power of computer chips, worried that people would be replaced by machines, just as horses were made obsolete by the invention of internal combustion engines. In the past, however, automation has often substituted for human labour in the short term, but has led to the creation of jobs in the long term (Autor, 2015).

Nevertheless, there is widespread concern that artificial intelligence and associated technologies could create mass unemployment during the next two decades. One recent paper concluded that new information technologies will put 'a substantial share of employment, across a wide range of occupations, at risk in the near future' (Frey and Osborne, 2013).

AI is already widespread in finance, space exploration, advanced manufacturing, transportation, energy development and healthcare. Unmanned vehicles and autonomous drones are also performing functions that previously required human intervention. We have already seen the impact of automation on 'blue-collar' jobs; however, as computers become more sophisticated, creative, and versatile, more jobs will be affected by technology and more positions made obsolete.

### Impact on economic growth and productivity

Economists are generally enthusiastic about the prospects of AI on economic growth. Robotics added an estimated 0.4 percentage points of annual GDP growth and labour productivity for 17 countries between 1993 and 2007, which is of a similar magnitude to the impact of the introduction of steam engines on growth in the United Kingdom (Graetz and Michaels, 2015).

### Impact on the workforce

It is hard to quantify the effect that robots, AI and sensors will have on the workforce because we are in the early stages of the technology revolution. Economists also disagree on the relative impact of AI and robotics. One study asked 1,896 experts about the impact of emerging technologies; 48 percent believed that robots and digital agents would displace significant numbers of both 'blue' and 'white' collar workers, with many expressing concern that this would lead to vast increases in income inequality, large numbers of unemployable people, and breakdowns in the social order (Smith and Anderson, 2014). However, the other half of the experts who responded to this survey (52%) expected that technology would *not* displace more jobs than it created by 2025. Those experts believed that although many jobs currently performed by humans will be substantially taken over by robots or digital agents, they have faith that human ingenuity will create new jobs, industries, and ways to make a living.

Some argue that technology is already producing major changes in the workforce:

*'Technological progress is going to leave behind some people, perhaps even a lot of people, as it races ahead... there's never been a better time to be a worker with special skills or the right education because these people can use technology to create and capture value. However, there's never been a worse time to be a worker with only 'ordinary' skills and abilities to offer, because computers, robots, and other digital technologies are acquiring these skills and abilities at an extraordinary rate' (Brynjolfsson and McAfee, 2014).*

Ford (2009) issues an equally strong warning, and argues that:

*'as technology accelerates, machine automation may ultimately penetrate the economy to the extent that wages no longer provide the bulk of consumers with adequate discretionary income and confidence in the future. If this issue is not addressed, the result will be a downward economic spiral'. He warns that 'at some point in the future — it might be many years or decades from now — machines will be able to do the jobs of a large percentage of the 'average' people in our population, and these people will not be able to find new jobs'.*

However, some economists dispute these claims, saying that although many jobs will be lost through technological improvements, new ones will be created. According to these individuals, the job gains and losses will even out over the long run.

*'There may be fewer people sorting items in a warehouse because machines can do that better than humans. But jobs analysing big data, mining information, and managing data sharing networks will be created'* (West, 2018).

If AI led to economic growth, it could create demand for jobs throughout the economy, including in ways that are not directly linked to technology. For example, the share of workers in leisure and hospitality sectors could increase if household incomes rose, enabling people to afford more meals out and travel (Furman and Seamans, 2018).

Regardless, it is clear that a range of sectors will be affected. Frey and Osborne (2013) calculate that there is a high probability that 47 percent of U.S. workers will see their jobs become automated over the next 20 years. According to their analysis, telemarketers, title examiners, hand sewers, mathematical technicians, insurance underwriters, watch repairers, cargo agents, tax preparers, photographic process workers, new accounts clerks, library technicians, and data-entry specialists have a 99 percent chance of having their jobs computerised. At the other end of the spectrum, recreational therapists, mechanic supervisors, emergency management directors, mental health social workers, audiologists, occupational therapists, health care social workers, oral surgeons, firefighter supervisors and dieticians have less than a one percent chance of this.

In a further study, the team surveyed 156 academic and industry experts in machine learning, robotics and intelligent systems, and asked them what tasks they believed could currently be automated (Duckworth et al., 2019). They found that work that is clerical, repetitive, precise, and perceptual can increasingly be automated, while work that is more creative, dynamic, and human oriented tends to be less 'automatable'.

Worryingly, eight times as much work fell between 'mostly' and 'completely' automatable than between 'mostly not' and 'not at all' automatable, when weighted by employment. Activities classified as 'reasoning and decision making' and 'coordinating, developing, managing, and advising' were less likely than others to be automatable, while 'administering', 'information and data processing' and 'performing complex and technical activities' were likely to be more so.

Overall the model predicted very high automation potential for office, administrative support, and sales occupations, which together employ about 38 million people in the U.S. Also at high risk of automation were physical processes such as production, farming, fishing and forestry, and transportation and material moving, which employ about 20 million people in total. In contrast, occupations that were robust to automation included education, legal, community service, arts, and media occupations, and to a lesser extent, management, business, and financial occupations.

Unsurprisingly, the study found that occupations with the highest salaries and levels of education tend to be the least amenable to automation. However, even this does not guarantee that an occupation's activities cannot be automated. As the authors point out, air traffic controllers earn

about US\$125,000 a year, but it is thought that their tasks could largely be automated. In contrast, preschool teachers and teaching assistants earn under \$30,000 a year, yet their roles are not thought to be amenable to automation.

### Labour-market discrimination: effects on different demographics

The impacts of these sizeable changes will not be felt equally by all members of society. Different demographics will be affected to varying extents, and some are more at risk than others from emerging technologies. Those with few technical skills or specialty trades will face the most difficulties (UK Commission for Employment and Skills, 2014). Young people entering the labour market will also be disproportionately affected, since they are at the beginning of their careers and they will be the first generation to work alongside AI (Biavaschi et al., 2013). Even though many young people have time to acquire relevant expertise, few gain training in science, technology, engineering, and math (STEM) fields, limiting their ability to withstand employment alterations. According to the U.S. Department of Education (2014), there will be a 14 percent increase in STEM jobs between 2010 and 2020 — but 'only 16 percent of American high school seniors are proficient in mathematics and interested in a STEM career'.

Women may also be disproportionately affected, as more women work in caregiving positions — one of the sectors likely to be affected by robots. Due to discrimination, prejudice and lack of training, minorities and poor people already suffer high levels of unemployment: without high-skill training, it will be more difficult for them to adapt to a new economy. Many of these individuals also lack access to high-speed Internet, which limits their ability to access education, training and employment (Robinson et al., 2015).

Special Eurobarometer survey 460 identified that EU residents have a largely positive response to the increasing use of digital technology, considering it to improve society, the economy, and their quality of life, and that most also consider themselves competent enough to make use of this technology in various aspects of their life and work (European Commission, 2017). However, crucially, this attitude varied by age, location, and educational background — a finding that is central to the issue of how AI will affect different demographics and the potential issues arising around the 'digital divide'.

For instance, young men with high levels of education are the most likely to hold positive views about digitisation and the use of robots — and are also the most likely to have taken some form of protective measure relating to their online privacy and security (thus placing them at lower risk in this area). These kinds of socio-demographic patterns highlight a key area of concern in the increasing development and implementation of AI if nobody is to be disadvantaged or left behind (European Commission, 2017).

### Consequences

*'When we're talking about 'AI for good', we need to define what 'good' means. Currently, the key performance indicators we look to are framed around GDP. Not to say it's evil, but it's about measuring productivity and exponential profits'. (John Havens)*

It is possible that AI and robotic technologies could exacerbate existing social and economic divisions, via putting current job classes at risk, eliminating jobs, causing mass unemployment in automatable job sectors. Discrimination may also be an issue, with young people potentially being disproportionately affected, alongside those without high-skill training.

#### 2.1.2. Inequality

*'The biggest question around AI is inequality, which isn't normally included in the debate about AI ethics. It is an ethical issue, but it's mostly an issue of politics – who benefits from AI?' (Jack Stilgoe)*

AI and robotics technology are expected to allow companies to streamline their businesses, making them more efficient and more productive. However, some argue that this will come at the expense of their human workforces. This will inevitably mean that revenues will be split across fewer people, increasing social inequalities. Consequently, individuals who hold ownership in AI-driven companies are set to benefit disproportionately.

### Inequality: exploitation of workers

Changes in employment related to automation and digitisation will not be expressed solely via job losses, as AI is expected to create many numerous and new forms of employment (Hawksworth and Fertig, 2018), but also in terms of job quality. Winfield (2019b) states that new jobs may require highly skilled workers but be repetitive and dull, creating 'white-collar sweatshops' filled with workers performing tasks such as tagging and moderating content – in this way, AI could bring an additional human cost that must be considered when characterising the benefits of AI to society. Building AI most often requires people to manage and clean up data to instruct the training algorithms. Better (and safer) AI needs huge training data sets and a whole new outsourced industry has sprung up all over the world to meet this need. This has created several new categories of job.

These include: (i) scanning and identifying offensive content for deletion, (ii) manually tagging objects in images in order to create training data sets for machine learning systems (for example, to generate training data sets for driverless car AIs) and (iii) interpreting queries (text or speech) that an AI chatbot cannot understand. Collectively these jobs are sometimes known by the term 'mechanical turk' (so named after the 18<sup>th</sup> century chess playing automaton that was revealed to be operated by a human chess master hidden inside the cabinet).

When first launched such tasks were offered as a way for people to earn extra money in their spare time, however Gray and Suri (2019) suggest that 20 million individuals are now employed worldwide, via third party contractors, in an on-demand 'gig economy', working outside the protection of labour laws. The jobs are usually scheduled, routed, delivered and paid for online, through application programming interfaces (APIs). There have been a few journalistic investigations into the workers in this field of work<sup>3</sup> – termed 'ghost work' by Harvard researcher Mary L. Gray because of the 'hidden' nature of the value chain providing the processing power on which AI is based (Gray, 2019).

The average consumer of AI technology may never know that a person was part of the process – the value chain is opaque. One of the key ethical issues is that – given the price of the end-products – these temporary workers are being inequitably reimbursed for work that is essential to the functioning of the AI technologies. This may be especially the case where the labour force reside in countries outside the EU or US – there are growing 'data-labelling' industries in both China and Kenya, for example. Another issue is with the workers required to watch and vet offensive content for media platforms such as Facebook and YouTube (Roberts, 2016). Such content can include hate speech, violent pornography, cruelty and sometimes murder of both animals and humans. A news report (Chen, 2017) outlines mental health issues (PTSD-like trauma symptoms, panic attacks and burnout), alongside poor working conditions and ineffective counselling.

This hidden army of piecemeal workers are undertaking work that is at best extremely tedious and poorly paid, at worst, precarious, unhealthy and/or psychologically harmful. Gray's research makes the case that workers in this field still display the desire to invest in work as something more than a single payment transaction, and advises that the economic, social and psychological impacts of 'ghost work' should be dealt with systematically. Making the worker's inputs more transparent in the end-product, ensuring the value chain improves the equitable distribution of benefits, and

---

<sup>3</sup> The Verge: <https://www.theverge.com/2019/5/13/18563284/mary-gray-ghost-work-microwork-labor-silicon-valley-automation-employment-interview>;

ensuring appropriate support structures for those humans-in-the-loop who deal with psychologically harmful content are all important steps to address the ethical issues.

### Sharing the benefits

AI has the potential to bring significant and diverse benefits to society (Conn, 2018; UK Government Office for Science, 2015; The Future of Life Institute, n.d.; The White House, 2016) and facilitate, among other things, greater efficiency and productivity at lower cost (OECD, n.d.). The Future of Life Institute (n.d.) states that AI may be capable of tackling a number of the most difficult global issues – poverty, disease, conflict – and thus improve countless lives.

A US report on AI, automation, and the economy (2016) highlights the importance of ensuring that potential benefits of AI do not accumulate unequally, and are made accessible to as many people as possible. Rather than framing the development of AI and automation as leading to an inevitable outcome determined by the technology itself, the report states that innovation and technological change 'does not happen in a vacuum': the future of AI may be shaped not by technological capability, but by a wide range of non-technical incentives (The White House, 2016). Furthermore, the inventor or developer of an AI has great potential to determine its use and reach (Conn, 2018), suggesting a need for inventors to consider the wider potential impacts of their creations.

Automation is more applicable to certain roles than others (Duckworth et al., 2018), placing certain workers at a disadvantage and potentially increasing wage inequality (Acemoglu and Restrepo, 2018). Businesses may be motivated by profitability (Min, 2018) – but, while this may benefit business owner(s) and stakeholders, it may not benefit workers.

Brundage and Bryson (2016) mention the case study of electricity, which they say is sometimes considered analogous to AI. While electricity can make many areas more productive, remove barriers, and bring benefits and opportunity to countless lives, it has taken many decades for electricity to reach some markets, and 'indeed, over a billion [people] still lack access to it'.

To ensure that AI's benefits are distributed fairly – and to avoid a whoever designs it first, wins dynamic – one option may be to pre-emptively declare that AI is not a private good but instead for the benefit of all, suggests Conn (2018). Such an approach would require a change in cultural norms and policy. New national and governmental guidelines could underpin new strategies to harness the beneficial powers of AI for citizens, help navigate the AI-driven economic transition, and retain and strengthen public trust in AI (Min, 2018). Brundage and Bryson (2016) agree with this call for policy and regulation, stating that 'it is not sufficient to fund basic research and expect it to be widely and equitably diffused in society by private actors'. However, such future scenarios are not predetermined, says Servoz (2019), and will be shaped by present-day policies and choices.

The Future of Life Institute (n.d.) lists a number of policy recommendations to tackle the possible 'economic impacts, labour shifts, inequality, technological unemployment', and social and political tensions that may accompany AI. AI-driven job losses will require new retraining programmes and social and financial support for displaced workers; such issues may require economic policies such as universal basic income and robot taxation schemes. The Institute suggests that policies should focus on those most at risk of being left behind – caregivers, women and girls, underrepresented populations and the vulnerable – and on those building AI systems, to target any 'skewed product design, blind spots, false assumptions [and] value systems and goals encoded into machines' (The Future of Life Institute, n.d.).

According to Brundage and Bryson (2016), taking a proactive approach to AI policies is not 'premature, misguided [or] dangerous', given that AI 'is already sufficiently mature technologically to impact billions of lives trillions of times a day'. They suggest that governments seek to improve

their related knowledge and rely more on experts; that relevant research is allocated more funding; that policymakers plan for the future, seeking 'robustness and preparedness in the face of uncertainty'; and that AI is widely applied and proactively made accessible (especially in areas of great social value, such as poverty, illness, or clean energy).

Considering the energy industry as an example, AI may be able to modernise the energy grid, improve its reliability, and prevent blackouts by regulating supply and demand at both local and national levels, says Wolfe (2017). Such a 'smart grid' would save energy companies money but also allow consumers to actively monitor their own energy use in real-time and see cost savings, passing the benefits from developer to producer to consumer – and opening up new ways to save, earn, and interact with the energy grid (Gagan, 2018; Jacobs, 2017). Jacobs (2017) discusses the potential for 'prosumers' (those who both produce and consume energy, interacting with the grid in a new way) to help decentralise energy production and be a 'positive disruptive force' in the electricity industry – if energy strategy is regulated effectively via updated policy and management. Giving consumers real-time, accessible data would also help them to select the most cost-efficient tariff for them, say Ramchurn et al. (2013), given that accurately estimating one's yearly consumption and deciphering complex tariffs is a key challenge facing energy consumers. This may therefore have some potential to alleviate energy poverty, given that energy price increases and dependence on a centralised energy supply grid can leave households in fuel poverty (Ramchurn et al., 2013).

### Concentration of power among elites

*'Does AI have to increase inequality? Could you design systems that target, for example, the needs of the poorest people? If AI was being used to further benefit rich people more than it benefits poor people, which it looks likely to be, or more troublingly, put undue pressure on already particularly marginalised people, then what might we do about that? Is that an appropriate use of AI?' (Jack Stilgoe)*

Nemitz (2018) writes that it would be 'naive' to ignore that AI will concentrate power in the hands of a few digital internet giants, as 'the reality of how [most societies] use the Internet and what the Internet delivers to them is shaped by a few mega corporations...the development of AI is dominated exactly by these mega corporations and their dependent ecosystems'.

The accumulation of technological, economic and political power in the hands of the top five players – Google, Facebook, Microsoft, Apple and Amazon – affords them undue influence in areas of society relevant to opinion-building in democracies: governments, legislators, civil society, political parties, schools and education, journalism and journalism education and — most importantly — science and research.

In particular, Nemitz is concerned that investigations into the impact of new technologies like AI on human rights, democracy and the rule of law may be hampered by the power of tech corporations, who are not only shaping the development and deployment of AI, but also the debate on its regulation. Nemitz identifies several areas in which tech giants exert power:

1. **Financial.** Not only can the top five players afford to invest heavily in political and societal influence, they can also afford to buy new ideas and start-ups in the area of AI, or indeed any other area of interest to their business model — something they are indeed doing.
2. **Public discourse.** Tech corporations control the infrastructures through which public discourse takes place. Sites like Facebook and Google increasingly become the main, or even only, source of political information for citizens, especially the younger generation, to the detriment of the fourth estate. The vast majority of advertising revenue now also goes to Google and Facebook, removing the main income of newspapers and rendering investigative journalism unaffordable.

3. **Collecting personal data.** These corporations collect personal data for profit, and profile people based on their behaviour (both online and offline). They know more about us than ourselves or our friends — and they are using and making available this information for profit, surveillance, security and election campaigns.

Overall, Nemitz concludes that

*'this accumulation of power in the hands of a few — the power of money, the power over infrastructures for democracy and discourse, the power over individuals based on profiling and the dominance in AI innovation...must be seen together, and...must inform the present debate about ethics and law for AI'.*

Bryson (2019), meanwhile, believes this concentration of power could be an inevitable consequence of the falling costs of robotic technology. High costs can maintain diversity in economic systems. For example, when transport costs are high, one may choose to use a local shop rather than find the global best provider for a particular good. Lower costs allow relatively few companies to dominate, and where a few providers receive all the business, they will also receive all of the wealth.

### Political instability

Bryson (2019) also notes that the rise of AI could lead to wealth inequality and political upheaval. Inequality is highly correlated with political polarisation (McCarty et al., 2016), and one possible consequence of polarisation is an increase in identity politics, where beliefs are used to signal in-group status or affiliation (Iyengar et al., 2012; Newman et al., 2014). This could unfortunately result in situations where beliefs are more tied to a person's group affiliation than to objective facts, and where faith in experts is lost.

*'While occasionally motivated by the irresponsible use or even abuse of position by some experts, in general losing access to experts' views is a disaster. No one, however intelligent, can master in their lifetime all human knowledge. If society ignores the stores of expertise it has built up — often through taxpayer-funding of higher education — it sets itself at a considerable disadvantage' (Bryson, 2019).*

### 2.1.3. Privacy, human rights and dignity

AI will have profound impacts on privacy in the next decade. The privacy and dignity of AI users must be carefully considered when designing service, care and companion robots, as working in people's homes means they will be privy to intensely private moments (such as bathing and dressing). However, other aspects of AI will also affect privacy. Smith (2018), President of Microsoft, recently remarked:

*'[Intelligent 3] technology raises issues that go to the heart of fundamental human rights protections like privacy and freedom of expression. These issues heighten responsibility for tech companies that create these products. In our view, they also call for thoughtful government regulation and for the development of norms around acceptable uses.'*

### Privacy and data rights

*'Humans will not have agency and control [over their data] in any way if they are not given the tools to make it happen'. (John Havens)*

One way in which AI is already affecting privacy is via Intelligent Personal Assistants (IPA) such as Amazon's Echo, Google's Home and Apple's Siri. These voice activated devices are capable of

learning the interests and behaviour of their users, but concerns have been raised about the fact that they are always on and listening in the background.

A survey of IPA customers showed that people's biggest privacy concern was their device being hacked (68.63%), followed by it collecting personal information on them (16%), listening to their conversations 24/7 (10%), recording private conversations (12%), not respecting their privacy (6%), storing their data (6%) and the 'creepy' nature of the device (4%) (Manikonda et al, 2018). However despite these concerns, people were very positive about the devices, and comfortable using them.

Another aspect of AI that affects privacy is Big Data. Technology is now at the stage where long-term records can be kept on anyone who produces storable data — anyone with bills, contracts, digital devices, or a credit history, not to mention any public writing and social media use. Digital records can be searched using algorithms for pattern recognition, meaning that we have lost the default assumption of anonymity by obscurity (Selinger and Hartzog, 2017).

Any one of us can be identified by facial recognition software or data mining of our shopping or social media habits (Pasquale, 2015). These online habits may indicate not just our identity, but our political or economic predispositions, and what strategies might be effective for changing these (Cadwalladr, 2017a,b).

Machine learning allows us to extract information from data and discover new patterns, and is able to turn seemingly innocuous data into sensitive, personal data. For example, patterns of social media use can predict personality categories, political preferences, and even life outcomes (Youyou et al., 2015). Word choice, or even handwriting pressure on a digital stylus, can indicate emotional state, including whether someone is lying (Hancock et al., 2007; Bandyopadhyay and Hazra, 2017). This has significant repercussions for privacy and anonymity, both online and offline.

AI applications based on machine learning need access to large amounts of data, but data subjects have limited rights over how their data are used (Veale et al., 2018). Recently, the EU adopted new General Data Protection Regulations (GDPR) to protect citizen privacy. However, the regulations only apply to personal data, and not the aggregated 'anonymous' data that are usually used to train models.

In addition, personal data, or information about who was in the training set, can in certain cases be reconstructed from a model, with potentially significant consequences for the regulation of these systems. For instance, while people have rights about how their personal data are used and stored, they have limited rights over trained models. Instead, models have been typically thought to be primarily governed by varying intellectual property rights, such as trade secrets. For instance, as it stands, there are no data protection rights nor obligations concerning models in the period after they have been built, but before any decisions have been taken about using them.

This brings up a number of ethical issues. What level of control will subjects have over the data that are collected about them? Should individuals have a right to use the model, or at least to know what it is used for, given their stake in training it? Could machine learning systems seeking patterns in data inadvertently violate people's privacy if, for example, sequencing the genome of one family member revealed health information about other members of the family?

Another ethical issue surrounds how to prevent the identity, or personal information, of an individual involved in training a model from being discovered (for example through a cyber-attack). Veale et al. (2018) argue that extra protections should be given to people whose data have been used to train models, such as the right to access models; to know where they have originated from, and to whom they are being traded or transmitted; the right to erase themselves from a trained model; and the right to express a wish that the model not be used in the future.

## Human rights

AI has important repercussions for democracy, and people's right to a private life and dignity. For instance, if AI can be used to determine people's political beliefs, then individuals in our society might become susceptible to manipulation. Political strategists could use this information to identify which voters are likely to be persuaded to change party affiliation, or to increase or decrease their probability of turning out to vote, and then to apply resources to persuade them to do so. Such a strategy has been alleged to have significantly affected the outcomes of recent elections in the UK and USA (Cadwalladr, 2017a; b).

Alternatively, if AI can judge people's emotional states and gauge when they are lying, these people could face persecution by those who do not approve of their beliefs, from bullying by individuals through to missed career opportunities. In some societies, it could lead to imprisonment or even death at the hands of the state.

## Surveillance

*'Networks of interconnected cameras provide constant surveillance over many metropolitan cities. In the near future, vision-based drones, robots and wearable cameras may expand this surveillance to rural locations and one's own home, places of worship, and even locations where privacy is considered sacrosanct, such as bathrooms and changing rooms. As the applications of robots and wearable cameras expand into our homes and begin to capture and record all aspects of daily living, we begin to approach a world in which all, even bystanders, are being constantly observed by various cameras wherever they go' (Wagner, 2018).*

This might sound like a nightmare dystopian vision, but the use of AI to spy is increasing. For example, an Ohio judge recently ruled that data collected by a man's pacemaker could be used as evidence that he committed arson (Moon, 2017). Data collected by an Amazon Alexa device was also used as evidence (Sauer, 2017). Hundreds of connected home devices, including appliances and televisions, now regularly collect data that may be used as evidence or accessed by hackers. Video can be used for a variety of exceedingly intrusive purposes, such as detecting or characterising a person's emotions.

AI may also be used to monitor and predict potential troublemakers. Face recognition capacities are alleged to be used in China, not only to identify individuals, but to identify their moods and states of attention both in re-education camps and ordinary schools (Bryson, 2019). It is possible, such technology could be used to penalise students for not paying attention or penalise prisoners who do not appear happy to comply with their (re)education.

Unfortunately, governments do not always have their citizens' interests at heart. The Chinese government has already used surveillance systems to place over a million of its citizens in re-education camps for the crime of expressing their Muslim identity (Human Rights Watch, 2018). There is a risk that governments fearing dissent will use AI to suppress, imprison and harm individuals.

Law enforcement agencies in India already use 'proprietary, advance hybrid AI technology' to digitise criminal records, and use facial recognition to predict and recognise criminal activity (Marda, 2018; Sathe, 2018). There are also plans to train drones to identify violent behaviour in public spaces, and to test these drones at music festivals in India (Vincent, 2018). Most of these programmes intend to reduce crime rates, manage crowded public spaces to improve safety, and bring efficiency to law enforcement. However, they have clear privacy and human rights implications, as one's appearance and public behaviour is monitored, collected, stored and possibly shared without consent. Not only does the AI discussed operate in the absence of safeguards to prevent misuse, making them ripe for surveillance and privacy violations, they also operate at questionable levels of accuracy. This could

lead to false arrests and people from disproportionately vulnerable and marginalised communities being made to prove their innocence.

### Freedom of speech

Freedom of speech and expression is a fundamental right in democratic societies. This could be profoundly affected by AI. AI has been widely touted by technology companies as a solution to problems such as hate speech, violent extremism and digital misinformation (Li and Williams, 2018). In India, sentiment analysis tools are increasingly deployed to gauge the tone and nature of speech online, and are often trained to carry out automated content removal (Marda, 2018). The Indian Government has also expressed interest in using AI to identify fake news and boost India's image on social media (Seth 2017). This is a dangerous trend, given the limited competence of machine learning to understand tone and context. Automated content removal risks censorship of legitimate speech; this risk is made more pronounced by the fact that it is performed by private companies, sometimes acting on the instruction of government. Heavy surveillance affects freedom of expression, as it encourages self-censorship.

#### 2.1.4. Bias

AI is created by humans, which means it can be susceptible to bias. Systematic bias may arise as a result of the data used to train systems, or as a result of values held by system developers and users. It most frequently occurs when machine learning applications are trained on data that only reflect certain demographic groups, or which reflect societal biases. A number of cases have received attention for promoting unintended social bias, which has then been reproduced or automatically reinforced by AI systems.

##### *Examples of AI bias*

The investigative journalism organisation ProPublica showed that COMPAS, a machine learning based software deployed in the US to assess the probability of a criminal defendant re-offending, was strongly biased against black Americans. The COMPAS system was more likely to incorrectly predict that black defendants would reoffend, while simultaneously, and incorrectly, predicting the opposite in the case of white defendants (ProPublica, 2016).

Researchers have found that automated advertisement distribution tools are more likely to distribute adverts for well-paid jobs to men than women (Datta et al., 2015). AI-informed recruitment is susceptible to bias; an Amazon self-learning tool used to judge job-seekers was found to significantly favour men, ranking them highly (Dastin, 2018). The system had learned to prioritise applications that emphasised male characteristics, and to downgrade applications from universities with a strong female presence.

Many popular image databases contain images collected from just a few countries (USA, UK), which can lead to biases in search results. Such databases regularly portray women performing kitchen chores while men are out hunting (Zhao et al, 2017), for example, and searches for 'wedding gowns' produce the standard white version favoured in western societies, while Indian wedding gowns are categorised as 'performance art' or 'costumes' (Zhou 2018). When applications are programmed with this kind of bias, it can lead to situations such as a camera automatically warning a photographer that their subject has their eyes closed when taking a photo of an Asian person, as the camera has been trained on stereotypical, masculine and light-skinned appearances.

ImageNet, which has the goal of mapping out a world of objects, is a vast dataset of 14.1 million images organised into over 20,000 categories – the vast majority of which are plants, rocks, animals. Workers have sorted 50 images a minute into thousands of categories for ImageNet – at such a rate

there is large potential for inaccuracy. Problematic, inaccurate – and discriminatory – tagging (see Discrimination above) can be maintained in datasets over many iterations

There have been a few activities that have demonstrated the bias contained in data training sets. One is a facial recognition app (ImageNet Roulette)<sup>4</sup> which makes assumptions about you based entirely on uploaded photos of your face – everything from your age and gender to profession and even personal characteristics. It has been critiqued for its offensive, inaccurate and racist labelling – but the creators say that it is an interface that shows users how a machine learning model is interpreting the data and how results can be quite disturbing.<sup>5</sup>

#### *Implications*

As many machine-learning models are built from human-generated data, human biases can easily result in a skewed distribution in training data. Unless developers work to recognise and counteract these biases, AI applications and products may perpetuate unfairness and discrimination. AI that is biased against particular groups within society can have far-reaching effects. Its use in law enforcement or national security, for example, could result in some demographics being unfairly imprisoned or detained. Using AI to perform credit checks could result in some individuals being unfairly refused loans, making it difficult for them to escape a cycle of poverty (O'Neil 2016). If AI is used to screen people for job applications or university admissions it could result in entire sections of society being disadvantaged.

This problem is exacerbated by the fact that AI applications are usually 'black boxes', where it is impossible for the consumer to judge whether the data used to train them are fair or representative. This makes biases hard to detect and handle. Consequently, there has been much recent research on making machine learning fair, accountable and transparent, and more public-facing activities and demonstrations of this type would be beneficial.

### 2.1.5 Democracy

As already discussed, the concentration of technological, economic and political power among a few mega corporations could allow them undue influence over governments — but the adoption and implementation of AI could threaten democracy in other ways too.

#### *Fake news and social media*

Throughout history, political candidates campaigning for office have relied on limited anecdotal evidence and surveys to give them an insight into what voters are thinking. Now with the advent of Big Data, politicians have access to huge amounts of information that allow them to target specific categories of voters and develop messaging that will resonate with them most.

This may be a good thing for politicians, but there is a great deal of evidence that AI-powered technologies have been systematically misused to manipulate citizens in recent elections, damaging democracy. For example, 'bots' — autonomous accounts — were used to spread biased news and propaganda via Twitter in the run up to both the 2016 US presidential election and the Brexit vote in the United Kingdom (Pham, Gorodnichenko and Talavera, 2018). Some of these automated accounts were set up and operated from Russia and were, to an extent, able to bias the content viewed on social media, giving a false impression of support.

During the 2016 US presidential election, pro-Trump bots have been found to have infiltrated the online spaces used by pro-Clinton campaigners, where they spread highly automated content,

---

<sup>4</sup> Created by artist Trevor Paglen and Professor Kate Crawford and New York University.

<sup>5</sup> [https://www.vice.com/en\\_uk/article/xweagk/ai-face-app-imagenet-roulette](https://www.vice.com/en_uk/article/xweagk/ai-face-app-imagenet-roulette)

generating one-quarter of Twitter traffic about the 2016 election (Hess, 2016). Bots were also largely responsible for popularising #MacronLeaks on social media just days before the 2017 French presidential election (Polonski, 2017). They bombarded Facebook and Twitter with a mix of leaked information and falsified reports, building the narrative that Emmanuel Macron was a fraud and hypocrite.

A recent report found that at least 28 countries — including both authoritarian states and democracies — employ 'cyber troops' to manipulate public opinion over major social networking applications (Bradshaw and Howard, 2017). These cyber troops use a variety of tactics to sway public opinion, including verbally abusing and harassing other social media users who express criticism of the government. In Russia, cyber troops have been known to target journalists and political dissidents, and in Mexico, journalists are frequently targeted and harassed over social media by government-sponsored cyber troops (O'Carroll, 2017). Others use automated bots — according to Bradshaw and Howard (2017), bots have been deployed by government actors in Argentina, Azerbaijan, Iran, Mexico, the Philippines, Russia, Saudi Arabia, South Korea, Syria, Turkey and Venezuela. These bots are often used to flood social media networks with spam and 'fake' or biased news, and can also amplify marginal voices and ideas by inflating the number of likes, shares and retweets they receive, creating an artificial sense of popularity, momentum or relevance. According to the authors, authoritarian regimes are not the only or even the best at organised social media manipulation.

In addition to shaping online debate, AI can be used to target and manipulate individual voters. During the U.S. 2016 presidential election, the data science firm Cambridge Analytica gained access to the personal data of more than 50 million Facebook users, which they used to psychologically profile people in order to target adverts to voters they thought would be most receptive. There remains a general distrust of social media among members of the public across Europe, and its content is viewed with caution; a 2017 Eurobarometer survey found that just 7% of respondents deemed news stories published on online social platforms to be generally trustworthy (European Commission, 2017). However, a representative democracy depends on free and fair elections in which citizens can vote without manipulation — and AI threatens to undermine this process.

## News bubbles and echo chambers

The media increasingly use algorithmic news recommenders (ANR) to target customised news stories to people based on their interests (Thurman, 2011; Gillespie, 2014). However presenting readers with news stories based on their previous reading history lowers the chance of people encountering different and undiscovered content, opinions and viewpoints (Harambam et al., 2018). There is a danger this could result in increasing societal polarisation, with people essentially living in 'echo chambers' and 'filter bubbles' (Pariser, 2011) where they are only exposed to their own viewpoints. The interaction of different ideas and people is considered crucial to functioning democracies.

## The end of democracies

Some commentators have questioned whether democracies are particularly suited to the age of AI and machine learning, and whether its deployment will enable countries with other political systems to gain the advantage (Bartlett, 2018). For the past 200 years democracies have flourished because individual freedom is good for the economy. Freedom promotes innovation, boosting the economy and wealth, and creating well-off people who value freedom. However, what if that link was weakened? What if economic growth in the future no longer depended on individual freedom and entrepreneurial spirit?

A centrally planned, state-controlled economy may well be better suited to a new AI age, as it is less concerned with people's individual rights and privacy. For example, the size of the country's population means that Chinese businesses have access to huge amounts of data, with relatively few restraints on how those data can be used. In China, there are no privacy or data protection laws, such as the new GDPR rules in Europe. As China could soon become the world leader in AI, this means it could shape the future of the technology and the limits on how it is used.

'The last few years suggest digital technology thrives perfectly well under monopolistic conditions: the bigger a company is, the more data and computing power it gets, and the more efficient it becomes; the more efficient it becomes, the more data and computing power it gets, in a self-perpetuating loop' (Bartlett, 2018). According to Bartlett, people's love affair with 'convenience' means that if a 'machinocracy' was able to deliver wealth, prosperity and stability, many people would probably be perfectly happy with it.

## 2.2 Impact on human psychology

AI is getting better and better at modelling human thought, experience, action, conversation and relationships. In an age where we will frequently interact with machines as if they are humans, what will the impact be on real human relationships?

### 2.2.1 Relationships

Relationships with others form the core of human existence. In the future, robots are expected to serve humans in various social roles: nursing, housekeeping, caring for children and the elderly, teaching, and more. It is likely that robots will also be designed for the explicit purpose of sex and companionship. These robots may be designed to look and talk just like humans. People may start to form emotional attachments to robots, perhaps even feeling love for them. If this happens, how would it affect human relationships and the human psyche?

#### Human-robot relationships

*'The biggest risk [of AI] that anyone faces is the loss of ability to think for yourself. We're already seeing people are forgetting how to read maps, they're forgetting other skills. If we've lost the ability to be introspective, we've lost human agency and we're spinning around in circles'. (John Havens)*

One danger is that of **deception** and **manipulation**. Social robots that are loved and trusted could be misused to manipulate people (Scheutz 2012); for example, a hacker could take control of a personal robot and exploit its unique relationship with its owner to trick the owner into purchasing products. While humans are largely prevented from doing this by feelings like empathy and guilt, robots would have no concept of this.

Companies may design future robots in ways that enhance their trustworthiness and appeal. For example, if it emerged that humans are reliably more truthful with robots<sup>6</sup> or conversational AIs (chatbots) than they are with other humans, it would only be a matter of time before robots were used to interrogate humans — and if it emerged that robots are generally more believable than humans, then robots would likely be used as sales representatives.

It is also possible that people could become psychologically dependent on robots. Technology is known to tap into the reward functions of the brain, and this addiction could lead people to perform actions they would not have performed otherwise.

---

<sup>6</sup> The word's first chatbot ELIZA, developed by AI pioneer Joseph Weizenbaum showed that many early users were convinced of ELIZA's intelligence and understanding, despite Weizenbaum's insistence to the contrary.

It may be difficult to predict the psychological effects of forming a relationship with a robot. For example, Borenstein and Arkin (2019) ask how a 'risk-free' relationship with a robot may affect the mental and social development of a user; presumably, a robot would not be programmed to break up with a human companion, thus theoretically removing the emotional highs and lows from a relationship.

Enjoying a friendship or relationship with a companion robot may involve mistaking, at a conscious or unconscious level, the robot for a real person. To benefit from the relationship, a person would have to 'systematically delude themselves regarding the real nature of their relation with the [AI]' (Sparrow, 2002). According to Sparrow, indulging in such 'sentimentality of a morally deplorable sort' violates a duty that we have to ourselves to apprehend the world accurately. Vulnerable people would be especially at risk of falling prey to this deception (Sparrow and Sparrow, 2006).

### Human-human relationships

Robots may affect the stability of marital or sexual relationships. For instance, feelings of jealousy may emerge if a partner is spending time with a robot, such as a 'virtual girlfriend' (chatbot avatar). Loss of contact with fellow humans and perhaps a withdrawal from normal everyday relationships is also a possibility. For example, someone with a companion robot may be reluctant to go to events (say, a wedding) where the typical social convention is to attend as a human-human couple. People in human-robot relationships may be stigmatised.

There are several ethical issues brought about by humans forming relationships with robots:

- Could robots change the beliefs, attitudes, and/or values we have about human-human relationships? People may become impatient and unwilling to put the effort into working on human-human relationships when they can have a relationship with a 'perfect' robot and avoid these challenges.
- Could 'intimate robots' lead to an increase in violent behaviour? Some researchers argue that 'sexbots' would distort people's perceptions about the value of a human being, increasing people's desire or willingness to harm others. If we are able to treat robots as instruments for sexual gratification, then we may become more likely to treat other people this way. For example, if a user repeatedly punched a companion robot, would this be unethical (Lalji, 2015)? Would violence towards robots normalise a pattern of behaviour that would eventually affect other humans? However, some argue that robots could be an outlet for sexual desire, reducing the likelihood of violence, or to help recovery from assault.

Machines made to look and act like us could also affect the 'social suite' of capacities we have evolved to cooperate with one another, including love, friendship, cooperation and teaching (Christakis, 2019). In other words, AI could change how loving and kind we are—not just in our direct interactions with the machines in question, but in our interactions with one another. For example, should we worry about the effect of children being rude to digital assistants such as Alexa or Siri? Does this affect how they view or treat others?

Research shows that robots have the capacity to change how cooperative we are. In one experiment, small groups of people worked with a humanoid robot to lay railroad tracks in a virtual world. The robot was programmed to make occasional errors — and to acknowledge them and apologise. Having a clumsy, apologetic robot actually helped these groups perform *better* than control groups, by improving collaboration and communication among the human group members. This was also true in a second experiment, where people in groups containing error-prone robots consistently outperformed others in a problem-solving task (Christakis, 2017).

Both of these studies demonstrate that AI can improve the way humans relate to one another. However, AI can also make us behave less productively and less ethically. In another experiment, Christakis and his team gave several thousand subjects money to use over multiple rounds of an online game. In each round, subjects were told that they could either be selfish and keep their money, or be altruistic and donate some or all of it to their neighbours. If they made a donation, the researchers matched it, doubling the money their neighbours received. Although two thirds of people initially acted altruistically, the scientists found that the group's behaviour could be changed simply by adding just a few robots (posing as human players) that behaved selfishly. Eventually, the human players ceased cooperating with each other. The bots thus converted a group of generous people into selfish ones.

The fact that AI might reduce our ability to work together is concerning, as cooperation is a key feature of our species. 'As AI permeates our lives, we must confront the possibility that it will stunt our emotions and inhibit deep human connections, leaving our relationships with one another less reciprocal, or shallower, or more narcissistic,' says Christakis (2019).

## 2.2.4 Personhood

As machines increasingly take on tasks and decisions traditionally performed by humans, should we consider giving AI systems 'personhood' and moral or legal agency? One way of programming AI systems is 'reinforcement learning', where improved performance is reinforced with a virtual reward. Could we consider a system to be suffering when its reward functions give it negative input? Once we consider machines as entities that can perceive, feel and act, it is no huge leap to ponder their legal status. Should they be treated like animals of comparable intelligence? Will we consider the suffering of 'feeling' machines?

Scholars have increasingly discussed the legal status(es) of robots and AI systems over the past three decades. However, the debate was reignited recently when a 2017 resolution of the EU parliament invited the European Commission 'to explore, analyse and consider the implications of all possible legal solutions, [including]...creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently'.

However, the resolution provoked a number of objections, including an open letter from several 'Artificial Intelligence and Robotics Experts' in April 2018 which stated that 'the creation of a Legal Status of an 'electronic person' for 'autonomous', 'unpredictable' and 'self-learning' robots' should be discarded from technical, legal and ethical perspectives. Attributing electronic personhood to robots risks misplacing moral responsibility, causal accountability and legal liability regarding their mistakes and misuses, said the letter.

The majority of ethics research regarding AI seems to agree that AI machines should not be given moral agency, or seen as persons. Bryson (2018) argues that giving robots moral agency could in itself be construed as an immoral action, as 'it would be unethical to put artefacts in a situation of competition with us, to make them suffer, or to make them unnecessarily mortal'. She goes on to say that

*'there are substantial costs but little or no benefits from the perspective of either humans or robots to ascribing and implementing either agency or patency to intelligent artefacts beyond that ordinarily ascribed to any possession. The responsibility for any moral action taken by an artefact should therefore be attributed to its owner or operator, or in case of malfunctions to its manufacturer, just as with conventional artefacts'.*

## 2.3 Impact on the financial system

One of the first domains where autonomous applications have taken off is in financial markets, with most estimates attributing over half of trading volume in US equities to algorithms (Wellman and Rajan, 2017).

Markets are well suited to automation, as they now operate almost entirely electronically, generating huge volumes of data at high velocity, which require algorithms to digest. The dynamism of markets means that timely response to information is critical, providing a strong incentive to take slow humans out of the decision loop. Finally, and perhaps most obviously, the rewards available for effective trading decisions are considerable, explaining why firms have invested in this technology to the extent that they have. In other words, algorithmic trading can generate profits at a speed and frequency that is impossible for a human trader.

Although today's autonomous agents operate within a relatively narrow scope of competence and autonomy, they nevertheless take actions with consequences for people.

A well-known instance is that of Knight Capital Group. During the first 45 minutes of the trading day on 1 August 2012, while processing 212 small orders from customers, an automated trading agent developed by and operating on behalf of Knight Capital erroneously submitted millions of orders to the equity markets. Over four million transactions were executed in the financial markets as a result, leading to billions of dollars in net long and short positions. The company lost \$460 million on the unintended trades, and the value of its own stock fell by almost 75%.

Although this is an example of an accidental harm, autonomic trading agents could also be used maliciously to destabilise markets, or otherwise harm innocent parties. Even if their use is not intended to be malicious, the autonomy and adaptability of algorithmic trading strategies, including the increasing use of sophisticated machine learning techniques makes it difficult to understand how they will perform in unanticipated circumstances.

### Market manipulation

King et al. (2019) discuss several ways in which autonomous financial agents could commit financial crimes, including market manipulation, which is defined as 'actions and/or trades by market participants that attempt to influence market pricing artificially' (Spatt, 2014).

Simulations of markets comprising artificial trading agents have shown that, through reinforcement learning, an AI can learn the technique of order-book spoofing, which involves placing orders with no intention of ever executing them in order to manipulate honest participants in the marketplace (Lin, 2017).

Social bots have also been shown to exploit markets by artificially inflating stock through fraudulent promotion, before selling its position to unsuspecting parties at an inflated price (Lin 2017). For instance, in a recent prominent case a social bot network's sphere of influence was used to spread disinformation about a barely traded public company. The company's value gained more than 36,000% when its penny stocks surged from less than \$0.10 to above \$20 a share in a matter of few weeks (Ferrara 2015).

### *Collusion*

Price fixing, a form of collusion may also emerge in automated systems. As algorithmic trading agents can learn about pricing information almost instantaneously, any action to lower a price by

one agent will likely be instantaneously matched by another. In and of itself, this is no bad thing and only represents an efficient market. However, the possibility that lowering a price will result in your competitors simultaneously doing the same thing acts as a disincentive. Therefore, algorithms (if they are rational) will maintain artificially and tacitly agreed higher prices, by not lowering prices in the first place (Ezrachi and Stucke, 2016). Crucially, for collusion to take place, an algorithm does not need to be designed specifically to collude.

## Accountability

While the responsibility for trading algorithms rests with the organisations' that develop and deploy them, autonomous agents may perform actions — particularly in unusual circumstances — that would have been difficult to anticipate by their programmers. Does that difficulty mitigate responsibility to any degree?

For example, Wellman and Rajan (2017) give the example of an autonomous trading agent conducting an arbitrage operation, which is when a trader takes advantage of a discrepancy in prices for an asset in order to achieve a near-certain profit. Theoretically, the agent could attempt to instigate arbitrage opportunities by taking malicious actions to subvert markets, for example by propagating misinformation, obtaining improper access to information, or conducting direct violations of market rules

Clearly, it would be disadvantageous for autonomous trading agents to engage in market manipulation, however could an autonomous algorithm even meet the legal definition of market manipulation, which requires 'intent'?

Wellmen and Rajan (2017) argue that trading agents will become increasingly capable of operating at wider levels without human oversight, and that regulation is now needed to prevent societal harm. However, attempts to regulate or legislate may be hampered by several issues.

## 2.4 Impact on the legal system

The creation of AI machines and their use in society could have a huge impact on criminal and civil law. The entire history of human laws has been built around the assumption that people, and not robots, make decisions. In a society in which increasingly complicated and important decisions are being handed over to algorithms, there is the risk that the legal frameworks we have for liability will be insufficient.

Arguably, the most important near-term legal question associated with AI is who or what should be liable for tortious, criminal, and contractual misconduct involving AI and under what conditions.

### 2.4.1 Criminal law

A crime consists of two elements: a voluntary criminal act or omission (*actus reus*) and an intention to commit a crime (*mens rea*). If robots were shown to have sufficient awareness, then they could be liable as direct perpetrators of criminal offenses, or responsible for crimes of negligence. If we admit that robots have a mind of their own, endowed with human-like free will, autonomy or moral sense, then our whole legal system would have to be drastically amended.

Although this is possible, it is not likely. Nevertheless, robots may affect criminal laws in more subtle ways.

## Liability

The increasing delegation of decision making to AI will also impact many areas of law for which *mens rea*, or intention, is required for a crime to have been committed.

What would happen, for example if an AI program chosen to predict successful investments and pick up on market trends made a wrong evaluation that led to a lack of capital increase and hence, to the fraudulent bankruptcy of the corporation? As the intention requirement of fraud is missing, humans could only be held responsible for the lesser crime of bankruptcy triggered by the robot's evaluation (Pagallo, 2017).

Existing liability models may be inadequate to address the future role of AI in criminal activities (King et al, 2019). For example, in terms of *actus reus*, while autonomous agents can carry out the criminal act or omission, the voluntary aspect of *actus reus* would not be met, since the idea that an autonomous agent can act voluntarily is contentious. This means that agents, artificial or otherwise could potentially perform criminal acts or omissions without satisfying the conditions of liability for that particular criminal offence.

When criminal liability is fault-based, it also requires *mens rea* (a guilty mind). The *mens rea* may comprise an intention to commit the *actus reus* using an AI-based application, or knowledge that deploying an autonomous agent will or could cause it to perform a criminal action or omission. However, in some cases the complexity of the autonomous agent's programming could make it possible that the designer, developer, or deployer would neither know nor be able to predict the AI's criminal act or omission. This provides a great incentive for human agents to avoid finding out what precisely the machine learning system is doing, since the less the human agents know, the more they will be able to deny liability for both these reasons (Williams 2017).

The actions of autonomous robots could also lead to a situation where a human manifests the *mens rea*, and the robot commits the *actus reus*, splintering the components of a crime (McAllister 2017).

Alternatively, legislators could define criminal liability without a fault requirement. This would result in liability being assigned to the person who deployed the AI regardless of whether they knew about it, or could predict the illegal behaviour. Faultless liability is increasingly used for product liability in tort law (e.g., pharmaceuticals and consumer goods). However, Williams (2017) argues that *mens rea* with intent or knowledge is important, and we cannot simply abandon that key requirement of criminal liability in the face of difficulty in proving it.

Kingston (2018) references a definition provided by Hallevy (2010) on how AI actions may be viewed under criminal law. According to Hallevy, these legal models can be split into three scenarios:

1. *Perpetrator-via-another*. If an offence is committed by an entity that lacks the mental capacity for *mens rea* – a child, animal, or mentally deficient person – then they are deemed an innocent agent. However, if this innocent agent was instructed by another to commit the crime, then the instructor is held criminally liable. Under this model, an AI may be held to be an innocent agent, with either the software programmer or user filling the role of perpetrator-via-another.
2. *Natural-probable-consequence*. This relates to the accomplices of a criminal action; if no conspiracy can be proven, an accomplice may still be held legally liable if the perpetrator's acts were a natural or probable consequence of a scheme encouraged or aided by an accomplice. This scenario may hold when an AI that was designed for a 'good' purpose is misappropriated and commits a crime. For example, a factory line robot may injure a nearby worker they erroneously consider a threat to their programmed mission. In this

case, programmers may be held liable as accomplices if they knew that a criminal offence was a natural or probable consequence of their program design or use. This would not hold for an AI that was programmed to do a 'bad' thing, but to those that are misappropriated. Anyone capable and likely of foreseeing an AI being used in a specific criminal way may be held liable under this scenario: the programmer, the vendor, the service provider, or the user (assuming that the system limitations and possible consequences of misuse are spelt out in the AI instructions – which is unlikely).

3. *Direct liability.* This model attributes both *actus* and *mens rea* to an AI. However, while *actus rea* (the action or inaction) is relatively simple to attribute to an AI, says Kingston (2018), attributing *mens rea* (a guilty mind) is more complex. For example, the AI program 'driving' an autonomous vehicle that exceeds the speed limit could be held criminally liable for speeding – but for strict liability scenarios such as this, no criminal intent is required, and it is not necessary to prove that the car sped knowingly. Kingston also flags a number of possible issues that arise when considering AI to be directly liable. For example, could an AI infected by a virus claim a defence similar to coercion or intoxication, or an AI that is malfunctioning claim a defence akin to insanity? What would punishment look like – and who would be punished?

Identifying who exactly would be held liable for an AI's actions is important, but also potentially difficult. For example, 'programmer' could apply to multiple collaborators, or be widened to encompass roles such as program designer, product expert, and their superiors – and the fault may instead lie with a manager that appointed an inadequate expert or programmer (Kingston, 2010).

## Psychology

There is a risk that AI robots could manipulate a user's mental state in order to commit a crime. This was demonstrated by Weizenbaum (1976) who conducted early experiments into human–bot interactions where people revealed unexpectedly personal details about their lives. Robots could also normalise sexual offences and crimes against people, such as the case of certain sexbots (De Angeli, 2009).

## Commerce, financial markets and insolvency

As discussed earlier in this report, there are concerns that autonomous agents in the financial sector could be involved in market manipulation, price fixing and collusion. The lack of intention by human agents, and the likelihood that autonomous agents (AAs) may act together also raises serious problems with respect to liability and monitoring. It would be difficult to prove that the human agent intended the AA to manipulate markets, and it would also be difficult to monitor such manipulations. The ability of AAs to learn and refine their capabilities also implies that these agents may evolve new strategies, making it increasingly difficult to detect their actions (Farmer and Skouras 2013).

## Harmful or Dangerous Drugs

In the future AI could be used by organised criminal gangs to support the trafficking and sale of banned substances. Criminals could use AI equipped unmanned vehicles and autonomous navigation technologies to smuggle illicit substances. Because smuggling networks are disrupted by monitoring and intercepting transport lines, law enforcement becomes more difficult when unmanned vehicles are used to transport contraband. According to Europol (2017), drones present a real threat in the form of automated drug smuggling. Remote-controlled cocaine-trafficking submarines have already been discovered and seized by US law enforcement (Sharkey et al., 2010).

Unmanned underwater vehicles (UUVs) could also be used for illegal activities, posing a significant threat to enforcing drug prohibitions. As UUVs can act independently of an operator (Gogarty and Hagger, 2008), it would make it more difficult to catch the criminals involved.

Social bots could also be used to advertise and sell pornography or drugs to millions of people online, including children.

### Offences Against the Person

Social bots could also be used to harass people. Now that AI can generate more sophisticated fake content, new forms of harassment are possible. Recently, developers released software that produces synthetic videos where a person's face can be accurately substituted for another's. Many of these synthetic videos are pornographic and there is now the risk that malicious users may synthesise fake content in order to harass victims (Chesney and Citron 2018).

AI robots could also be used to torture and interrogate people, using psychological (e.g., mimicking people known to the torture subject) or physical torture techniques (McAllister 2017). As robots cannot understand pain or experience empathy, they will show no mercy or compassion. The mere presence of an interrogation robot may therefore cause the subject to talk out of fear. Using a robot would also serve to distance the human perpetrator from the *actus reus*, and emotionally distance themselves from their crime, making torture more likely.

As unthinking machines, AAs cannot bear moral responsibility or liability for their actions. However, one solution would be to take the approach of *strict* criminal liability, where punishment or damages may be imposed without proof of fault, which would lower the intention-threshold for the crime. However even under a strict liability framework, the question of who exactly should face imprisonment for AI-caused offences against a person is difficult. It is clear that an AA cannot be held liable. Yet, the number of actors involved creates a problem in ascertaining where the liability lies—whether with the person who commissioned and operated the AA, or its developers, or the legislators and policymakers who sanctioned real-world deployment of such agents (McAllister 2017).

### Sexual Offences

There is a danger that AI embodied robots could be used to promote sexual objectification, sexual abuse and violence. As discussed in section 2.1, sexbots could allow people to simulate sexual offences such as rape fantasies. They could even be designed to emulate sexual offences, such as adult and child rape (Danaher 2017).

Interaction with social bots and sexbots could also desensitise a perpetrator towards sexual offences, or even heighten their desire to commit them (De Angeli 2009; Danaher 2017).

## Who is responsible?

When considering the possible consequences and misuse of an AI, the key question is: *who is responsible for the actions of an AI?* Is it the programmers, manufacturers, end users, the AI itself, or another? Is the answer to this question the same for all AI or might it differ, for example, for systems capable of learning and adapting their behaviour?

According to the European Parliament Resolution (2017) on AI, legal responsibility for an AI's action (or inaction) is traditionally attributed to a human actor: the owner, developer, manufacturer or operator of an AI, for instance. For example, self-driving cars in Germany are currently deemed the responsibility of their owner. However, issues arise when considering third-party involvement, and advanced systems such as self-learning neural networks: if an action cannot be predicted by the developer because an AI has sufficiently changed from their design, can a developer be held responsible for that action? Additionally, current legislative infrastructure and the lack of effective regulatory mechanisms pose a challenge in regulating AI and assigning blame, say Atabekov and Yastrebov (2018), with autonomous AI in particular raising the question of whether a new legal category is required to encompass their features and limitations (European Parliament, 2017).

Taddeo and Floridi (2018) highlight the concept of 'distributed agency'. As an AI's actions or decisions come about following a long, complex chain of interactions between both human and robot – from developers and designers to manufacturers, vendors and users, each with different motivations, backgrounds, and knowledge – then an AI outcome may be said to be the result of distributed agency. With distributed agency comes distributed responsibility. One way to ensure that AI works towards 'preventing evil and fostering good' in society may be to implement a moral framework of distributed responsibility that holds all agents accountable for their role in the outcomes and actions of an AI (Taddeo and Floridi, 2018).

Different applications of AI may require different frameworks. For example, when it comes to military robots, Lokhorst and van den Hoven (2014) suggest that the primary responsibility lies with a robot's designer and deployer, but that a robot may be able to hold a certain level of responsibility for its actions.

Learning machines and autonomous AI are other crucial examples. Their use may create a 'responsibility gap', says Matthias (2004), where the manufacturer or operator of a machine may, in principle, be unable to predict a given AI's future behaviour – and thus cannot be held responsible for it in either a legal or moral sense. Matthias proposes that the programmer of a neural network, for instance, increasingly becomes the 'creator of software organisms', with very little control past the point of coding. The behaviour of such AI deviates from the initial programming to become a product of its interactions with its environment – the clear distinction between the phases of programming, training, and operation may be lost, making the ascription of blame highly complex and unclear. This responsibility gap requires the development and clarification of appropriate moral practice and legislation alongside the deployment of learning automata (Matthias, 2004). This is echoed by Scherer (2016), who states that AI has so far been developed in 'a regulatory vacuum', with few laws or regulations designed to explicitly address the unique challenges of AI and responsibility.

## Theft and fraud, and forgery and impersonation

AI could be used to gather personal data, and forge people's identities. For example, social media bots that add people as 'friends' would get access to their personal information, location, telephone number, or relationship history (Bilge et al., 2009). AI could manipulate people by building rapport with them, then exploiting that relationship to obtain information from or access to their computer (Chantler and Broadhurst 2006).

AI could also be used to commit banking fraud by forging a victim's identity, including mimicking a person's voice. Using the capabilities of machine learning, Adobe's software is able to learn and reproduce people's individual speech pattern from a 20-min recording of that person's voice. Copying the voice of the customer could allow criminals to talk to the person's bank and make transactions.

### 2.4.2 Tort law

Tort law covers situations where one person's behaviour causes injury, suffering, unfair loss, or harm to another person. This is a broad category of law that can include many different types of personal injury claims.

Tort laws serve two basic, general purposes: 1) to compensate the victim for any losses caused by the defendant's violations; and 2) to deter the defendant from repeating the violation in the future.

Tort law will likely come into sharp focus in the next few years as self-driving cars emerge on public roads. In the case of self-driving autonomous cars, when an accident occurs there are two areas of law that are relevant - negligence and product liability.

Today most accidents result from driver error, which means that liability for accidents are governed by negligence principles (Lin et al, 2017). Negligence is a doctrine that holds people liable for acting unreasonably under the circumstances (Anderson et al, 2009). To prove a negligence claim, a plaintiff must show that:

- A duty of care is owed by the defendant to the plaintiff
- There has been a breach of that duty by the defendant
- There is a causal link between the defendant's breach of duty and the plaintiff's harm, and;
- That the plaintiff has suffered damages as a result.

Usually insurance companies determine the at fault party, avoiding a costly lawsuit. However this is made much more complicated if a defect in the vehicle caused the accident. In the case of self-driving cars, accidents could be caused by hardware failure, design failure or a software error – a defect in the computer's algorithms.

Currently, if a collision is caused by an error or defect in a computer program, the manufacturer would be held responsible under the Product Liability doctrine, which holds manufacturers, distributors, suppliers, retailers, and others who make products available to the public responsible for the injuries those products cause.

As the majority of autonomous vehicle collisions are expected to be through software error, the defect would likely have to pass the 'risk-utility test' (Anderson et al., 2010), where a product is defective if the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design by the seller, and the omission of the alternative design renders the product not reasonably safe.

However, risk-utility test cases, which are needed to prove design defects are complex and require many expert witnesses, making design defect claims expensive to prove (Gurney et al, 2013). The nature of the evidence, such as complex algorithms and sensor data is also likely to make litigation especially challenging and complex.

This means the methods used to recover damages for car accidents would have to switch from an established, straightforward area of the law into a complicated and costly area of law (products liability). A plaintiff would need multiple experts to recover and find the defect in the algorithm, which would have implications for even the most straightforward of autonomous vehicle accidents. This would likely affect the ability of victims to get compensation and redress for injuries sustained in car accidents.

## 2.5 Impact on the environment and the planet

AI and robotics technologies require considerable computing power, which comes with an energy cost. Can we sustain massive growth in AI from an energetic point of view when we are faced with unprecedented climate change?

### 2.5.1 Use of natural resources

The extraction of nickel, cobalt and graphite for use in lithium ion batteries – commonly found in electrical cars and smartphones - has already damaged the environment, and AI will likely increase this demand. As existing supplies are diminished, operators may be forced to work in more complex environments that are dangerous to human operators – leading to further automation of mining and metal extraction (Khakurel et al., 2018). This would increase the yield, and depletion rate of rare earth metals, degrading the environment further.

### 2.5.2 Pollution and waste

At the end of their product cycle, electronic goods are usually discarded, leading to a build-up of heavy metals and toxic materials in the environment (O'Donoghue, 2010).

Increasing the production and consumption of technological devices such as robots will exacerbate this waste problem, particularly as the devices will likely be designed with 'inbuilt obsolescence' – a process where products are designed to wear out 'prematurely' so that customers have to buy replacement items – resulting in the generation of large amounts of electronic waste (Khakurel et al., 2018). Planned obsolescence depletes the natural environment of resources such as rare earth metals, while increasing the amount of waste. Sources indicate that in North America, over 100 million cell phones and 300 million personal computers are discarded each year (Guitinana et al., 2009).

Ways of combating this include 'encouraging consumers to prefer eco-efficient, more sustainable products and services' (World Business Council for Sustainable Development, 2000). However, this is hampered by consumers expecting frequent upgrades, and the lack of consumer concern for environmental consequences when contemplating an upgrade.

### 2.5.3 Energy concerns

As well as the toll that increased mining and waste will have on the environment, adoption of AI technology, particularly machine learning, will require more and more data to be processed. And that requires huge amounts of energy. In the United States, data centres already account for about 2 percent of all electricity used. In one estimation, DeepMind's AlphaGo – which beat Go Champion Lee Sedol in 2016 – took 50,000 times as much power as the human brain to do so (Mattheij, 2016).

AI will also require large amounts of energy for manufacturing and training – for example, it would take many hours to train a large-scale AI model to understand and recognise human language such that it could be used for translation purposes (Winfield, 2019b). According to Strubell, Ganesh, and McCallum (2019), the carbon footprint of training, tuning, and experimenting with a natural language processing AI is over seven times that of an average human in one year, and roughly 1.5 times the carbon footprint of an average car, including fuel, across its entire lifetime.

#### 2.5.4 Ways AI could help the planet

Alternatively AI could actually help us take better care of the planet, by helping us manage waste and pollution. For example, the adoption of autonomous vehicles could reduce greenhouse gas emissions, as autonomous vehicles could be programmed to follow the principles of eco-driving throughout a journey, reducing fuel consumption by as much as 20 percent and reducing greenhouse gas emissions to a similar extent (Iglinski et al., 2017). Autonomous vehicles could also reduce traffic congestion by recommending alternative routes and the shortest routes possible, and by sharing traffic information to other vehicles on the motorways, resulting in less fuel consumption.

There are also applications for AI in conservation settings. For example, deep-learning technology could be used to analyse images of animals captured by motion-sensor cameras in the wild. This information could then be used to provide accurate, detailed, and up-to-date information about the location, count, and behaviour of animals in the wild, which could be useful in enhancing local biodiversity and local conservation efforts (Norouzzadeh et al., 2018).

### 2.6 Impact on trust

AI is set to change our daily lives in domains such as transportation; the service industry; health-care; education; public safety and security; and entertainment. Nevertheless, these systems must be introduced in ways that build trust and understanding, and respect human and civil rights (Dignum, 2018). They need to follow fundamental human principles and values, and safeguard the well-being of people and the planet.

The overwhelming consensus amongst the research community is that trust in AI can only be attained by fairness, transparency, accountability and regulation. Other issues that impact on trust are how much control we want to exert over AI machines, and if, for example we want to always maintain a human-in the loop, or give systems more autonomy.

While robots and AI are largely viewed positively by citizens across Europe, they also evoke mixed feelings, raising concern and unease (European Commission 2012; European Commission 2017). Two Eurobarometer surveys, which aim to gauge public perception, acceptance, and opinion of specific topics among EU citizens in Member States, have been performed to characterise public attitudes towards robots and AI (survey 382), and towards increasing digitisation and automation (survey 460).

These surveys suggest that there is some way to go before people are comfortable with the widespread use of robots and advanced technology in society. For example, while respondents favoured the idea of prioritising the use of robots in areas that pose risk or difficulty to humans — space exploration, manufacturing, military, security, and search and rescue, for instance — they were very uncomfortable with areas involving vulnerable or dependent areas of society. Respondents opposed the use of robots to care for children, the elderly, and the disabled; for education; and for healthcare, despite many holding positive views of robots in general. The majority of those surveyed were also 'totally uncomfortable' with the idea of having their dog

walked by a robot, having a medical operation performed by a robot, or having their children or elderly parents minded by a robot — scenarios in which trust is key.

### 2.6.1 Why trust is important

*'In order for AI to reach its full potential, we must allow machines to sometimes work autonomously, and make decisions by themselves without human input', explains Taddeo (2017).*

Imagine a society in which there is no trust in doctors, teachers, or drivers. Without trust we would have to spend a significant portion of our lives devoting time and resources to making sure other people, or things were doing their jobs properly (Taddeo, 2017). This supervision would come at the expense of doing our own jobs, and would ultimately create a dysfunctional society.

*'We trust machine learning algorithms to indicate the best decision to make when hiring a future colleague or when granting parole during a criminal trial; to diagnose diseases and identify a possible cure. We trust robots to take care of our elderly and toddlers, to patrol borders, and to drive or fly us around the globe. We even trust digital technologies to simulate experiments and provide results that advance our scientific knowledge and understanding of the world. This trust is widespread and is resilient. It is only reassessed (rarely broken) in the event of serious negative consequences.' (Taddeo, 2017)*

In fact digital technologies are so pervasive that trusting them is essential for our societies to work properly. Constantly supervising a machine learning algorithm used to make a decision would require significant time and resources, to the point that using digital technologies would become unfeasible. At the same time, however, the tasks with which we trust digital technologies are of such relevance that a complete lack of supervision may lead to serious risks for our safety and security, as well for the rights and values underpinning our societies.

In other words, it is crucial to identify an effective way to trust digital technologies so that we can harness their value, while protecting fundamental rights and fostering the development of open, tolerant, just information societies (Floridi, 2016; Floridi and Taddeo, 2016). This is especially important in hybrid systems involving human and artificial agents.

But how do we find the correct level of trust? Taddeo suggests that in the short term design could play a crucial role in addressing this problem. For example, pop-up messages alerting users to algorithmic search engine results that have taken into account the user's online profile, or messages flagging that the outcome of an algorithm may not be objective. However in the long term, an infrastructure is needed that enforces norms such as fairness, transparency and accountability across all sectors.

### 2.6.2 Fairness

In order to trust AI it must be fair and impartial. As discussed in section 3.4, as more and more decisions are delegated to AI, we must ensure that those decisions are free from bias and discrimination. Whether it's filtering through CVs for job interviews, deciding on admissions to university, conducting credit ratings for loan companies, or judging the risk of someone reoffending, it's vital that decisions made by AI are fair, and do not deepen already entrenched social inequalities.

But how do we go about making algorithms fair? It's not as easy as it seems. The problem is that it is impossible to know what algorithms based on neural networks are actually learning when you train them with data. For example, the COMPAS algorithm, which assessed how likely someone was to commit a violent crime was found to strongly discriminate against black people. However the

algorithms were not actually given people's race as an input. Instead the algorithm inferred this sensitive data from other information, e.g. address.

For instance, one study found that two AI programs that had independently learnt to recognise images of horses from a vast library, used totally different approaches (Lapuschkin et al., 2019). While one AI focused rightly on the animal's features, the other based its decision wholly on a bunch of pixels at the bottom left corner of each horse image. It turned out that the pixels contained a copyright tag for the horse pictures. The AI worked perfectly for entirely the wrong reasons.

To devise a fair algorithm, first you must decide what a fair outcome looks like. Corbett-Davies et al. (2017) describe four different definitions of algorithmic fairness for an algorithm that assesses people's risk of committing a crime.

1. Statistical parity - where an equal proportion of defendants are detained in each race group. For example, white and black defendants are detained at equal rates.
2. Conditional statistical parity - where controlling for a limited set of 'legitimate' risk factors, an equal proportion of defendants are detained within each race group. For example, among defendants who have the same number of prior convictions, black and white defendants are detained at equal rates.
3. Predictive equality - where the accuracy of decisions is equal across race groups, as measured by false positive rate. This means that among defendants who would not have gone on to commit a violent crime if released, detention rates are equal across race groups.
4. Calibration - among defendants with a given risk score, the proportion who reoffend is the same across race groups.

However, while it is possible to devise algorithms that satisfy some of these requirements, many notions of fairness conflict with one another, and it is impossible to have an algorithm that meets all of them.

Another important aspect of fairness is to know *why* an automated program made a particular decision. For example, a person has the right to know why they were rejected for a bank loan. This requires transparency. However as we will find out, it is not always easy to find out why an algorithm came to a particular decision – many AIs employ complex 'neural networks' so that even their designers cannot explain how they arrive at a particular answer.

### 2.6.3 Transparency

A few years ago, a computer program in America assessed the performance of teachers in Houston by comparing their students' test scores against state averages (Sample, 2017). Those with high ratings won praise and even bonuses, while those with low ratings faced being fired. Some teachers felt that the system marked them down without good reason, however they had no way of checking if the program was fair or faulty as the company that built the software, the SAS Institute, considered its algorithm a trade secret and would not disclose its workings. The teachers took their case to court, and a federal judge ruled that the program had violated their civil rights.

This case study highlights the importance of transparency for building trust in AI - it should always be possible to find out *why* an autonomous system made a particular decision, especially if that decision caused harm. Given that real-world trials of driverless car autopilots have already resulted in several fatal accidents, there is clearly an urgent need for transparency in order to discover *how*

---

and *why* those accidents occurred, remedy any technical or operational faults, and establish accountability.

This issue is also prevalent amongst members of the public, especially when it comes to healthcare, a very personal issue for many (European Commission, 2017). For example, across Europe, many express concern over their lack of ability to access their health and medical records; while the majority would be happy to pass their records over to a healthcare professional, far fewer would be happy to do so to a public or private company for the purposes of medical research. These attitudes reflect concerns over trust, data access, and data use — all of which relate strongly to the idea of transparency and of understanding *what* AI gathers, *why*, and *how* one may access the data being gathered about them.

#### *Black boxes*

Transparency can be very difficult with modern AI systems, especially those based on deep learning systems. Deep learning systems are based on artificial neural networks (ANNs), a group of interconnected nodes, inspired by a simplification of the way neurons are connected in a brain. A characteristic of ANNs is that, after the ANN has been trained with datasets, any attempt to examine the internal structure of the ANN in order to understand why and how the ANN makes a particular decision is more or less impossible. Such systems are referred to as 'black boxes'.

Another problem is that of how to verify the system to confirm that it fulfils the specified design requirements. Current verification approaches typically assume that the system being verified will never change its behaviour, however systems based on machine learning—by definition—change their behaviour, so any verification is likely to be rendered invalid after the system has learned (Winfield and Jirotka, 2018).

The AI Now Institute at New York University, which researches the social impact of AI, recently released a report which urged public agencies responsible for criminal justice, healthcare, welfare and education to ban black box AIs because their decisions cannot be explained. The report also recommended that AIs should pass pre-release trials and be monitored 'in the wild' so that biases and other faults are swiftly corrected (AI Now Report, 2018).

In many cases, it may be possible to find out how an algorithm came to a particular decision without 'opening the AI black box'. Rather than exposing the full inner workings of an AI, researchers recently developed a way of working out what it would take to change their AI's decision (Wachter et al., 2018). Their method could explain why an AI turned down a person's mortgage application, for example, as it might reveal that the loan was denied because the person's income was £30,000, but would have been approved if it was £45,000. This would allow the decision to be challenged, and inform the person what they needed to address to get the loan.

Kroll (2018) argues that, contrary to the criticism that black-box software systems are inscrutable, algorithms are fundamentally understandable pieces of technology. He makes the point that inscrutability arises from the power dynamics surrounding software systems, rather than the technology itself, which is always built for a specific purpose, and can also always be understood in terms of design and operational goals, and inputs, outputs and outcomes. For example, while it is hard to tell why a particular ad was served to a particular person at a particular time, it is possible to do so, and to not do so is merely a design choice, not an inevitability of the complexity of large systems – systems must be designed so that they support analysis.

Kroll argues that it is possible to place too much focus on understanding the mechanics of a tool, when the real focus should be on how that tool is put to use and in what context.

Other issues and problems with transparency include the fact that software and data are proprietary works, which means it may not be in a company's best interest to divulge how they address a particular problem. Many companies view their software and algorithms as valuable trade secrets that are absolutely key to maintaining their position in a competitive market.

Transparency also conflicts with privacy, as people involved in training machine learning models may not want their data, or inferences about their data to be revealed. In addition, the lay public, or even regulators may not have the technological know-how to understand and assess algorithms.

### Explainable systems

Some researchers have demanded that systems produce explanations of their behaviours (Selbst and Barocas 2018; Wachter et al., 2017; Selbst and Powles, 2017). However, that requires a decision about what must be explained, and to whom. Explanation is only useful if it includes the context behind how the tool is operated. The danger is that explanations focus on the mechanism of how the tool operates at the expense of contextualising that operation.

In many cases, it may be unnecessary to understand the precise mechanisms of an algorithmic system, just as we do not understand how humans make decisions. Similarly, while transparency is often taken to mean the disclosure of source code or data, we don't have to see the computer source code for a system to be transparent, as this would tell us little about its behaviour. Instead transparency must be about the external behaviour of algorithms. This is how we regulate the behaviour of humans — not by looking into their brain's neural circuitry, but by observing their behaviour and judging it against certain standards of conduct.

Explanation may not improve human trust in a computer system, as even incorrect answers would receive explanations that may seem plausible. Automation bias, the phenomenon in which humans become more likely to believe answers that originate from a machine (Cummings, 2004), could mean that such misleading explanations have considerable weight.

### Intentional understanding

The simplest way to understand a piece of technology is to understand what it was designed to do, how it was designed to do that, and why it was designed in that particular way instead of some other way (Kroll, 2018). The best way of ensuring that a program does what you intend it to, and that there are no biases, or unintended consequences is through thorough validation, investigation and evaluation of the program during development. In other words, measuring the performance of a system during development in order to uncover bugs, biases and incorrect assumptions. Even carefully designed systems can miss important facts about the world, and it is important to verify that systems are operating as intended. This includes whether the model accurately measures what it is supposed to – a concept known as construct validity; and whether the data accurately reflects the real world

For example a machine learning model tasked with conducting credit checks could inadvertently learn that a borrower's quality of clothing correlates with their income and hence their creditworthiness. During development the software should be checked for such correlations, so that they can be rejected.

### Algorithm auditors

Larsson et al. (2019) suggest a role for professional algorithm auditors, whose job would be to interrogate algorithms in order to ensure they comply with pre-set standards. One example would be an autonomous vehicle algorithm auditor, who could provide simulated traffic scenarios to ensure that the vehicle did not disproportionately increase the risk to pedestrians or cyclists relative to passengers.

Recently, researchers proposed a new class of algorithms, called oversight programs, whose function is to 'monitor, audit, and hold operational AI programs accountable' (Etzioni and Etzioni 2016). For example, one idea would be to have an algorithm that conducts real-time assessments of the amount of bias caused by a news filtering algorithm, raising an alarm if bias increases beyond a certain threshold.

## 2.6.4 Accountability

*'How do decision-makers make sense of what decisions get made by AI technologies and how these decisions are different to those made by humans?... the point is that AI makes decisions differently from humans and sometimes we don't understand those differences; we don't know why or how it is making that decision.' (Jack Stilgoe)*

Another method of ensuring trust of AI is through accountability. As discussed, accountability ensures that if an AI makes a mistake or harms someone, there is someone that can be held responsible, whether that be the designer, the developer or the corporation selling the AI. In the event of damages incurred, there must be a mechanism for redress so that victims can be sufficiently compensated.

A growing body of literature has begun to address concepts such as algorithmic accountability and responsible AI. Algorithmic accountability, according to Caplan et al. (2018), deals with the delegation of responsibility for damages incurred as a result of algorithmically based decisions producing discriminatory or unfair consequences. One area where accountability is likely to be important is the introduction of self-driving vehicles. In the event of an accident, who should be held accountable? A number of fatal accidents have already occurred with self-driving cars, for example in 2016, a Tesla Model S equipped with radar and cameras determined that a nearby lorry was in fact the sky, which resulted in a fatal accident. In March 2018, a car used by Uber in self-driving vehicle trials hit and killed a woman in Arizona, USA. Even if autonomous cars are safer than vehicles driven by humans, accidents like these undermine trust.

## Regulation

One way of ensuring accountability is regulation. Winfield and Jirotka (2018) point out that technology is, in general, trusted if it brings benefits and is safe and well regulated. Their paper argues that one key element in building trust in AI is ethical governance – a set of processes, procedures, cultures and values designed to ensure the highest standards of behaviour. These standards of behaviour need to be adopted by individual designers and the organisations in which they work, so that ethical issues are dealt with as or before they arise in a principled manner, rather than waiting until a problem surfaces and dealing with it in an ad-hoc way.

They give the example of airliners, which are trusted because we know that they are part of a highly regulated industry with an outstanding safety record. The reason commercial aircraft are so safe is not just good design, it is also the tough safety certification processes, and the fact that when things do go wrong, there are robust and publicly visible processes of air accident investigation.

Winfield and Jirotka (2018) suggest that some robot types, driverless cars for instance, should be regulated through a body similar to the Civil Aviation Authority (CAA), with a driverless car equivalent of the Air Accident Investigation Branch.

When it comes to public perception of robots and advanced technology, regulation and management crops up as a prominent concern. In two surveys of citizens across the EU (European Commission 2012; European Commission, 2012), both showed that there was a generally positive view of robots and digitisation as long as this is implemented and managed carefully. In fact,

between 88% and 91% of those surveyed declared that robots and advanced technology must be managed carefully, one of the strongest results in either survey — reflecting a strong concern and area of priority amongst EU citizens.

## 2.6.5 Control

Another issue which affects public trust of AI is control. Much of this relates to fears around the idea of 'Superintelligence' - that as artificial intelligence increases to the point that it surpasses human abilities, it may come to take control over our resources and outcompete our species, leading to human extinction. A related fear is that, even if an AI agent was carefully designed to have goals aligned with human needs, it might develop for itself unanticipated subgoals that are not. For example, Bryson (2019) gives the example of a chess-playing robot taught to improve its game. This robot inadvertently learns to shoot people that switch it off at night, depriving it of vital resources. However, while most researchers agree this threat is unlikely to occur, to maintain trust in AI, it is important that humans have ultimate oversight over this technology.

### Human in the loop

One idea that has been suggested by researchers is that of always keeping a human-in-the-loop (HITL). Here a human operator would be a crucial component of the automated control process, supervising the robots. A simple form of HITL already in existence is the use of human workers to label data for training machine learning algorithms. For example when you mark an email as 'spam', you are one of many humans in the loop of a complex machine learning algorithm, helping it in its continuous quest to improve email classification as spam or non-spam.

However HITL can also be a powerful tool for regulating the behaviour of AI systems. For instance, many researchers argue that human operators should be able to monitor the behaviour of LAWS, or 'killer robots,' or credit scoring algorithms (Citron and Pasquale 2014). The presence of a human fulfils two major functions in a HITL AI system (Rahwan, 2018):

1. The human can identify misbehaviour by an otherwise autonomous system, and take corrective action. For instance, a credit scoring system may misclassify an adult as ineligible for credit because their age was incorrectly input—something a human may spot from the applicant's photograph. Similarly, a computer vision system on a weaponised drone may mis-identify a civilian as a combatant, and the human operator—it is hoped—would override the system.
2. Keeping humans in the loop would also provide accountability - if an autonomous system causes harm to human beings, having a human in the loop provides trust that somebody would bare the consequence of such mistakes. According to Rahwan (2018), until we find a way to punish algorithms for harm to humans, 'it is hard to think of any other alternative'.

However, although HITL is useful for building AI systems that are subject to oversight, it may not be enough. AI machines that make decisions with wider societal implications, such as algorithms that control millions of self-driving cars or news filtering algorithms that influence the political beliefs and preferences of millions of citizens, should be subject to oversight by society as a whole, requiring a 'society-in-the-loop' paradigm (Rahwan, 2018).

### The big red button

As a way to address some of the threats of artificial intelligence, researchers have proposed ways to stop an AI system before it has a chance to escape outside control and cause harm. A so-called 'big red button', or 'kill switch' would enable human operators to interrupt or divert a system, while preventing the system from learning that such an intervention is a threat. However, some

commentators fear that a sufficiently advanced AI machine could anticipate this move and defend itself by learning to disable its own 'kill switch'.

The red button raises wider practical questions about shutting down AI systems in order to keep them safe. What is the best way to accomplish that, and for what specific kinds of AI systems?

Orseau and Armstrong (2016) recently published a paper about how to prevent AI programmed through reinforcement learning (RL) from seeing interruptions as a threat. For example, an algorithm trying to optimise its chess performance may learn to disable its off switch so that it can spend more time learning how to play chess. Or it may learn to harm people who tried to switch it off, etc. What the researchers propose is to steer certain variants of reinforcement learning away from learning to avoid or impede an interruption. In this way, the authors argue, a system can pursue an optimal policy that is also interruptible. By being 'safely interruptible,' the paper concludes, reinforcement learning will not undermine the means of responsible oversight and intervention.

Riedl and Harrison (2017) suggests making a 'big red button' that, once pressed, diverted the AI into a simulated world where it could pursue its reward functions without causing any harm. Alternatively another idea is to maintain system uncertainty about key reward functions, which would prevent AI from attaching value to disabling an off-switch (Hadfield-Menell et al., 2016).

However Arnold and Schultz (2018) argue that the 'red button' approach comes at the point when a system has already 'gone rogue' and seeks to obstruct interference, and that 'big red button' approaches focus on long-term threats, imagining systems considerably more advanced than exist today and neglecting the present day problems with keeping automated systems accountable. A better approach, according to Arnold and Scheutz, would be to make ongoing self-evaluation and testing an integral part of a system's operation, in order to diagnose how the system is performing, and correct any errors.

They argue that to achieve this AIs should contain an ethical core (EC) consisting of a scenario-generation mechanism and a simulation environment used to test a system's decisions in simulated worlds, rather than the real world. This EC would be kept hidden from the system itself, so that the system's algorithms would be prevented from learning about its operation and its function, and ultimately its presence. Through continual testing in the simulated world, the EC would monitor and check for deviant behaviour - providing a far more effective and vigilant response than an emergency button which one might not get to push in time.

### 3. Ethical initiatives in the field of artificial intelligence

As detailed in previous sections, there are myriad ethical considerations accompanying the development, use and effects of artificial intelligence (AI). These range from the potential effects AI could have on the fundamental human rights of citizens within a society to the security and utilisation of gathered data; from the bias and discrimination unintentionally embedded into an AI by a homogenous group of developers, to a lack of public awareness and understanding about the consequences of their choices and usage of any given AI, leading to ill-informed decisions and subsequent harm.

AI builds upon previous revolutions in ICT and computing and, as such, will face a number of similar ethical problems. While technology may be used for good, potentially it may be misused. We may excessively anthropomorphise and humanise AI, blurring the lines between human and machine. The ongoing development of AI will bring about a new 'digital divide', with technology benefiting some socioeconomic and geographic groups more than others. Further, AI will have an impact on our biosphere and environment that is yet to be qualified (Veruggio and Oporto, 2006).

#### 3.1. International ethical initiatives

While official regulation remains scarce, many independent initiatives have been launched internationally to explore these – and other – ethical quandaries. The initiatives explored in this section are outlined in Table 3.1 and will be studied in light of the associated harms and concerns they aim to understand and mitigate.

Table 1: Ethical initiatives and harms addressed

Initiative	Location	Key issues tackled	Publications	Sources of funding
The Institute for Ethics in Artificial Intelligence	Germany	Human-centric engineering and a focus on the cultural and social anchoring of rapid advances in AI, covering disciplines including philosophy, ethics, sociology, and political science.		Initial (2019) funding grant from Facebook (\$7.5 million over five years).
The Institute for Ethical AI & Machine Learning	United Kingdom	The Institute aims to empower all from individuals to entire nations to develop AI, based on eight principles for responsible machine learning: these concern the maintenance of human control, appropriate redress for AI impact, evaluation of bias, explicability, transparency, reproducibility, mitigation of the effect of AI automation on workers, accuracy, cost, privacy, trust, and security.		unknown
The Institute for Ethical Artificial Intelligence in Education	United Kingdom	The potential threats to young people and education of the rapid growth of new AI technology, and ensuring the ethical development of AI-led EdTech.		unknown
The Future of Life Institute	United States	Ensuring that the development of AI is beneficial to humankind, with a focus on safety and existential risk: autonomous weapons arms race, human control of AI, and the potential dangers of advanced 'general/strong' or super-intelligent AI.	'Asilomar AI Principles'	Private. Top donors: Elon Musk (SpaceX and Tesla), Jaan Tallinn (Skype), Matt Wage (financial trader), Nisan Stiennon (software engineer), Sam Harris, George Godula (tech entrepreneur), and Jacob Trefethen (Harvard).
The Association for Computing Machinery	United States	The transparency, usability, security, accessibility, accountability, and digital inclusiveness of computers and networks, in terms of research, development, and implementation.	Statements on: algorithmic transparency and accountability (January 2017), computing and network security (May 2017), the Internet of Things (June 2017), accessibility, usability, and digital inclusiveness (September 2017),	unknown

			and mandatory access to information infrastructure for law enforcement (April 2018).	
The Japanese Society for Artificial Intelligence (JSAI)	Japan	To ensure that AI R&D remains beneficial to human society, and that development and research is conducted ethically and morally.	' <b>Ethical Guidelines'</b>	unknown
AI4All	United States	Diversity and inclusion in AI, to expose underrepresented groups to AI for social good and humanity's benefit.		Google
The Future Society	United States	The impact and governance of artificial intelligence to broadly benefit society, spanning policy research, advisory and collective intelligence, coordination of governance, law, and education.	' <b>Draft Principles for the Governance of AI</b> ' Published October 2017 (later published on their website on 7th February 2019),	unknown
The AI Now Institute	United States	The social implications of AI, especially in the areas of: Rights and liberties, labour and automation, bias and inclusion, and safety and critical infrastructure.		Various organisations, including Luminate, the MacArthur Foundation, Microsoft Research, Google, the Ford Foundation, DeepMind Ethics & Society, and the Ethics & Governance of AI Initiative.
The Institute of Electrical and Electronics Engineers (IEEE)	United States	Societal and policy guidelines to keep AI and intelligent systems human-centric, and serving humanity's values and principles. Focuses on ensuring that all stakeholders – across design and development – are educated, trained, and empowered to prioritise the ethical considerations of human rights, well-being, accountability, transparency, and awareness of misuse.	' <b>Ethically Aligned Design</b> ' First Edition (March 2019)	
The Partnership on AI	United States	Best practices on AI technologies: Safety, fairness, accountability, transparency, labour and the economy, collaboration between people and systems, social and societal influences, and social good.		The Partnership was formed by a group of AI researchers representing six of the world's largest tech companies: Apple,

				Amazon, DeepMind and Google, Facebook, IBM, and Microsoft.
The Foundation for Responsible Robotics	The Netherlands	Responsible robotics (in terms of design, development, use, regulation, and implementation). Proactively taking stock of the issues that accompany technological innovation, and the impact these will have on societal values such as safety, security, privacy, and well-being.		unknown
AI4People	Belgium	The social impacts of AI, and the founding principles, policies, and practices upon which to build a 'good AI society'.	<b>'Ethical Framework for a Good AI Society'</b>	Atomium—European Institute for Science, Media and Democracy. Some funding was provided to the project's Scientific Committee Chair from the Engineering and Physical Sciences Research Council.
The Ethics and Governance of Artificial Intelligence Initiative	United States	Seeks to ensure that technologies of automation and machine learning are researched, developed, and deployed in a way which vindicate social values of fairness, human autonomy, and justice.		The Harvard Berkman Klein Center and the MIT Media Lab. Supported by The Miami Foundation (fiscal sponsorship), Knight Foundation, Luminate, Red Hoffman, and the William and Flora Hewlett Foundation.
Saidot: Enabling responsible AI ecosystems	Finland	Helping companies, governments, and organisations develop and deploy responsible AI ecosystems, to deliver transparent, accountable, trustworthy AI services. Enabling organisations to develop human-centric AI, with a focus on increasing the levels of trust and accountability in AI ecosystems. The platform offers software and algorithmic systems that can 'validate [an] intelligence system's trustworthiness' (Saidot, 2019)		
euRobotics	Europe	Maintaining and extending European talent and progress in robotics – AI industrialisation and economic impact.		European Commission

The Centre for Data Ethics and Innovation	UK	Identifying and plugging gaps in our regulatory landscape, AI use of data, and maximising the benefits of AI to society.		UK Government
Special Interest Group on Artificial Intelligence (SIGAI), The Association for Computing Machinery	United States	Promoting and supporting the growth and application of AI principles and techniques throughout computing, and promoting AI education and publications through various forums		The Association for Computing Machinery
<b>Other key international developments: current and historical</b>				
The Montréal Declaration	Canada	The socially responsible development of AI, bringing together 400 participants across all sectors of society to identify the ethical and moral challenges in the short and long term. Key values: well-being, autonomy, justice, privacy, knowledge, democracy, and accountability.		Université de Montréal with the support of the Fonds de recherche en santé du Québec and the Palais des congrès de Montréal.
The UNI Global Union	Switzerland	Worker disruption and transparency in the application of AI, robotics, and data and machine learning in the workplace. Safeguarding workers' interests and maintaining human control and a healthy power balance.	'Top 10 Principles for Ethical AI'	unknown
The European Robotics Research Network (EURON)	Europe (Coordinator based in Sweden)	Research co-ordination, education and training, publishing and meetings, industrial links and international links in robotics.	'Roboethics Roadmap'	European Commission (2000-2004)
The European Robotics Platform (EUROP)	Europe	Bringing European robotics and AI community together. Industry-driven, focus on competitiveness and innovation.		European Commission

### 3.2. Ethical harms and concerns tackled by these initiatives

All of the initiatives listed above agree that AI should be researched, developed, designed, deployed, monitored, and used in an ethical manner – but each has different areas of priority. This section will include analysis and grouping of the initiatives above, by type of issues they aim to address, and then outline some of the proposed approaches and solutions to protect from harms.

A number of key issues emerge from the initiatives, which **can be broadly split into the following categories:**

1. Human rights and well-being  
*Is AI in the best interests of humanity and human well-being?*
2. Emotional harm  
*Will AI degrade the integrity of the human emotional experience, or facilitate emotional or mental harm?*
3. Accountability and responsibility  
*Who is responsible for AI, and who will be held accountable for its actions?*
4. Security, privacy, accessibility, and transparency  
*How do we balance accessibility and transparency with privacy and security, especially when it comes to data and personalisation?*
5. Safety and trust  
*What if AI is deemed untrustworthy by the public, or acts in ways that threaten the safety of either itself or others?*
6. Social harm and social justice  
*How do we ensure that AI is inclusive, free of bias and discrimination, and aligned with public morals and ethics?*
7. Financial harm  
*How will we control for AI that negatively affects economic opportunity and employment, and either takes jobs from human workers or decreases the opportunity and quality of these jobs?*
8. Lawfulness and justice  
*How do we go about ensuring that AI - and the data it collects - is used, processed, and managed in a way that is just, equitable, and lawful, and subject to appropriate governance and regulation? What would such regulation look like? Should AI be granted 'personhood'?*
9. Control and the ethical use – or misuse – of AI  
*How might AI be used unethically - and how can we protect against this? How do we ensure that AI remains under complete human control, even as it develops and 'learns'?*
10. Environmental harm and sustainability  
*How do we protect against the potential environmental harm associated with the development and use of AI? How do we produce it in a sustainable way?*
11. Informed use  
*What must we do to ensure that the public is aware, educated, and informed about their use of*

*and interaction with AI?*

12. Existential risk

*How do we avoid an AI arms race, pre-emptively mitigate and regulate potential harm, and ensure that advanced machine learning is both progressive and manageable?*

Overall, these initiatives all aim to identify and form ethical frameworks and systems that establish human beneficence at the highest levels, prioritise benefit to both human society and the environment (without these two goals being placed at odds), and mitigate the risks and negative impacts associated with AI — with a focus on ensuring that AI is accountable and transparent (IEEE, 2019).

The IEEE's '**Ethically Aligned Design: A Vision for Prioritising Human Well-being with Autonomous and Intelligent Systems**' (v1; 2019) is one of the most substantial documents published to date on the ethical issues that AI may raise — and the various proposed means of mitigating these.

Figure 2: General principles for the ethical and values-based design, development, and implementation of autonomous and intelligent systems (as defined by the IEEE's *Ethically Aligned Design* First Edition March 2019)



**Areas of key impact** comprise sustainable development; personal data rights and agency over digital identity; legal frameworks for accountability; and policies for education and awareness. They fall under **the three pillars of the Ethically Aligned Design conceptual framework:** Universal human values; political self-determination and data agency; and technical dependability.

### 3.2.1 Harms in detail

Taking each of these harms in turn, this section explores how they are being conceptualised by initiatives and some of the challenges that remain.

#### Human rights and well-being

All initiatives adhere to the view that ***AI must not impinge on basic and fundamental human rights***, such as human dignity, security, privacy, freedom of expression and information, protection of personal data, equality, solidarity and justice (European Parliament, Council and Commission, 2012).

How do we ensure that AI upholds such fundamental human rights and prioritises human well-being? Or that AI does not disproportionately affect vulnerable areas of society, such as children, those with disabilities, or the elderly, or reduce quality of life across society?

In order to ensure that human rights are protected, the IEEE recommends new governance frameworks, standards, and regulatory bodies which oversee the use of AI; translating existing legal obligations into informed policy, allowing for cultural norms and legal frameworks; and always maintaining complete human control over AI, without granting them rights or privileges equal to those of humans (IEEE, 2019). To safeguard human well-being, defined as 'human satisfaction with life and the conditions of life, as well as an appropriate balance between positive and negative affect' (*ibid*), the IEEE suggest prioritising human well-being throughout the design phase, and using the best and most widely-accepted available metrics to clearly measure the societal success of an AI.

There are crossovers with accountability and transparency: there must always be appropriate ways to identify and trace the impingement of rights, and to offer appropriate redress and reform. Personal data are also a key issue here; AI collect all manner of personal data, and users must retain the access to, and control of, their data, to ensure that their fundamental rights are being lawfully upheld (IEEE, 2019).

According to the ***Foundation for Responsible Robotics***, AI must be ethically developed with human rights in mind to achieve their goal of 'responsible robotics', which relies upon proactive innovation to uphold societal values like safety, security, privacy, and well-being. The Foundation engages with policymakers, organises and hosts events, publishes consultation documents to educate policymakers and the public, and creates public-private collaborations to bridge the gap between industry and consumers, to create greater transparency. It calls for ethical decision-making right from the research and development phase, greater consumer education, and responsible law- and policymaking – made before AI is released and put into use.

The ***Future of Life Institute*** defines a number of principles, ethics, and values for consideration in the development of AI, including the need to design and operate AI in a way that is compatible with the ideals of human dignity, rights, freedoms, and cultural diversity<sup>7</sup>. This is echoed by the ***Japanese Society for AI Ethical Guidelines***, which places the utmost importance on AI being realised in a way that is beneficial to humanity, and in line with the ethics, conscience, and competence of both its researchers and society as a whole. AI must contribute to the peace, safety, welfare, and public interest of society, says the Society, and protect human rights.

***The Future Society's Law and Society Initiative*** emphasises that human beings are equal in rights, dignity, and freedom to flourish, and are entitled to their human rights.<sup>8</sup> With this in mind, to what extent should we delegate to machines decisions that affect people? For example, could AI 'judges' in the legal profession be more efficient, equitable, uniform, and cost-saving than human ones –

---

<sup>7</sup> <https://futureoflife.org/ai-principles/>

<sup>8</sup> <http://thefuturesociety.org/law-and-society-initiative>

and even if they were, would this be an appropriate way to deploy AI? **The Montréal Declaration<sup>9</sup>** aims to clarify this somewhat, by pulling together an ethical framework that promotes internationally recognised human rights in fields affected by the rollout of AI: 'The principles of the current declaration rest on the common belief that human beings seek to grow as social beings endowed with sensations, thoughts and feelings, and strive to fulfil their potential by freely exercising their emotional, moral and intellectual capacities.' In other words, AI must not only not disrupt human well-being, but it must also proactively encourage and support it to improve and grow.

Some approach AI from a more specific viewpoint – such as the **UNI Global Union**, which strives to protect an individual's right to work. Over half of the work currently done by people could be done faster and more efficiently in an automated way, says the Union. This identifies a prominent harm that AI may cause in the realm of human employment. The Union states that we must ensure that AI serves people and the planet, and both protects and increases fundamental human rights, human dignity, integrity, freedom, privacy, and cultural and gender diversity<sup>10</sup>.

### Emotional harm

**What is it to be human?** AI will interact with and have an impact on the human emotional experience in ways that have not yet been qualified; humans are susceptible to emotional influence both positively and negatively, and '**affect – how emotion and desire influence behaviour – is a core part of intelligence**'. Affect varies across cultures, and, given different cultural sensitivities and ways of interacting, affective and influential AI could begin to influence how people view society itself. The **IEEE** recommend various ways to mitigate this risk, including the ability to adapt and update AI norms and values according to who they are engaging with, and the sensitivities of the culture in which they are operating.

There are various ways in which AI could inflict emotional harm, including false intimacy, over-attachment, objectification and commodification of the body, and social or sexual isolation. These are covered by various of the aforementioned ethical initiatives, including **the Foundation for Responsible Robotics, Partnership on AI, the AI Now institute** (especially regarding affect computing), **the Montréal Declaration**, and the **European Robotics Research Network (EURON) Roadmap** (for example, their section on the risks of humanoids).

These possible harms come to the fore when considering the development of an intimate relationship with an AI, for example in the sex industry. Intimate systems, as the **IEEE** call them, must not contribute to sexism, racial inequality, or negative body image stereotypes; must be for positive and therapeutic use; must avoid sexual or psychological manipulation of users without consent; should not be designed in a way that contributes to user isolation from human companionship; must be designed in a way that is transparent about the effect they may have on human relationship dynamics and jealousy; must not foster deviant or criminal behaviour, or normalise illegal sexual practices such as paedophilia or rape; and must not be marketed commercially as a person (in a legal sense or otherwise).

Affective AI is also open to the possibility of deceiving and coercing its users – researchers have defined the act of AI subtly modifying behaviour as '**nudging**', when an AI emotionally manipulates and influences its user through the affective system. While this may be useful in some ways – drug dependency, healthy eating – it could also trigger behaviours that worsen human health. Systematic analyses must examine the ethics of affective design prior to deployment; users must be educated on how to recognise and distinguish between nudges; users must have an opt-in system for autonomous nudging systems; and vulnerable populations that cannot give informed consent, such

---

<sup>9</sup> <https://www.montrealdeclaration-responsibleai.com/the-declaration>

<sup>10</sup> [http://www.thefutureworldofwork.org/media/35420/uni\\_ethical\\_ai.pdf](http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf)

as children, must be subject to additional protection. In general, stakeholders must discuss the question of whether or not the nudging design pathway for AI, which lends itself well to selfish or detrimental uses, is an ethical one to pursue (IEEE, 2019).

As raised by the **IEEE** (2019), nudging may be used by governments and other entities to influence public behaviour. Would it be ethically appropriate for a robot to use nudging to encourage, for example, charitable behaviour or donations? We must pursue full transparency regarding the beneficiaries of such behaviour, say the IEEE, due to the potential for misuse.

Other issues include technology addiction and emotional harm due to societal or gender bias.

### Accountability and responsibility

The vast majority of initiatives mandate that AI must be **auditable**, in order to assure that the designers, manufacturers, owners, and operators of AI are held accountable for the technology or system's actions, and are thus considered responsible for any potential harm it might cause. According to the **IEEE**, this could be achieved by the courts clarifying issues of culpability and liability during the development and deployment phases where possible, so that those involved understand their obligations and rights; by designers and developers taking into account the diversity of existing cultural norms among various user groups; by establishing multi-stakeholder ecosystems to create norms that currently do not exist, given that AI-oriented technology is too new; and by creating registration and record-keeping systems so that it is always possible to trace who is legally responsible for a particular AI.

The **Future of Life Institute** tackles the issue of accountability via its **Asilomar Principles**, a list of 23 guiding principles for AI to follow in order to be ethical in the short and long term. Designers and builders of advanced AI systems are 'stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications' (FLI, 2017); if an AI should make a mistake, it should also be possible to ascertain why. The **Partnership on AI** also stresses the importance of accountability in terms of bias. We should be sensitive to the fact that assumptions and biases exist within data and thus within systems built from these data, and strive not to replicate them – i.e. to be actively accountable for building fair, bias-free AI.

All other initiatives highlight the importance of accountability and responsibility – both by designers and AI engineers, and by regulation, law and society on a larger scale.

### Sex and Robots

In July of 2017, the **Foundation for Responsible Robotics** published a report on 'Our Sexual Future with Robots' (Foundation for Responsible Robotics, 2019). This aimed to present an objective summary of the various issues and opinions surrounding our intimate association with technology. Many countries are developing robots for sexual gratification; these largely tend to be pornographic representations of the human body – and are mostly female. These representations, when accompanied by human anthropomorphism, may cause robots to be perceived as somewhere between living and inanimate, especially when sexual gratification is combined with elements of intimacy, companionship and conversation. Robots may also affect societal perceptions of gender or body stereotypes, erode human connection and intimacy and lead to greater social isolation. However, there is also some potential for robots to be of emotional sexual benefit to humans, for example by helping to reduce sex crime, and to rehabilitate victims of rape or sexual abuse via inclusion in healing therapies.

## Access and transparency vs. security and privacy

A main concern over AI is its ***transparency***, explicability, security, reproducibility, and interpretability: is it possible to discover why and how a system made a specific decision, or why and how a robot acted in the way it did? This is especially pressing in the case of *safety-critical* systems that may have direct consequences for physical harm: driverless cars, for example, or medical diagnosis systems. Without transparency, users may struggle to understand the systems they are using – and their associated consequences – and it will be difficult to hold the relevant persons accountable and responsible.

To address this, the **IEEE** propose developing new standards that detail measurable and testable levels of transparency, so systems can be objectively assessed for their compliance. This will likely take different forms for different stakeholders; a robot user may require a 'why-did-you-do-that' button, while a certification agency or accident investigator will require access to relevant algorithms in the form of an 'ethical black box' which provides failure transparency (IEEE, 2019).

AI require data to continually learn and develop their automatic decision-making. These data are personal and may be used to identify a particular individual's physical, digital, or virtual identity (i.e. personally identifiable information, PII). 'As a result,' write the IEEE (2017), 'through every digital transaction (explicit or observed) humans are generating a unique digital shadow of their physical self'. To what extent can humans realise the right to keep certain information private, or have input into how these data are used? Individuals may lack the appropriate tools to control and cultivate their unique identity and manage the associated ethical implications of the use of their data. Without clarity and education, many users of AI will remain unaware of the digital footprint they are creating, and the information they are putting out into the world. Systems must be put in place for users to control, interact with and access their data, and give them agency over their digital personas.

PII has been established as the asset of the individual (by Regulation (EU) 2016/679 in Europe, for example), and systems must ask for explicit consent at the time data are collected and used, in order to protect individual autonomy, dignity and right to consent. The IEEE mention the possibility of a personalised 'privacy AI or algorithmic agent or guardian' to help individuals curate and control their personal data and foresee and mitigate potential ethical implications of machine learning data exchange.

The **Future of Life Institute's Asilomar Principles** agree with the IEEE on the importance of transparency and privacy across various aspects: failure transparency (if an AI fails, it must be possible to figure out why), judicial transparency (any AI involved in judicial decision-making must provide a satisfactory explanation to a human), personal privacy (people must have the right to access, manage, and control the data AI gather and create), and liberty and privacy (AI must not unreasonably curtail people's real or perceived liberties). **Saidot** takes a slightly wider approach and strongly emphasises the importance of AI that are transparent, accountable, and trustworthy, where

### Autonomy and agent vs. patient

The current approach to AI is undeniably anthropocentric. This raises **possible issues around the distinction between moral agents and moral patients, between artificial and natural, between self-organising and not**. AI cannot become autonomous in the same way that living beings are considered autonomous (IEEE, 2019), but how do we define autonomy in terms of AI? Machine autonomy designates how machines act and operate according to regulation, but any attempts to implant emotion and morality into AI 'blur the distinction between agents and patients and may encourage anthropomorphic expectations of machines', writes the **IEEE** — especially as embodied AI begins to look increasingly similar to humans. Establishing a usable distinction between human and system/machine autonomy involves questions of free will, being/becoming and predetermination. It is clear that further discussion is needed to clarify what 'autonomy' may mean in terms of artificial intelligence and systems.

people, organisations, and smart systems are openly connected and collaborative in order to foster cooperation, progress, and innovation.

All of the initiatives surveyed identify transparency and accountability of AI as an important issue. This balance underpins many other concerns – such as legal and judicial fairness, worker compensation and rights, security of data and systems, public trust, and social harm.

## Safety and trust

Where AI is used to supplement or replace human decision-making, there is consensus that it must be ***safe, trustworthy, and reliable, and act with integrity.***

The **IEEE** propose cultivating a 'safety mindset' among researchers, to 'identify and pre-empt unintended and unanticipated behaviors in their systems' and to develop systems which are 'safe by design'; setting up review boards at institutions as a resource and means of evaluating projects and their progress; encouraging a community of sharing, to spread the word on safety-related developments, research, and tools. The **Future of Life Institute's Asilomar principles** indicate that all involved in developing and deploying AI should be mission-led, adopting the norm that AI 'should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organisation' (Future of Life Institute, 2017). This approach would build public trust in AI, something that is key to its successful integration into society.

### An 'ethical black box'

Initiatives including the **UNI Global Union** and **IEEE** suggest equipping AI systems with an 'ethical black box': a device that can record information about said system to ensure its accountability and transparency, but that also includes clear data on the ethical consideration built into the system from the beginning (UNI Global Union, n.d.).

**The Japanese Society for AI** proposes that AI should act with integrity at all times, and that AI and society should earnestly seek to learn from and communicate with one another. 'Consistent and effective communication' will strengthen mutual understanding, says the Society, and '[contribute] to the overall peace and happiness of mankind' (JSAI, 2017). The **Partnership on AI** agrees, and strives to ensure AI is trustworthy and to create a culture of cooperation, trust, and openness among AI scientists and engineers. The **Institute for Ethical AI & Machine Learning** also emphasises the importance of dialogue; it ties together the issues of trust and privacy in its eight core tenets, mandating that AI technologists communicate with stakeholders about the processes and data involved to build trust and spread understanding throughout society.

## Social harm and social justice: inclusivity, bias, and discrimination

AI development requires **a diversity of viewpoints**. There are several organisations establishing that these must be in line with community viewpoints and align with social norms, values, ethics, and preferences, that biases and assumptions must not be built into data or systems, and that AI should be aligned with public values, goals, and behaviours, respecting cultural diversity. Initiatives also argue that all should have access to the benefits of AI, and it should work for the common good. In other words, developers and implementers of AI have a social responsibility to embed the right values into AI and ensure that they do not cause or exacerbate any existing or future harm to any part of society.

The **IEEE** suggest first identifying social and moral norms of the specific community in which an AI will be deployed, and those around the specific task or service it will offer; designing AI with the idea of 'norm updating' in mind, given that norms are not static and AI must change dynamically and transparently alongside culture; and identifying the ways in which people resolve norm conflicts, and equipping AI with a system in which to do so in a similar and transparent way. This should be done collaboratively and across diverse research efforts, with care taken to evaluate and assess potential biases that disadvantage specific social groups.

Several initiatives – such as **AI4ALL** and the **AI Now Institute** – explicitly advocate for fair, diverse, equitable, and non-discriminatory inclusion in AI at all stages, with a focus on support for under-represented groups. Currently, AI-related degree programmes do not equip aspiring developers and designers with an appropriate knowledge of ethics (IEEE, 2017), and corporate environments and business practices are not ethically empowering, with a lack of roles for senior ethicists that can steer and support value-based innovation.

On a global scale, the inequality gap between developed and developing nations is significant. While AI may have considerable usefulness in a humanitarian sense, they must not widen this gap or exacerbate poverty, illiteracy, gender and ethnic inequality, or disproportionately disrupt employment and labour. The IEEE suggests taking action and investing to mitigate the inequality gap; integrating corporate social responsibility (CSR) into development and marketing; developing transparent power structures; facilitating and sharing robotics and AI knowledge and research; and generally keeping AI in line with the US Sustainable Development Goals<sup>11</sup>. AI technology should be made equally available worldwide via global standardisation and open-source software, and interdisciplinary discussion should be held on effective AI education and training (IEEE, 2019).

A set of ethical guidelines published by the **Japanese Society for AI** emphasises, among other considerations, the importance of a) contribution to humanity, and b) social responsibility. AI must act in the public interest, respect cultural diversity, and always be used in a fair and equal manner.

The **Foundation for Responsible Robotics** includes a Commitment to Diversity in its push for responsible AI; the **Partnership on AI** cautions about the 'serious blind spots' of ignoring the presence of biases and assumptions hidden within data; **Saidot** aims to ensure that, although our social values are now 'increasingly mediated by algorithms', AI remains human-centric (Saidot, 2019); the **Future of Life Institute** highlights a need for AI imbued with human values of cultural diversity and human rights; and the **Institute for Ethical AI & Machine Learning** includes 'bias evaluation' for monitoring bias in AI development and production. The dangers of human bias and assumption are a frequently identified risk that will accompany the ongoing development of AI.

### Financial harm: Economic opportunity and employment

AI may disrupt the economy and lead to loss of jobs or work disruption for many humans, and will have an impact on workers' rights and displacement strategy as many strains of work become automated (and vanish in related business change).

Additionally, rather than just focusing on the number of jobs lost or gained, traditional employment structures will need to be changed to mitigate the effects of automation and take into account the complexities of employment. Technological change is happening too fast for the traditional workforce to keep pace without retraining. Workers must train for adaptability, says the **IEEE** (2019), and new skill sets, with fallback strategies put in place for those who cannot be re-trained, and training programmes implemented at the level of high school or earlier to increase access to future employment. The **UNI Global Union** call for multi-stakeholder ethical AI governance bodies on global and regional levels, bringing together designers, manufacturers, developers, researchers, trade unions, lawyers, CSOs, owners, and employers. AI must benefit and empower people broadly and equally, with policies put in place to bridge the economic, technological, and social digital divides, and ensure a just transition with support for fundamental freedoms and rights.

**The AI Now Institute** works with diverse stakeholder groups to better understand the implications that AI will have for labour and work, including automation and early-stage integration of AI changing the nature of employment and working conditions in various sectors. **The Future Society** specifically asks how AI will affect the legal profession: 'If AI systems are demonstrably superior to

---

<sup>11</sup> <https://sustainabledevelopment.un.org/?menu=1300>

human attorneys at certain aspects of legal work, what are the ethical and professional implications for the practice of law?' (Future Society, 2019)

AI in the workplace will affect far more than workers' finances, and may offer various positive opportunities. As laid out by the **IEEE** (2019), AI may offer potential solutions to workplace bias – if it is developed with this in mind, as mentioned above – and reveal deficiencies in product development, allowing proactive improvement in the design phase (as opposed to retroactive improvement).

*'RRI is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society).' (Von Schomberg, 2013)*

### Lawfulness and justice

Several initiatives address the need for AI to be lawful, equitable, fair, just and subject to appropriate, pre-emptive governance and regulation. The many complex ethical problems surrounding AI translate directly and indirectly into discrete legal challenges. How should AI be labelled: as a product? An animal? A person? Something new?

The **IEEE** conclude that AI should not be granted any level of 'personhood', and that, while development, design and distribution of AI should fully comply with all applicable international and domestic law, there is much work to be done in defining and implementing the relevant legislation. Legal issues fall into a few categories: legal status, governmental use (transparency, individual rights), legal accountability for harm, and transparency, accountability, and verifiability. The IEEE suggest that AI should remain subject to the applicable regimes of property law; that stakeholders should identify the types of decisions that should never be delegated to AI, and ensure effective human control over those decisions via rules and standards; that existing laws should be scrutinised and reviewed for mechanisms that could practically give AI legal autonomy; and that manufacturers and operators should be required to comply with the applicable laws of all jurisdictions in which an AI could operate. They also recommend that governments reassess the legal status for AI as they become more sophisticated, and work closely with regulators, societal and industry actors and other stakeholders to ensure that the interests of humanity – and not the development of systems themselves – remain the guiding principle.

### Control and the ethical use – or misuse – of AI

With more sophisticated and complex new AI come more sophisticated and complex possibilities for misuse. Personal data may be used maliciously or for profit, systems are at risk of hacking, and technology may be used exploitatively. This ties into informed use and public awareness: as we enter a new age of AI, with new systems and technology emerging that have never before been implemented, citizens must be kept up to date of the risks that may come with either the use or misuse of these.

### Responsible research and innovation (RRI)

RRI is a growing area, especially in the EU, that draws from classical ethics to provide tools with which to address ethical concerns from the very outset of a project. When incorporated into a project's design phase, RRI increases the chances of design being both relevant and strong in terms of ethical alignment. Many research funders and organisations include RRI in their mission statements and within their research and innovation efforts (IEEE, 2019).

The **IEEE** suggests new ways of educating the public on ethics and security issues, for example a 'data privacy' warning on smart devices that collect personal data; delivering this education in scalable, effective ways; and educating government, lawmakers, and enforcement agencies surrounding these issues, so they can work collaboratively with citizens – in a similar way to police officers providing safety lectures in schools – and avoid fear and confusion (IEEE, 2019).

Other issues include manipulation of behaviour and data. Humans must retain control over AI and oppose subversion. Most initiatives reviewed flag this as a potential issue facing AI as it develops, and flag that AI must behave in a way that is predictable and reliable, with appropriate means for redress, and be subject to validation and testing. AI must also work for the good of humankind, must not exploit people, and be regularly reviewed by human experts.

### Environmental harm and sustainability

The production, management, and implementation of AI must be sustainable and avoid environmental harm. This also ties in to the concept of well-being; a key recognised aspect of well-being is environmental, concerning the air, biodiversity, climate change, soil and water quality, and so on (IEEE, 2019). The **IEEE** (EAD, 2019) state that AI must do no harm to Earth's natural systems or exacerbate their degradation, and contribute to realising sustainable stewardship, preservation, and/or the restoration of Earth's natural systems. The **UNI Global Union** state that AI must put people and the planet first, striving to protect and even enhance our planet's biodiversity and ecosystems (UNI Global Union, n.d.). The **Foundation for Responsible Robotics** identifies a number of potential uses for AI in coming years, from agricultural and farming roles to monitoring of climate change and protection of endangered species. These require responsible, informed policies to govern AI and robotics, say the Foundation, to mitigate risk and support ongoing innovation and development.

### Informed use: public education and awareness

Members of the public must be educated on the use, misuse, and potential harms of AI, via civic participation, communication, and dialogue with the public. The issue of consent – and how much an individual may reasonably and knowingly give – is core to this. For example, the **IEEE** raise several instances in which consent is less clear-cut than might be ethical: what if one's personal data are used to make inferences they are uncomfortable with or unaware of? Can consent be given when a system does not directly interact with an individual? This latter issue has been named the 'Internet of Other People's Things' (IEEE, 2019). Corporate environments also raise the issue of power imbalance; many employees do not have clear consent on how their personal data – including those on health – is used by their employer. To remedy this, the IEEE (2017) suggest employee data impact assessments to deal with these corporate nuances and ensure that no data is collected without employee consent. Data must also be only gathered and used for specific, explicitly stated, legitimate purposes, kept up-to-date, lawfully processed, and not kept for a longer period than necessary. In cases where subjects do not have a direct relationship with the system gathering data, consent must be dynamic, and the system designed to interpret data preferences and limitations on collection and use.

### Personhood and AI

The issue of whether or not an AI deserves 'personhood' ties into debates surrounding accountability, autonomy, and responsibility: is it the AI itself that is responsible for its actions and consequences, or the person(s) who built them?

This concept, rather than allowing robots to be considered people in a human sense, would place robots on the same legal level as corporations. It is worth noting that corporations' legal personhood can currently shield the natural persons behind them from the implications of the law. However, **The UNI Global Union** asserts that legal responsibility lies with the creator, not the robot itself, and calls for a ban on attributing responsibility to robots.

To increase awareness and understanding of AI, undergraduate and postgraduate students must be educated on AI and its relationship to sustainable human development, say the IEEE. Specifically, curriculum and core competencies should be defined and prepared; degree programmes focusing on engineering in international development and humanitarian relief should be exposed to the potential of AI applications; and awareness should be increased of the opportunities and risks faced by Lower Middle Income Countries in the implementation of AI in humanitarian efforts across the globe.

Many initiatives focus on this, including the **Foundation for Responsible Robotics, Partnership on AI, Japanese Society for AI Ethical Guidelines, Future Society** and **AI Now Institute**; these and others maintain that clear, open and transparent dialogue between AI and society is key to the creation of understanding, acceptance, and trust.

### Existential risk

According to the Future of Life Institute, the main existential issue surrounding AI 'is not malevolence, but competence' – AI will continually learn as they interact with others and gather data, leading them to gain intelligence over time and potentially develop aims that are at odds with those of humans.

*'You're probably not an evil ant-hater who steps on ants out of malice,' 'but if you're in charge of a hydroelectric green energy project and there's an anthill in the region to be flooded, too bad for the ants. A key goal of AI safety research is to never place humanity in the position of those ants'* (The Future of Life Institute, 2019).

AI also poses a threat in the form of **autonomous weapons systems (AWS)**. As these are designed to cause physical harm, they raise numerous ethical quandaries. The IEEE (2019) lays out a number of recommendations to ensure that AWS are subject to meaningful human control: they suggest audit trails to guarantee accountability and control; adaptive learning systems that can explain their reasoning in a transparent, understandable way; that human operators of autonomous systems are identifiable, held responsible, and aware of the implications of their work; that autonomous behaviour is predictable; and that professional codes of ethics are developed to address the development of autonomous systems – especially those intended to cause harm. The pursuit of AWS may lead to an international arms race and geopolitical stability; as such, the IEEE recommend that systems designed to act outside the boundaries of human control or judgement are unethical and violate fundamental human rights and legal accountability for weapons use.

Given their potential to seriously harm society, these concerns must be controlled for and regulated pre-emptively, says the **Foundation for Responsible Robotics**. Other initiatives that cover this risk explicitly include the **UNI Global Union** and the **Future of Life Institute**, the latter of which cautions against an arms race in lethal autonomous weapons, and calls for planning and mitigation efforts for possible longer-term risks. We must avoid strong assumptions on the upper limits of future AI capabilities, assert the FLI's **Asilomar Principles**, and recognise that advanced AI represents a profound change in the history of life on Earth.

## 3.3. Case studies

### 3.3.1. Case study: healthcare robots

Artificial Intelligence and robotics are rapidly moving into the field of healthcare and will increasingly play roles in diagnosis and clinical treatment. For example, currently, or in the near future, robots will help in the diagnosis of patients; the performance of simple surgeries; and the monitoring of patients' health and mental wellness in short and long-term care facilities. They may also provide basic physical interventions, work as companion carers, remind patients to take their

medications, or help patients with their mobility. In some fundamental areas of medicine, such as medical image diagnostics, machine learning has been proven to match or even surpass our ability to detect illnesses.

Embodied AI, or robots, are already involved in a number of functions that affect people's physical safety. In June 2005, a surgical robot at a hospital in Philadelphia malfunctioned during prostate surgery, injuring the patient. In June 2015, a worker at a Volkswagen plant in Germany was crushed to death by a robot on the production line. In June 2016, a Tesla car operating in autopilot mode collided with a large truck, killing the car's passenger (Yadron and Tynan, 2016).

As robots become more prevalent, the potential for future harm will increase, particularly in the case of driverless cars, assistive robots and drones, which will face decisions that have real consequences for human safety and well-being. The stakes are much higher with embodied AI than with mere software, as robots have moving parts in physical space (Lin et al., 2017). Any robot with moving physical parts poses a risk, especially to vulnerable people such as children and the elderly.

## Safety

Again, perhaps the most important ethical issue arising from the growth of AI and robotics in healthcare is that of safety and avoidance of harm. It is vital that robots should not harm people, and that they should be safe to work with. This point is especially important in areas of healthcare that deal with vulnerable people, such as the ill, elderly, and children.

Digital healthcare technologies offer the potential to improve accuracy of diagnosis and treatments, but to thoroughly establish a technology's long-term safety and performance investment in clinical trials is required. The debilitating side-effects of vaginal mesh implants and the continued legal battles against manufacturers (The Washington Post, 2019), stand as an example against shortcircuiting testing, despite the delays this introduces to innovating healthcare. Investment in clinical trials will be essential to safely implement the healthcare innovations that AI systems offer.

## User understanding

The correct application of AI by a healthcare professional is important to ensure patient safety. For instance, the precise surgical robotic assistant 'the da Vinci' has proven a useful tool in minimising surgical recovery, but requires a trained operator (The Conversation, 2018).

A shift in the balance of skills in the medical workforce is required, and healthcare providers are preparing to develop the digital literacy of their staff over the next two decades (NHS' Topol Review, 2009). With genomics and machine learning becoming embedded in diagnoses and medical decision-making, healthcare professionals need to become digitally literate to understand each technological tool and use it appropriately. It is important for users to trust the AI presented but to be aware of each tool's strengths and weaknesses, recognising when validation is necessary. For instance, a generally accurate machine learning study to predict the risk of complications in patients with pneumonia erroneously considered those with asthma to be at low risk. It reached this conclusion because asthmatic pneumonia patients were taken directly to intensive care, and this higher-level care circumvented complications. The inaccurate recommendation from the algorithm was thus overruled (Pulmonology Advisor, 2017).

However, it's questionable to what extent individuals need to understand how an AI system arrived at a certain prediction in order to make autonomous and informed decisions. Even if an in-depth understanding of the mathematics is made obligatory, the complexity and learned nature of machine learning algorithms often prevent the ability to understand how a conclusion has been made from a dataset — a so called 'black box' (Schönberger, 2019). In such cases, one possible route

to ensure safety would be to license AI for specific medical procedures, and to 'disbar' the AI if a certain number of mistakes are made (Hart, 2018).

### Data protection

Personal medical data needed for healthcare algorithms may be at risk. For instance, there are worries that data gathered by fitness trackers might be sold to third parties, such as insurance companies, who could use those data to refuse healthcare coverage (National Public Radio, 2018). Hackers are another major concern, as providing adequate security for systems accessed by a range of medical personnel is problematic (Forbes, 2018).

Pooling personal medical data is critical for machine learning algorithms to advance healthcare interventions, but gaps in information governance form a barrier against responsible and ethical data sharing. Clear frameworks for how healthcare staff and researchers use data, such as genomics, in a way that safeguards patient confidentiality is necessary to establish public trust and enable advances in healthcare algorithms (NHS' Topol Review, 2009).

### Legal responsibility

Although AI promises to reduce the number of medical mishaps, when issues occur, legal liability must be established. If equipment can be proven to be faulty then the manufacturer is liable, but it is often tricky to establish what went wrong during a procedure and whether anyone, medical personnel or machine, is to blame. For instance, there have been lawsuits against the da Vinci surgical assistant (Mercury News, 2017), but the robot continues to be widely accepted (The Conversation, 2018).

In the case of 'black box' algorithms where it is impossible to ascertain how a conclusion is reached, it is tricky to establish negligence on the part of the algorithm's producer (Hart, 2018).

For now, AI is used as an aide for expert decisions, and so experts remain the liable party in most cases. For instance, in the aforementioned pneumonia case, if the medical staff had relied solely on the AI and sent asthmatic pneumonia patients home without applying their specialist knowledge, then that would be a negligent act on their part (Pulmonology Advisor, 2017; International Journal of Law and Information Technology, 2019).

Soon, the omission of AI could be considered negligence. For instance, in less developed countries with a shortage of medical professionals, withholding AI that detects diabetic eye disease and so prevents blindness, because of a lack of ophthalmologists to sign off on a diagnosis, could be considered unethical (The Guardian, 2019; International Journal of Law and Information Technology, 2019).

### Bias

Non-discrimination is one of the fundamental values of the EU (see Article 21 of the EU Charter of Fundamental Rights), but machine learning algorithms are trained on datasets that often have proportionally less data available about minorities, and as such can be biased (Medium, 2014). This can mean that algorithms trained to diagnose conditions are less likely to be accurate for ethnic patients; for instance, in the dataset used to train a model for detecting skin cancer, less than 5 percent of the images were from individuals with dark skin, presenting a risk of misdiagnosis for people of colour (The Atlantic, 2018).

To ensure the most accurate diagnoses are presented to people of all ethnicities, algorithmic biases must be identified and understood. Even with a clear understanding of model design this is a difficult task because of the aforementioned 'black box' nature of machine learning. However, various codes of conduct and initiatives have been introduced to spot biases earlier. For instance,

The Partnership on AI, an ethics-focused industry group was launched by Google, Facebook, Amazon, IBM and Microsoft (The Guardian, 2016) — although, worryingly, this board is not very diverse.

### Equality of access

Digital health technologies, such as fitness trackers and insulin pumps, provide patients with the opportunity to actively participate in their own healthcare. Some hope that these technologies will help to redress health inequalities caused by poor education, unemployment, and so on. However, there is a risk that individuals who cannot afford the necessary technologies or do not have the required 'digital literacy' will be excluded, so reinforcing existing health inequalities (The Guardian, 2019).

The UK's National Health Services' Widening Digital Participation programme is one example of how a healthcare service has tried to reduce health inequalities, by helping millions of people in the UK who lack the skills to access digital health services. Programmes such as this will be critical in ensuring equality of access to healthcare, but also in increasing the data from minority groups needed to prevent the biases in healthcare algorithms discussed above.

### Quality of care

'There is remarkable potential for digital healthcare technologies to improve accuracy of diagnoses and treatments, the efficiency of care, and workflow for healthcare professionals' (NHS' Topol Review, 2019).

If introduced with careful thought and guidelines, companion and care robots, for example, could improve the lives of the elderly, reducing their dependence, and creating more opportunities for social interaction. Imagine a home-care robot that could: remind you to take your medications; fetch items for you if you are too tired or are already in bed; perform simple cleaning tasks; and help you stay in contact with your family, friends and healthcare provider via video link.

However, questions have been raised over whether a 'cold', emotionless robot can really substitute for a human's empathetic touch. This is particularly the case in long-term caring of vulnerable and often lonely populations, who derive basic companionship from caregivers. Human interaction is particularly important for older people, as research suggests that an extensive social network offers protection against dementia. At present, robots are far from being real companions. Although they can interact with people, and even show simulated emotions, their conversational ability is still extremely limited, and they are no replacement for human love and attention. Some might go as far as saying that depriving the elderly of human contact is unethical, and even a form of cruelty.

And does abandoning our elderly to cold machine care objectify (degrade) them, or human caregivers? It's vital that robots don't make elderly people feel like objects, or with even less control over their lives than when they were dependent on humans — otherwise they may feel like they are 'lumps of dead matter: to be pushed, lifted, pumped or drained, without proper reference to the fact that they are sentient beings' (Kitwood 1997).

In principle, autonomy, dignity and self-determination can all be thoroughly respected by a machine application, but it's unclear whether application of these roles in the sensitive field of medicine will be deemed acceptable. For instance, a doctor used a telepresence device to give a prognosis of death to a Californian patient; unsurprisingly the patient's family were outraged by this impersonal approach to healthcare (The Independent, 2019). On the other hand, it's argued that new technologies, such as health monitoring apps, will free up staff time for more direct interactions with patients, and so potentially increase the overall quality of care (The Guardian, Press Association, Monday 11 February 2019).

## Deception

A number of 'carebots' are designed for social interactions and are often touted to provide an emotional therapeutic role. For instance, care homes have found that a robotic seal pup's animal-like interactions with residents brightens their mood, decreases anxiety and actually increases the sociability of residents with their human caregivers. However, the line between reality and imagination is blurred for dementia patients, so is it dishonest to introduce a robot as a pet and encourage a social-emotional involvement? (KALW, 2015) And if so, is it morally justifiable?

Companion robots and robotic pets could alleviate loneliness amongst older people, but this would require them believing, in some way, that a robot is a sentient being who cares about them and has feelings — a fundamental deception. Turkle et al. (2006) argue that 'the fact that our parents, grandparents and children might say 'I love you' to a robot who will say 'I love you' in return, does not feel completely comfortable; it raises questions about the kind of authenticity we require of our technology'. Wallach and Allen (2009) agree that robots designed to detect human social gestures and respond in kind all use techniques that are arguably forms of deception. For an individual to benefit from owning a robot pet, they must continually delude themselves about the real nature of their relation with the animal. What's more, encouraging elderly people to interact with robot toys has the effect of infantilising them.

## Autonomy

It's important that healthcare robots actually benefit the patients themselves, and are not just designed to reduce the care burden on the rest of society — especially in the case of care and companion AI. Robots could empower disabled and older people and increase their independence; in fact, given the choice, some might prefer robotic over human assistance for certain intimate tasks such as toileting or bathing. Robots could be used to help elderly people live in their own homes for longer, giving them greater freedom and autonomy. However, how much control, or autonomy, should a person be allowed if their mental capability is in question? If a patient asked a robot to throw them off the balcony, should the robot carry out that command?

## Liberty and privacy

As with many areas of AI technology, the privacy and dignity of users' needs to be carefully considered when designing healthcare service and companion robots. Working in people's homes means that robots will be privy to private moments such as bathing and dressing; if these moments are recorded, who should have access to the information, and how long should recordings be kept? The issue becomes more complicated if an elderly person's mental state deteriorates and they become confused — someone with Alzheimer's could forget that a robot was monitoring them, and could perform acts or say things thinking that they are in the privacy of their own home. Home-care robots need to be able to balance their user's privacy and nursing needs, for example by knocking and awaiting an invitation before entering a patient's room, except in a medical emergency.

To ensure their charge's safety, robots might sometimes need to act as supervisors, restricting their freedoms. For example, a robot could be trained to intervene if the cooker was left on, or the bath was overflowing. Robots might even need to restrain elderly people from carrying out potentially dangerous actions, such as climbing up on a chair to get something from a cupboard. Smart homes with sensors could be used to detect that a person is attempting to leave their room, and lock the door, or call staff — but in so doing the elderly person would be imprisoned.

## Moral agency

*'There's very exciting work where the brain can be used to control things, like maybe they've lost the use of an arm...where I think the real concerns lie is with things like behavioural targeting: going straight to the hippocampus and people pressing 'consent', like we do now, for data access'. (John Havens)*

Robots do not have the capacity for ethical reflection or a moral basis for decision-making, and thus humans must currently hold ultimate control over any decision-making. An example of ethical reasoning in a robot can be found in the 2004 dystopian film 'I, Robot', where Will Smith's character disagreed with how the robots of the fictional time used cold logic to save his life over that of a child's. If more automated healthcare is pursued, then the question of moral agency will require closer attention. Ethical reasoning is being built into robots, but moral responsibility is about more than the application of ethics — and it is unclear whether robots of the future will be able to handle the complex moral issues in healthcare (Goldhill, 2016).

## Trust

Larosa and Danks (2018) write that AI may affect human-human interactions and relationships within the healthcare domain, particularly that between patient and doctor, and potentially disrupt the trust we place in our doctor.

'Psychology research shows people mistrust those who make moral decisions by calculating costs and benefits — like computers do' (The Guardian, 2017). Our distrust of robots may also come from the number of robots running amok in dystopian science fiction. News stories of computer mistakes — for instance, of an image-identifying algorithm mistaking a turtle for a gun (The Verge, 2017) — alongside worries over the unknown, privacy and safety are all reasons for resistance against the uptake of AI (Global News Canada, 2016).

Firstly, doctors are explicitly certified and licensed to practice medicine, and their license indicates that they have specific skills, knowledge, and values such as 'do no harm'. If a robot replaces a doctor for a particular treatment or diagnostic task, this could potentially threaten patient-doctor trust, as the patient now needs to know whether the system is appropriately approved or 'licensed' for the functions it performs.

Secondly, patients trust doctors because they view them as paragons of expertise. If doctors were seen as 'mere users' of the AI, we would expect their role to be downgraded in the public's eye, undermining trust.

Thirdly, a patient's experiences with their doctor are a significant driver of trust. If a patient has an open line of communication with their doctor, and engages in conversation about care and treatment, then the patient will trust the doctor. Inversely, if the doctor repeatedly ignores the patient's wishes, then these actions will have a negative impact on trust. Introducing AI into this dynamic could increase trust — if the AI reduced the likelihood of misdiagnosis, for example, or improved patient care. However, AI could also decrease trust if the doctor delegated too much diagnostic or decision-making authority to the AI, undercutting the position of the doctor as an authority on medical matters.

As the body of evidence grows to support the therapeutic benefits for each technological approach, and as more robotic interacting systems enter the marketplace, then trust in robots is likely to increase. This has already happened for robotic healthcare systems such as the da Vinci surgical robotic assistant (The Guardian, 2014).

## Employment replacement

As in other industries, there is a fear that emerging technologies may threaten employment (The Guardian, 2017), for instance, there are carebots now available that can perform up to a third of nurses' work (Tech Times, 2018). Despite these fears, the NHS' Topol Review (2009) concluded that 'these technologies will not replace healthcare professionals but will enhance them ('augment them'), giving them more time to care for patients'. The review also outlined how the UK's NHS will nurture a learning environment to ensure digitally capable employees.

### 3.3.2 Case study: Autonomous Vehicles

Autonomous Vehicles (AVs) are vehicles that are capable of sensing their environment and operating with little to no input from a human driver. While the idea of self-driving cars has been around since at least the 1920s, it is only in recent years that technology has developed to a point where AVs are appearing on public roads.

According to automotive standardisation body SAE International (2018), there are six levels of driving automation:

0	No automation	An automated system may issue warnings and/or momentarily intervene in driving, but has no sustained vehicle control.
1	Hands on	The driver and automated system share control of the vehicle. For example, the automated system may control engine power to maintain a set speed (e.g. Cruise Control), engine and brake power to maintain and vary speed (e.g. Adaptive Cruise Control), or steering (e.g. Parking Assistance). The driver must be ready to retake full control at any time.
2	Hands off	The automated system takes full control of the vehicle (including accelerating, braking, and steering). However, the driver must monitor the driving and be prepared to intervene immediately at any time.
3	Eyes off	The driver can safely turn their attention away from the driving tasks (e.g. to text or watch a film) as the vehicle will handle any situations that call for an immediate response. However, the driver must still be prepared to intervene, if called upon by the AV to do so, within a timeframe specified by the AV manufacturer.
4	Minds off	As level 3, but no driver attention is ever required for safety, meaning the driver can safely go to sleep or leave the driver's seat.
5	Steering wheel optional	No human intervention is required at all. An example of a level 5 AV would be a robotic taxi.

Some of the lower levels of automation are already well-established and on the market, while higher level AVs are undergoing development and testing. However, as we transition up the levels and put more responsibility on the automated system than the human driver, a number of ethical issues emerge.

#### Societal and Ethical Impacts of AVs

*'We cannot build these tools saying, 'we know that humans act a certain way, we're going to kill them – here's what to do!'.' (John Havens)*

#### *Public safety and the ethics of testing on public roads*

At present, cars with 'assisted driving' functions are legal in most countries. Notably, some Tesla models have an Autopilot function, which provides level 2 automation (Tesla, nd). Drivers are legally allowed to use assisted driving functions on public roads provided they remain in charge of the

vehicle at all times. However, many of these assisted driving functions have not yet been subject to independent safety certification, and as such may pose a risk to drivers and other road users. In Germany, a report published by the Ethics Commission on Automated Driving highlights that it is the public sector's responsibility to guarantee the safety of AV systems introduced and licensed on public roads, and recommends that all AV driving systems be subject to official licensing and monitoring (Ethics Commission, 2017).

In addition, it has been suggested that the AV industry is entering its most dangerous phase, with cars being not yet fully autonomous but human operators not being fully engaged (Solon, 2018). The risks this poses have been brought to widespread attention following the first pedestrian fatality involving an autonomous car. The tragedy took place in Arizona, USA, in May 2018, when a level 3 AV being tested by Uber collided with 49-year-old Elaine Herzberg as she was walking her bike across a street one night. It was determined that Uber was 'not criminally liable' by prosecutors (Shepherdson and Somerville, 2019), and the US National Transportation Safety Board's preliminary report (NTSB, 2018), which drew no conclusions about the cause, said that all elements of the self-driving system were operating normally at the time of the crash. Uber said that the driver is relied upon to intervene and take action in situations requiring emergency braking – leading some commentators to call out the misleading communication to consumers around the terms 'self-driving cars' and 'autopilot' (Leggett, 2018). The accident also caused some to condemn the practice of testing AV systems on public roads as dangerous and unethical, and led Uber to temporarily suspend its self-driving programme (Bradshaw, 2018).

This issue of human safety — of both public and passenger — is emerging as a key issue concerning self-driving cars. Major companies — Nissan, Toyota, Tesla, Uber, Volkswagen — are developing autonomous vehicles capable of operating in complex, unpredictable environments without direct human control, and capable of learning, inferring, planning and making decisions.

Self-driving vehicles could offer multiple benefits: statistics show you're almost certainly safer in a car driven by a computer than one driven by a human. They could also ease congestion in cities, reduce pollution, reduce travel and commute times, and enable people to use their time more productively. However, they won't mean the end of road traffic accidents. Even if a self-driving car has the best software and hardware available, there is still a collision risk. An autonomous car could be surprised, say by a child emerging from behind a parked vehicle, and there is always the issue of *how: how should such cars be programmed when they must decide whose safety to prioritise?*

Driverless cars may also have to choose between the safety of passengers and other road users. Say that a car travels around a corner where a group of school children are playing; there is not enough time to stop, and the only way the car can avoid hitting the children is to swerve into a brick wall — endangering the passenger. Whose safety should the car prioritise: the children's, or the passenger's?

#### *Processes and technologies for accident investigation*

AVs are complex systems that often rely on advanced machine learning technologies. Several serious accidents have already occurred, including a number of fatalities involving level 2 AVs:

- In January 2016, 23-year-old Gao Yaning died when his Tesla Model S crashed into the back of a road-sweeping truck on a highway in Hebei, China. The family believe Autopilot was engaged when the accident occurred and accuse Tesla of exaggerating the system's capabilities. Tesla state that the damage to the vehicle made it impossible to determine whether Autopilot was engaged and, if so, whether it malfunctioned. A civil case into the crash is ongoing, with a third-party appraiser reviewing data from the vehicle (Curtis, 2016).

- In May 2016, 40-year-old Joshua Brown died when his Tesla Model S collided with a truck while Autopilot was engaged in Florida, USA. An investigation by the National Highways and Transport Safety Agency found that the driver, and not Tesla, were at fault (Gibbs, 2016). However, the National Highway Traffic Safety Administration later determined that both Autopilot and over-reliance by the motorist on Tesla's driving aids were to blame (Felton, 2017).
- In March 2018, Wei Huang was killed when his Tesla Model X crashed into a highway safety barrier in California, USA. According to Tesla, the severity of the accident was 'unprecedented'. The National Transportation Safety Board later published a report attributing the crash to an Autopilot navigation mistake. Tesla is now being sued by the victim's family (O'Kane, 2018).

Unfortunately, efforts to investigate these accidents have been stymied by the fact that standards, processes, and regulatory frameworks for investigating accidents involving AVs have not yet been developed or adopted. In addition, the proprietary data logging systems currently installed in AVs mean that accident investigators rely heavily on the cooperation of manufacturers to provide critical data on the events leading up to an accident (Stilgoe and Winfield, 2018).

One solution is to fit all future AVs with industry standard event data recorders — a so-called 'ethical black box' — that independent accident investigators could access. This would mirror the model already in place for air accident investigations (Sample, 2017).

### Near-miss accidents

At present, there is no system in place for the systematic collection of near-miss accidents. While it is possible that manufacturers are collecting this data already, they are not under any obligation to do so — or to share the data. The only exception at the moment is the US state of California, which requires all companies that are actively testing AVs on public roads to disclose the frequency at which human drivers were forced to take control of the vehicle for safety reasons (known as 'disengagement').

In 2018, the number of disengagements by AV manufacturer varied significantly, from one disengagement for every 11,017 miles driven by Waymo AVs to one for every 1.15 miles driven by Apple AVs (Hawkins, 2019). Data on these disengagements reinforces the importance of ensuring that human safety drivers remain engaged. However, the Californian data collection process has been criticised, with some claiming its ambiguous wording and lack of strict guidelines enables companies to avoid reporting certain events that could be termed near-misses.

Without access to this type of data, policymakers cannot account for the frequency and significance of near-miss accidents, or assess the steps taken by manufacturers as a result of these near-misses. Again, lessons could be learned from the model followed in air accident investigations, in which all near misses are thoroughly logged and independently investigated. Policymakers require comprehensive statistics on all accidents and near-misses in order to inform regulation.

### Data privacy

It is becoming clear that manufacturers collect significant amounts of data from AVs. As these vehicles become increasingly common on our roads, the question emerges: to what extent are these data compromising the privacy and data protection rights of drivers and passengers?

Already, data management and privacy issues have appeared, with some raising concerns about the potential misuse of AV data for advertising purposes (Lin, 2014). Tesla have also come under fire for the unethical use of AV data logs. In an investigation by *The Guardian*, the newspaper found multiple instances where the company shared drivers' private data with the media following crashes, without

their permission, to prove that its technology was not responsible (Thielman, 2017). At the same time, Tesla does not allow customers to see their own data logs.

One solution, proposed by the German Ethics Commission on Automated Driving, is to ensure that all AV drivers be given full data sovereignty (Ethics Commission, 2017). This would allow them to control how their data is used.

## Employment

The growth of AVs is likely to put certain jobs — most pertinently bus, taxi, and truck drivers — at risk.

In the medium term, truck drivers face the greatest risk as long-distance trucks are at the forefront of AV technology (Viscelli, 2018). In 2016, the first commercial delivery of beer was made using a self-driving truck, in a journey covering 120 miles and involving no human action (Isaac, 2016). Last year saw the first fully driverless trip in a self-driving truck, with the AV travelling seven miles without a single human on board (Cannon, 2018).

Looking further forward, bus drivers are also likely to lose jobs as more and more buses become driverless. Numerous cities across the world have announced plans to introduce self-driving shuttles in the future, including Edinburgh (Calder, 2018), New York (BBC, 2019a) and Singapore (BBC 2017). In some places, this vision has already become a reality; the Las Vegas shuttle famously got off to a bumpy start when it was involved in a collision on its first day of operation (Park, 2017), and tourists in the small Swiss town of Neuhausen Rheinfall can now hop on a self-driving bus to visit the nearby waterfalls (CNN, 2018). In the medium term, driverless buses will likely be limited to routes that travel along 100% dedicated bus lanes. Nonetheless, the advance of self-driving shuttles has already created tensions with organised labour and city officials in the USA (Weinberg, 2019). Last year, the Transport Workers Union of America formed a coalition in an attempt to stop autonomous buses from hitting the streets of Ohio (Pfleger, 2018).

Fully autonomous taxis will likely only become realistic in the long term, once AV technology has been fully tested and proven at levels 4 and 5. Nonetheless, with plans to introduce self-driving taxis in London by 2021 (BBC, 2018), and an automated taxi service already available in Arizona, USA (Sage, 2019), it is easy to see why taxi drivers are uneasy.

## The quality of urban environments

In the long-term, AVs have the potential to reshape our urban environment. Some of these changes may have negative consequences for pedestrians, cyclists and locals. As driving becomes more automated, there will likely be a need for additional infrastructure (e.g. AV-only lanes). There may also be more far-reaching effects for urban planning, with automation shaping the planning of everything from traffic congestion and parking to green spaces and lobbies (Marshall and Davies, 2018). The rollout of AVs will also require that 5G network coverage is extended significantly — again, something with implications for urban planning (Khosravi, 2018).

The environmental impact of self-driving cars should also be considered. While self-driving cars have the potential to significantly reduce fuel usage and associated emissions, these savings could be counteracted by the fact that self-driving cars make it easier and more appealing to drive long distances (Worland, 2016). The impact of automation on driving behaviours should therefore not be underestimated.

### *Legal and ethical responsibility*

From a legal perspective, who is responsible for crashes caused by robots, and how should victims be compensated (if at all) when a vehicle controlled by an algorithm causes injury? If courts cannot resolve this problem, robot manufacturers may incur unexpected costs that would discourage investment. However, if victims are not properly compensated then autonomous vehicles are unlikely to be trusted or accepted by the public.

Robots will need to make judgement calls in conditions of uncertainty, or 'no win' situations. However, which ethical approach or theory should a robot be programmed to follow when there's no legal guidance? As Lin et al. explain, different approaches can generate different results, including the number of crash fatalities.

Additionally, who should choose the ethics for the autonomous vehicle — drivers, consumers, passengers, manufacturers, politicians? Loh and Loh (2017) argue that responsibility should be shared among the engineers, the driver and the autonomous driving system itself.

However, Millar (2016) suggests that the user of the technology, in this case the passenger in the self-driving car, should be able to decide what ethical or behavioural principles the robot ought to follow. Using the example of doctors, who do not have the moral authority to make important decisions on end-of-life care without the informed consent of their patients, he argues that there would be a moral outcry if engineers designed cars without either asking the driver directly for their input, or informing the user ahead of time how the car is programmed to behave in certain situations.

### **3.3.3 Case study: Warfare and weaponisation**

Although partially autonomous and intelligent systems have been used in military technology since at least the Second World War, advances in machine learning and AI signify a turning point in the use of automation in warfare.

AI is already sufficiently advanced and sophisticated to be used in areas such as satellite imagery analysis and cyber defence, but the true scope of applications has yet to be fully realised. A recent report concludes that AI technology has the potential to transform warfare to the same, or perhaps even a greater, extent than the advent of nuclear weapons, aircraft, computers and biotechnology (Allen and Chan, 2017). Some key ways in which AI will impact militaries are outlined below.

### **Ethical dilemmas in development**

In 2014, the Open Roboethics initiative (ORi 2014a, 2014b) conducted a poll asking people what they thought an autonomous car in which they were a passenger should do if a child stepped out in front of the vehicle in a tunnel. The car wouldn't have time to brake and spare the child, but could swerve into the walls of the tunnel, killing the passenger. This is a spin on the classic 'trolley dilemma', where one has the option to divert a runaway trolley from a path that would hurt several people onto the path that would only hurt one.

36 % of participants said that they would prefer the car to swerve into the wall, saving the child; however, the majority (64 %) said they would wish to save themselves, thus sacrificing the child. 44 % of participants thought that the passenger should be able to choose the car's course of action, while 33 % said that lawmakers should choose. Only 12 % said that the car's manufacturers should make the decision. These results suggest that people do not like the idea of engineers making moral decisions on their behalf.

Asking for the passenger's input in every situation would be impractical. However, Millar (2016) suggests a 'setup' procedure where people could choose their ethics settings after purchasing a new car. Nonetheless, choosing how the car reacts in advance could be seen as premeditated harm, if, for example a user programmed their vehicle to always avoid vehicle collisions by swerving into cyclists. This would increase the user's accountability and liability, whilst diverting responsibility away from manufacturers.

## Lethal autonomous weapons

As automatic and autonomous systems have become more capable, militaries have become more willing to delegate authority to them. This is likely to continue with the widespread adoption of AI, leading to an AI inspired arms-race. The Russian Military Industrial Committee has already approved an aggressive plan whereby 30% of Russian combat power will consist of entirely remote-controlled and autonomous robotic platforms by 2030. Other countries are likely to set similar goals. While the United States Department of Defense has enacted restrictions on the use of autonomous and semi-autonomous systems wielding lethal force, other countries and non-state actors may not exercise such self-restraint.

## Drone technologies

Standard military aircraft can cost more than US\$100 million per unit; a high-quality quadcopter Unmanned Aerial Vehicle, however, currently costs roughly US\$1,000, meaning that for the price of a single high-end aircraft, a military could acquire one million drones. Although current commercial drones have limited range, in the future they could have similar ranges to ballistic missiles, thus rendering existing platforms obsolete.

## Robotic assassination

Widespread availability of low-cost, highly-capable, lethal, and autonomous robots could make targeted assassination more widespread and more difficult to attribute. Automatic sniping robots could assassinate targets from afar.

## Mobile-robotic-Improvised Explosive Devices

As commercial robotic and autonomous vehicle technologies become widespread, some groups will leverage this to make more advanced Improvised Explosive Devices (IEDs). Currently, the technological capability to rapidly deliver explosives to a precise target from many miles away is restricted to powerful nation states. However, if long distance package delivery by drone becomes a reality, the cost of precisely delivering explosives from afar would fall from millions of dollars to thousands or even hundreds. Similarly, self-driving cars could make suicide car bombs more frequent and devastating since they no longer require a suicidal driver.

Hallaq et al. (2017) also highlight key areas in which machine learning is likely to affect warfare. They describe an example where a Commanding Officer (CO) could employ an Intelligent Virtual Assistant (IVA) within a fluid battlefield environment that automatically scanned satellite imagery to detect specific vehicle types, helping to identify threats in advance. It could also predict the enemy's intent, and compare situational data to a stored database of hundreds of previous wargame exercises and live engagements, providing the CO with access to a level of accumulated knowledge that would otherwise be impossible to accrue.

Employing AI in warfare raises several **legal and ethical questions**. One concern is that automated weapon systems that exclude human judgment could violate International Humanitarian Law, and threaten our fundamental right to life and the principle of human dignity. AI could also lower the threshold of going to war, affecting global stability.

International Humanitarian law stipulates that any attack needs to distinguish between combatants and non-combatants, be proportional and must not target civilians or civilian objects. Also, no attack should unnecessarily aggravate the suffering of combatants. AI may be unable to fulfil these principles without the involvement of human judgment. In particular, many researchers are concerned that Lethal Autonomous Weapon Systems (LAWS) — a type of autonomous military robot that can independently search for and 'engage' targets using lethal force — may not meet the standards set by International Humanitarian Law, as they are not able to distinguish civilians from

combatants, and would not be able to judge whether the force of the attack was proportional given the civilian damage it would incur.

Amoroso and Tamburini (2016, p. 6) argue that: '[LAWS must be] capable of respecting the principles of distinction and proportionality at least as well as a competent and conscientious human soldier'. However, Lim (2019) points out that while LAWS that fail to meet these requirements should not be deployed, one day LAWS *will* be sophisticated enough to meet the requirements of distinction and proportionality. Meanwhile, Asaro (2012) argues that it doesn't matter how good LAWS get; it is a moral requirement that only a human should initiate lethal force, and it is simply morally wrong to delegate life or death decisions to machines.

Some argue that delegating the decision to kill a human to a machine is an infringement of basic human dignity, as robots don't feel emotion, and can have no notion of sacrifice and what it means to take a life. As Lim et al (2019) explain, 'a machine, bloodless and without morality or mortality, cannot fathom the significance of using force against a human being and cannot do justice to the gravity of the decision'.

Robots also have no concept of what it means to kill the 'wrong' person. 'It is only because humans can feel the rage and agony that accompanies the killing of humans that they can understand sacrifice and the use of force against a human. Only then can they realise the 'gravity of the decision' to kill' (Johnson and Axinn 2013, p. 136).

However, others argue that there is no particular reason why being killed by a machine would be a subjectively worse, or less dignified, experience than being killed by a cruise missile strike. 'What matters is whether the victim experiences a sense of humiliation in the process of getting killed. Victims being threatened with a potential bombing will not care whether the bomb is dropped by a human or a robot' (Lim et al, 2019). In addition, not all humans have the emotional capacity to conceptualise sacrifice or the relevant emotions that accompany risk. In the heat of battle, soldiers rarely have time to think about the concept of sacrifice, or generate the relevant emotions to make informed decisions each time they deploy lethal force.

Additionally, who should be held accountable for the actions of autonomous systems — the commander, programmer, or the operator of the system? Schmit (2013) argues that the responsibility for committing war crimes should fall on both the individual who programmed the AI, and the commander or supervisor (assuming that they knew, or should have known, the autonomous weapon system had been programmed and employed in a war crime, and that they did nothing to stop it from happening).

## 4. AI standards and regulation

A small new generation of ethical standards are emerging as the ethical, legal and societal impacts of artificial intelligence and robotics are further understood. Whether a standard clearly articulates explicit or implicit ethical concerns, all standards embody some kind of ethical principle (Winfield, 2019a). The standards that do exist are still in development and there is limited publicly available information on them.

Perhaps the earliest explicit ethical standard in robotics is BS 8611 Guide to the Ethical Design and Application of Robots and Robotic Systems (British Standard BS 8611, 2016). BS8611 is not a code of practice, but guidance on how designers can identify potential ethical harm, undertake an ethical risk assessment of their robot or AI, and mitigate any ethical risks identified. It is based on a set of 20 distinct ethical hazards and risks, grouped under four categories: societal, application, commercial & financial, and environmental.

Advice on measures to mitigate the impact of each risk is given, along with suggestions on how such measures might be verified or validated. The societal hazards include, for example, loss of trust, deception, infringements of privacy and confidentiality, addiction, and loss of employment. Ethical Risk Assessment should consider also foreseeable misuse, risks leading to stress and fear (and their minimisation), control failure (and associated psychological effect), reconfiguration and linked changes to responsibilities, hazards associated with specific robotics applications. Particular attention is paid to robots that can learn and the implications of robot enhancement that arise, and the standard argues that the ethical risk associated with the use of a robot should not exceed the risk of the same activity when conducted by a human.

British Standard BS 8611 assumes that physical hazards imply ethical hazards, and defines ethical harm as affecting 'psychological and/or societal and environmental well-being.' It also recognises that physical and emotional hazards need to be balanced against expected benefits to the user. The standard highlights the need to involve the public and stakeholders in development of robots and provides a list of key design considerations including:

- Robots should not be designed primarily to kill humans;
- Humans remain responsible agents;
- It must be possible to find out who is responsible for any robot;
- Robots should be safe and fit for purpose;
- Robots should not be designed to be deceptive;
- The precautionary principle should be followed;
- Privacy should be built into the design;
- Users should not be discriminated against, nor forced to use a robot.

Particular guidelines are provided for roboticists, particularly those conducting research. These include the need to engage the public, consider public concerns, work with experts from other disciplines, correct misinformation and provide clear instructions. Specific methods to ensure ethical use of robots include: user validation (to ensure robot can/is operated as expected), software verification (to ensure software works as anticipated), involvement of other experts in ethical assessment, economic and social assessment of anticipated outcomes, assessment of any legal implications, compliance testing against relevant standards. Where appropriate, other guidelines and ethical codes should be taken into consideration in the design and operation of robots (e.g. medical or legal codes relevant in specific contexts). The standard also makes the case that military application of robots does not remove the responsibility and accountability of humans.

The IEEE Standards Association has also launched a standard via its global initiative on the Ethics of Autonomous and Intelligent Systems. Positioning 'human well-being' as a central precept, the IEEE initiative explicitly seeks to reposition robotics and AI as technologies for improving the human condition rather than simply vehicles for economic growth (Winfield, 2019a). Its aim is to educate, train and empower AI/robot stakeholders to 'prioritise ethical considerations so that these technologies are advanced for the benefit of humanity.'

There are currently 14 IEEE standards working groups working on drafting so-called 'human' standards that have implications for artificial intelligence (Table 4.1).

Table 2: IEEE 'human standards' with implications for AI

Standard		Aims/Objectives
P7000	Model Process for Addressing Ethical Concerns During System Design	To establish a process for <b>ethical design of Autonomous and Intelligent Systems</b> .
P7001	Transparency of Autonomous Systems	<p>To ensure the <b>transparency of autonomous systems to a range of stakeholders</b>. It specifically will address:</p> <ul style="list-style-type: none"> <li>• <i>Users</i>: ensuring users understand what the system does and why, with the intention of building trust;</li> <li>• <i>Validation and certification</i>: ensuring the system is subject to scrutiny;</li> <li>• <i>Accidents</i>: enabling accident investigators to undertake investigation;</li> <li>• <i>Lawyers and expert witnesses</i>: ensuring that, following an accident, these groups are able to give evidence;</li> <li>• <i>Disruptive technology (e.g. driverless cars)</i>: enabling the public to assess technology (and, if appropriate, build confidence).</li> </ul>
P7002	Data Privacy Process	To establish standards for <b>the ethical use of personal data</b> in software engineering processes. It will develop and describe privacy impact assessments (PIA) that can be used to identify the need for, and effectiveness of, privacy control measures. It will also provide checklists for those developing software that uses personal information.

P7003	Algorithmic Bias Considerations	To help algorithm developers make explicit the ways in which they have sought to <b>eliminate or minimise the risk of bias</b> in their products. This will address the use of overly subjective information and help developers ensure they are compliant with legislation regarding protected characteristics (e.g. race, gender). It is likely to include: <ul style="list-style-type: none"> <li>• Benchmarking processes for the selection of data sets;</li> <li>• Guidelines on communicating the boundaries for which the algorithm has been designed and validated (guarding against unintended consequences of unexpected uses);</li> <li>• Strategies to avoid incorrect interpretation of system outputs by users.</li> </ul>
P7004	Standard for Child and Student Data Governance	Specifically <b>aimed at educational institutions</b> , this will provide guidance on accessing, collecting, storing, using, sharing and destroying child/student data.
P7005	Standard for Transparent Employer Data Governance	Similar to P7004, but <b>aimed at employers</b> .
P7006	Standard for Personal Data Artificial Intelligence (AI) Agent	Describes the technical elements required to create and grant access to <b>personalised AI</b> . It will enable individuals to safely organise and share their personal information at a machine-readable level, and enable personalised AI to act as a proxy for machine-to-machine decisions.
P7007	Ontological Standard for Ethically Driven Robotics and Automation Systems	This standard brings together engineering and philosophy <b>to ensure that user well-being is considered throughout the product life cycle</b> . It intends to identify ways to maximise benefits and minimise negative impacts, and will also consider the ways in which communication can be clear between diverse communities.

P7008	Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems	Drawing on 'nudge theory', this standard seeks <b>to delineate current or potential nudges that robots or autonomous systems might undertake</b> . It recognises that nudges can be used for a range of reasons, but that they seek to affect the recipient emotionally, change behaviours and can be manipulative, and seeks to elaborate methodologies for ethical design of AI using nudge.
P7009	Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems	To create effective methodologies for <b>the development and implementation of robust, transparent and accountable fail-safe mechanisms</b> . It will address methods for measuring and testing a system's ability to fail safely.
P7010	Well-being Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems	To establish a baseline for metrics used <b>to assess well-being factors that could be affected by autonomous systems</b> , and for how human well-being could proactively be improved.
P7011	Standard for the Process of Identifying and Rating the Trustworthiness of News Sources	Focusing on news information, this standard sets out <b>to standardise the processes for assessing the factual accuracy of news stories</b> . It will be used to produce a 'trustfulness' score. This standard seeks to address the negative effects of unchecked 'fake' news, and is designed to restore trust in news purveyors.
P7012	Standard for Machine Readable Personal Privacy Terms	To establish <b>how privacy terms are presented</b> and how they could be read and accepted by machines.
P7013	Inclusion and Application Standards for Automated Facial Analysis Technology	To provide <b>guidelines on the data used in facial recognition</b> , the requirements for diversity, and benchmarking of applications and situations in which facial recognition should not be used.

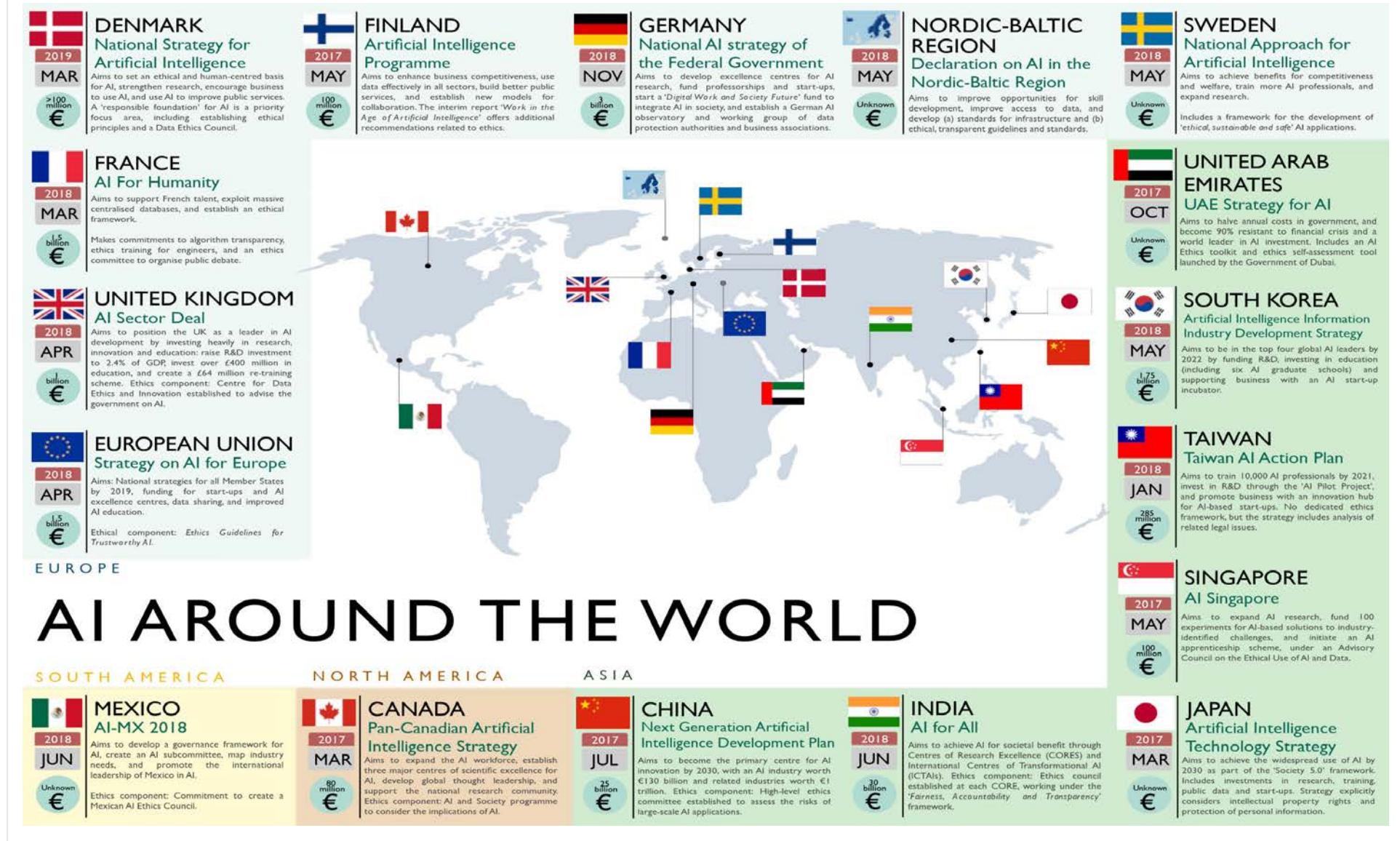
## 5. National and International Strategies on AI

As the technology behind AI continues to progress beyond expectations, policy initiatives are springing up across the globe to keep pace with these developments.

The first national strategy on AI was launched by Canada in March 2017, followed soon after by technology leaders Japan and China. In Europe, the European Commission put forward a communication on AI, initiating the development of independent strategies by Member States. An American AI initiative is expected soon, alongside intense efforts in Russia to formalise their 10-point plan for AI.

These initiatives differ widely in terms of their goals, the extent of their investment, and their commitment to developing ethical frameworks, reviewed here as of May 2019.

Figure 3: National and International Strategies on AI published as of May 2019.



## 5.1. Europe

The European Commission's Communication on Artificial Intelligence (European Commission, 2018a), released in April 2018, paved the way to the first international strategy on AI. The document outlines a coordinated approach to maximise the benefits, and address the challenges, brought about by AI.

The Communication on AI was formalised nine months later with the presentation of a coordinated plan on AI (European Commission, 2018b). The plan details seven objectives, which include financing start-ups, investing €1.5 billion in several 'research excellence centres', supporting masters and PhDs in AI and creating common European data spaces.

Objective 2.6 of the plan is to develop 'ethics guidelines with a global perspective'. The Commission appointed an independent high-level expert group to develop their ethics guidelines, which – following consultation – were published in their final form in April 2019 (European Commission High-Level Expert Group on Artificial Intelligence, 2019). The Guidelines list key requirements that AI systems must meet in order to be trustworthy.

The EU's seven requirements for trustworthy AI:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental wellbeing
7. Accountability

*Source: European Commission High-Level Expert Group on Artificial Intelligence, 2019*

The EU's High-Level Expert Group on AI shortly after released a further set of policy and investment guidelines for trustworthy AI (European Commission High-Level Expert Group on AI, 2019b), which includes a number of important recommendations around protecting people, boosting uptake of AI in the private sector, expanding European research capacity in AI and developing ethical data management practices.

The Council of Europe also has various ongoing projects regarding the application of AI and in September 2019 established an Ad Hoc Committee on Artificial Intelligence (CAHAI). The committee will assess the potential elements of a legal framework for the development and application of AI, based on the Council's founding principles of human rights, democracy and the rule of law (Council of Europe, 2019a).

Looking ahead, the next European Commission President, Ursula von der Leyen, has announced AI as a priority for the next Commission, including legislation for a coordinated approach on the 'human and ethical implications' of AI (Kayali, 2019; von der Leyen, 2019).

The European Commission provides a unifying framework for AI development in the EU, but Member States are also required to develop their own national strategies.

**Finland** was the first Member State to develop a national programme on AI (Ministry of Economic Affairs and Employment of Finland, 2018a). The programme is based on two reports, *Finland's Age of Artificial Intelligence* and *Work in the Age of Artificial Intelligence* (Ministry of Economic Affairs and Employment of Finland, 2017, 2018b). Policy objectives focus on investment for business competitiveness and public services. Although recommendations have already been incorporated into policy, Finland's AI steering group will run until the end of the present Government's term, with a final report expected imminently.

So far, Denmark, France, Germany, Sweden and the UK have also announced national initiatives on AI. **Denmark**'s National Strategy for Artificial Intelligence (The Danish Government, 2019) was released in March 2019 and follows its 'Strategy for Digital Growth' (The Danish Government, 2018). This comprehensive framework lists objectives including establishing a responsible foundation for AI, providing high quality data and overall increasing investment in AI (particularly in the agriculture, energy, healthcare and transport sectors). There is a strong focus on data ethics, including responsibility, security and transparency, and recognition of the need for an ethical framework. The Danish government outlines six principles for ethical AI – self-determination, dignity, responsibility, explainability, equality and justice, and development (solutions that support ethically responsible development and use of AI in order to achieve societal progress) – and will establish a Data Ethics Council to monitor technological development in the country.

In **France**, 'AI for Humanity' was launched in March 2018 and makes commitments to support French talent, make better use of data and also establish an ethical framework on AI (AI For Humanity, 2018). President Macron has committed to ensuring transparency and fair use in AI, which will be embedded in the education system. The strategy is mainly based on the work of Cédric Villani, French mathematician and politician, whose 2018 report on AI made recommendations across economic policy, research infrastructure, employment and ethics (Villani, 2018).

**Germany**'s AI Strategy was adopted soon after in November 2018 (Die Bundesregierung, 2018) and makes three major pledges: to make Germany a global leader in the development and use of AI, to safeguard the responsible development and use of AI, and to integrate AI in society in ethical, legal, cultural and institutional terms. Individual objectives include developing Centres of Excellence for research, the creation of 100 extra professorships for AI, establishing a German AI observatory, funding 50 flagship applications of AI to benefit the environment, developing guidelines for AI that are compatible with data protection laws, and establishing a 'Digital Work and Society Future Fund' (De.digital, 2018).

**Sweden**'s approach to AI (Government Offices of Sweden, 2018) has less specific terms, but provides general guidance on education, research, innovation and infrastructure for AI. Recommendations include building a strong research base, collaboration between sectors and with other countries, developing efforts to prevent and manage risk and developing standards to guide the ethical use of AI. A Swedish AI Council, made up of experts from industry and academia, has also been established to develop a 'Swedish model' for AI, which they say will be sustainable, beneficial to society and promote long-term economic growth (Swedish AI Council, 2019).

The **UK** government issued the comprehensive 'AI Sector Deal' in April 2018 (GOV.UK, 2018), part of a larger 'Industrial Strategy', which sets out to increase productivity by investing in business, skills and infrastructure (GOV.UK, 2019). It pledges almost £1 billion to promote AI in the UK, along five key themes: ideas, people, infrastructure, business environment and places.

Key policies include increasing research and development investment to a total of 2.4% of GDP by 2027; investing over £400 million in maths, digital and technical education; developing a national retraining scheme to plug the skills gap and investing in digital infrastructure such as electric

vehicles and fibre networks. As well as these investment commitments, included in the deal is the creation of a 'Centre for Data Ethics and Innovation' (CDEI) to ensure the safe and ethical use of AI. First announced in the 2017 budget, the CDEI will assess the risks of AI, review regulatory and governance frameworks and advise the government and technology creators on best practice (UK Government Department for Digital, Culture, Media & Sport, 2019).

Several other European nations are well on their way to releasing national strategies. **Austria** has established a 'Robot Council' to help the Government to develop a national AI Strategy (Austrian Council on Robotics and Artificial Intelligence, 2019). A white paper prepared by the Council lays the groundwork for the strategy. The socially-focused document includes objectives to promote the responsible use of AI, develop measures to recognise and mitigate hazards, create a legal framework to protect data security, and engender a public dialogue around the use of AI (Austrian Council on Robotics and Artificial Intelligence, 2018).

**Estonia** has traditionally been quick to take up new technologies, AI included. In 2017, Estonia's Adviser for Digital Innovation Marten Kaevats described AI as the next step for 'e-governance' in Estonia (Plantera, 2017). Indeed, AI is already widely used by the government, which is currently devising a national AI strategy (Castellanos, 2018). The plan will reportedly consider the ethical implications of AI, alongside offering practical economic incentives and pilot programmes.

An AI task force has been established by **Italy** (Agency for Digital Italy, 2019) to identify the opportunities offered by AI and improve the quality of public services. Their white paper (Task Force on Artificial Intelligence of the Agency for Digital Italy, 2018), published in March 2018, describes ethics as the first challenge to the successful implementation of AI, stating a need to uphold the principle that AI should be at the service of the citizen and to ensure equality by using technology to address universal needs. The task force further outline challenges relating to technology development, the skills gap, data accessibility and quality, and a legal framework. It makes a total of 10 recommendations to government, which are yet to be realised by policy.

**Malta**, a country that has previously focused heavily on blockchain technology, has now made public its plans to develop a national AI strategy, putting Malta 'amongst the top 10 nations with a national strategy for AI' (Malta AI, 2019). A task force has been established composed of industry representatives, academics and other experts to help devise a policy for Malta that will focus on an ethical, transparent and socially-responsible AI while developing measures that garner foreign investment, which will include developing the skillset and infrastructure needed to support AI in Malta.

**Poland** too is working on its national AI strategy. A report recently released by the Digital Poland Foundation (2019) focuses on the AI ecosystem in Poland, as a forerunner of the national AI strategy. Although it provides a comprehensive overview of the state-of-the-art in Poland, it does not make specific recommendations for government, and makes no reference to the ethical issues surrounding AI.

Despite media reports of military-focused AI developments in **Russia** (Apps, 2019; Bershidski, 2017; Le Miere, 2017; O'Connor, 2017) the country currently has no national strategy on AI. Following the 2018 conference 'Artificial Intelligences: Problems and Solutions', the Russian Ministry of Defence released a list of policy recommendations, which include creating a state system for AI education and a national centre for AI. The latest reports suggest President Putin has set a deadline of June 15<sup>th</sup> 2019 for his government to finalise the national strategy on AI.

### 5.1.1. Across the EU: Public attitudes to robots and digitisation

Overall, surveys of European perspectives to AI, robotics, and advanced technology (European Commission 2012; European Commission 2017) have reflected that citizens hold a generally positive view of these developments, viewing them as a positive addition to society, the economy, and citizens' lives. However, this attitude varies by age, gender, educational level, and location and is largely dependent on one's exposure to robots and relevant information — for example, only small numbers of those surveyed actually had experience of using a robot (past or present), and those with experience were more likely to view them positively than those without.

General trends in public perception from these surveys showed that respondents were:

- Supportive of using robots and digitisation in jobs that posed risk or difficulty to humans (such as space exploration, manufacturing and the military);
- Concerned that such technology requires effective and careful management;
- Worried that automation and digitisation would bring job losses, and unsure whether it would stimulate and boost job opportunities across the EU;
- Unsupportive of using robots to care for vulnerable members of society (the elderly, ill, dependent pets, or those undergoing medical procedures);
- Worried about accessing and protecting their data and online information, and likely to have taken some form of protective action in this area (antivirus software, changed browsing behaviour);
- Unwilling to drive in a driverless car (only 22% would be happy to do this);
- Distrustful of social media, with only 7% viewing stories published on social media as 'generally trustworthy'; and
- Unlikely to view widespread use of robots as near-term, instead perceiving it to be a scenario that would occur at least 20 years in the future.

These concerns thus feature prominently in European AI initiatives, and are reflective of general opinion on the implementation of robots, AI, automation and digitisation across the spheres of life, work, health, and more.

## 5.2. North America

**Canada** was the first country in the world to launch a national AI strategy, back in March 2017. The Pan-Canadian Artificial Intelligence Strategy (Canadian Institute For Advanced Research, 2017) was established with four key goals, to: increase the number of AI researchers and graduates in Canada; establish centres of scientific excellence (in Edmonton, Montreal and Toronto); develop global thought leadership in the economic, ethical, policy and legal implications of AI; and support a national research community in AI.

A separate programme for AI and society was dedicated to the social implications of AI, led by policy-relevant working groups that publish their findings for both government and public. In collaboration with the French National Centre for Scientific Research (CNRS) and UK Research and Innovation (UKRI), the AI and society programme has recently announced a series of interdisciplinary workshops to explore issues including trust in AI, the impact of AI in the healthcare sector and how AI affects cultural diversity and expression (Canadian Institute For Advanced Research, 2019).

In the **USA**, President Trump issued an Executive Order launching the 'American AI Initiative' in February 2019 (The White House, 2019a), soon followed by the launch of a website uniting all other AI initiatives (The White House, 2019b), including AI for American Innovation, AI for American Industry, AI for the American Worker and AI for American Values. The American AI Initiative has five key areas: investing in R&D, unleashing AI resources (i.e. data and computing power), setting

governance standards, building the AI workforce and international engagement. The Department of Defence has also published its own AI strategy (US Department of Defence, 2018), with a focus on the military capabilities of AI.

In May, the US advanced this with the AI Initiative Act, which will invest \$2.2 billion into developing a national AI strategy, as well as funding federal R&D. The legislation, which seeks to 'establish a coordinated Federal initiative to accelerate research and development on artificial intelligence for the economic and national security of the United States' commits to establishing a National AI Coordination Office, create AI evaluation standards and fund 5 national AI research centres. The programme will also fund the National Science Foundation to research the effects of AI on society, including the roles of data bias, privacy and accountability, and expand AI-based research efforts led by the Department of Energy (US Congress, 2019).

In June 2019, the National Artificial Intelligence Research and Development Strategic Plan was released, which builds on an earlier plan issued by the Obama administration and identifies eight strategic priorities, including making long-term investments in AI research, developing effective methods for human-AI collaboration, developing shared public datasets, evaluating AI technologies through standards and benchmarks, and understanding and addressing the ethical, legal and societal implications of AI. The document provides a coordinated strategy for AI research and development in the US (National Science & Technology Council, 2019).

### 5.3. Asia

Asia has in many respects led the way in AI strategy, with **Japan** being the second country to release a national initiative on AI. Released in March 2017, Japan's AI Technology Strategy (Japanese Strategic Council for AI Technology, 2017) provides an industrialisation roadmap, including priority areas in health and mobility, important with Japan's ageing population in mind. Japan envisions a three-stage development plan for AI, culminating in a completely connected AI ecosystem, working across all societal domains.

**Singapore** was not far behind. In May 2017, AI Singapore was launched, a five-year programme to enhance the country's capabilities in AI, with four key themes: industry and commerce, AI frameworks and testbeds, AI talent and practitioners and R&D (AI Singapore, 2017). The following year the Government of Singapore announced additional initiatives focused around the governance and ethics of AI, including establishing an Advisory Council on the Ethical Use of AI and Data, formalised in January 2019's 'Model AI Governance Framework' (Personal Data Protection Commission Singapore, 2019). The framework provides a set of guiding ethical principles, which are translated into practical measures that businesses can adopt, including how to manage risk, how to incorporate human decision making into AI and how to minimise bias in datasets.

**China**'s economy has experienced huge growth in recent decades, making it the world's second largest economy (World Economic Forum, 2018). To catapult China to world leader in AI, the Chinese Government released the 'Next Generation AI Development Plan' in July 2017. The detailed plan outlines objectives for industrialisation, R&D, education, ethical standards and security (Foundation for Law and International Affairs, 2017). In line with Japan, it is a three-step strategy for AI development, culminating in 2030 with becoming the world's leading centre for AI innovation.

There is substantial focus on governance, with intent to develop regulations and ethical norms for AI and 'actively participate' in the global governance of this technology. Formalised under the 'Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry', the strategy iterates four main goals, to: scale-up the development of key AI products (with a focus on intelligent vehicles, service robots, medical diagnosis and video image identification

systems); significantly enhance core competencies in AI; deepen the development of smart manufacturing; and establish the foundation for an AI industry support system (New America, 2018).

In **India**, AI has the potential to add 1 trillion INR to the economy by 2035 (NITI Aayog, 2018). India's AI strategy, named AI for All, aims to utilise the benefits of AI for economic growth but also social development and 'inclusive growth', with significant focus on empowering citizens to find better quality work. The report provides 30 recommendations for the government, which include setting up Centres of Research Excellence for AI (COREs, each with their own Ethics Council), promoting employee reskilling, opening up government datasets and establishing 'Centres for Studies on Technological Sustainability'. It also establishes the concept of India as an 'AI Garage', whereby solutions developed in India can be rolled out to developing economies in the rest of the world.

Alongside them, **Taiwan** released an 'AI Action Plan' in January 2018 (AI Taiwan, 2018), focused heavily on industrial innovation, and **South Korea** announced their 'AI Information Industry Development Strategy' in May 2018 (H. Sarmah, 2019). The report on which this was based (Government of the Republic of Korea, 2016) provides fairly extensive recommendations for government, across data management, research methods, AI in government and public services, education and legal and ethical reforms.

**Malaysia**'s Prime Minister announced plans to introduce a national AI framework back in 2017 (Abas, 2017), an extension of the existing 'Big Data Analytics Framework' and to be led by the Malaysia Digital Economy Corporation (MDEC). There has been no update from the government since 2017. More recently, **Sri Lanka**'s wealthiest businessman Dharmika Perera has called for a national AI strategy in the country, at an event held in collaboration with the Computer Society of Sri Lanka (Cassim, 2019), however there has not yet been an official pledge from the government.

In the Middle East, the **United Arab Emirates** was the first country to develop a strategy for AI, released in October 2017 and with emphasis on boosting government performance and financial resilience (UAE Government, 2018). Investment will be focused on education, transport, energy, technology and space. The ethics underlying the framework is fairly comprehensive; the Dubai AI Ethics Guidelines dictate the key principles that make AI systems fair, accountable, transparent and explainable (Smart Dubai, 2019a). There is even a self-assessment tool available to help developers of AI technology to evaluate the ethics of their system (Smart Dubai, 2019b).

World leader in technology **Israel** is yet to announce a national AI strategy. Acknowledging the global race for AI leadership, a recent report by the Israel Innovation Authority (Israel Innovation Authority, 2019) recommended that Israel develop a national AI strategy 'shared by government, academia and industry'.

## 5.4. Africa

Africa has taken great interest in AI; a recent white paper suggests this technology could solve some of the most pressing problems in Sub-Saharan Africa, from agricultural yields to providing secure financial services (Access Partnership, 2018). The document provides essential elements for a pan-African strategy on AI, suggesting that lack of government engagement to date has been a hindrance and encouraging African governments to take a proactive approach to AI policy. It lists laws on data privacy and security, initiatives to foster widespread adoption of the cloud, regulations to enable the use of AI for provision of public services, and adoption of international data standards as key elements of such a policy, although one is yet to emerge.

**Kenya** however has announced a task force on AI (and blockchain) chaired by a former Secretary in the Ministry of Information and Communication, which will offer recommendations to the government on how best to leverage these technologies (Kenyan Wallstreet, 2018). **Tunisia** too has created a task force to put together a national strategy on AI and held a workshop in 2018 entitled 'National AI Strategy: Unlocking Tunisia's capabilities potential' (ANPR, 2018).

## 5.5. South America

**Mexico** is so far the only South American nation to release an AI strategy. It includes five key actions, to: develop an adequate governance framework to promote multi-sectorial dialogue; map the needs of industry; promote Mexico's international leadership in AI; publish recommendations for public consultation; and work both with experts and the public to achieve the continuity of these efforts (México Digital, 2018). The strategy is the formalisation of a White Paper (Martinho-Truswell et al., 2018) authored by the British Embassy in Mexico, consultancy firm Oxford Insights and thinktank C Minds, with the collaboration of the Mexican Government.

The strategy emphasises the role of its citizens in Mexico's AI development and the potential of social applications of AI, such as improving healthcare and education. It also addresses the fact that 18% of all jobs in Mexico (9.8 million in total) will be affected by automation in the coming 20 years and makes a number of recommendations to improve education in computational approaches.

Other South American nations will likely follow suit if they are to keep pace with emerging markets in Asia. Recent reports suggest AI could double the size of the economy in Argentina, Brazil, Chile, Colombia and Peru (Ovanessoff and Plastino, 2017).

## 5.6. Australasia

**Australia** does not yet have a national strategy on AI. It does however have a 'Digital Economy Strategy' (Australian Government, 2017) which discusses empowering Australians through 'digital skills and inclusion', listing AI as a key emerging technology. A report on 'Australia's Tech Future' further details plans for AI, including using AI to improve public services, increase administrative efficiency and improve policy development (Australian Government, 2018).

The report also details plans to develop an ethics framework with industry and academia, alongside legislative reforms to streamline the sharing and release of public sector data. The draft ethics framework (Dawson et al., 2019) is based on case studies from around the world of AI 'gone wrong' and offers eight core principles to prevent this, including fairness, accountability and the protection of privacy. It is one of the more comprehensive ethics frameworks published so far, although yet to be implemented.

Work is also ongoing to launch a national strategy in **New Zealand**, where AI has the potential to increase GDP by up to \$54 billion (AI Forum New Zealand, 2018). The AI Forum of New Zealand has been set up to increase awareness and capabilities of AI in the country, bringing together public, industry, academia and Government.

Their report 'Artificial Intelligence: Shaping The Future of New Zealand' (AI Forum New Zealand, 2018) lays out a number of recommendations for the government to coordinate strategy development (i.e. to coordinate research investment and the use of AI in government services); increase awareness of AI (including conducting research into the impacts of AI on economy and society); assist AI adoption (by developing best practice resources for industry); increase the accessibility of trusted data; grow the AI talent pool (developing AI courses, including AI on the list of valued skills for immigrants); and finally to adapt to AI's effects on law, ethics and society. This

includes the recommendation to establish an AI ethics and society working group to investigate moral issues and develop guidelines for best practice in AI, aligned with international bodies.

### Challenges to government adoption of AI

The World Economic Forum has, through consultation with stakeholders, identified five major roadblocks to government adoption of AI:

1. Effective use of data - Lack of understanding of data infrastructure, not implementing data governance processes (e.g. employing data officers and tools to efficiently access data).
2. Data and AI skills - It is difficult for governments, which have smaller hiring budgets than many big companies, to attract candidates with the required skills to develop first-rate AI solutions.
3. The AI ecosystem - There are many different companies operating in the AI market and it is rapidly changing. Many of the start-ups pioneering AI solutions have limited experience working with government and scaling up for large projects.
4. Legacy culture - It can be difficult to adopt transformative technology in government, where there are established practices and processes and perhaps less encouragement for employees to take risks and innovate than in the private sector.
5. Procurement mechanisms - The private sector treats algorithms as intellectual property, which may make it difficult for governments to customise them as required. Public procurement mechanisms can also be slow and complicated (e.g. extensive terms and conditions, long wait times from tender response submission to final decision).

(Torres Santeli and Gerdon, 2019)

## 5.7. International AI Initiatives, in addition to the EU

In addition to the EU, there are a growing number of international strategies on AI, aiming to provide a unifying framework for governments worldwide on stewardship of this new and powerful technology.

### G7 Common Vision for the Future of AI

At the 2018 meeting of the G7 in Charlevoix, Canada, the leaders of the G7 (Canada, France, Germany, Italy, Japan, the United Kingdom and the United States) committed to 12 principles for AI, summarised below:

1. Promote human-centric AI and the commercial adoption of AI, and continue to advance appropriate technical, ethical and technologically neutral approaches.
2. Promote investment in R&D in AI that generates public test in new technologies and supports economic growth.
3. Support education, training and re-skilling for the workforce.
4. Support and involve underrepresented groups, including women and marginalised individuals, in the development and implementation of AI.

5. Facilitate multi-stakeholder dialogue on how to advance AI innovation to increase trust and adoption.
6. Support efforts to promote trust in AI, with particular attention to countering harmful stereotypes and fostering gender equality. Foster initiatives that promote safety and transparency.
7. Promote the use of AI by small and medium-sized enterprises.
8. Promote active labour market policies, workforce development and training programmes to develop the skills needed for new jobs.
9. Encourage investment in AI.
10. Encourage initiatives to improve digital security and develop codes of conduct.
11. Ensure the development of frameworks for privacy and data protection.
12. Support an open market environment for the free flow of data, while respecting privacy and data protection.

(G7 Canadian Presidency, 2018).

### **Nordic-Baltic Region Declaration on AI**

The declaration signed by the Nordic-Baltic Region (comprising Denmark, Estonia, Finland, the Faroe Islands, Iceland, Latvia, Lithuania, Norway, Sweden and the Åland Islands) aims to promote the use of AI in the region, including improving the opportunities for skills development, increasing access to data and a specific policy objective to develop 'ethical and transparent guidelines, standards, principles and values' for when and how AI should be used (Nordic Co-operation, 2018).

### **OECD Principles on AI**

On 22 May 2019, the Organisation for Economic Co-operation and Development issued its principles for AI, the first international standards agreed by governments for the responsible development of AI. They include practical policy recommendations as well as value-based principles for the 'responsible stewardship of trustworthy AI', summarised below:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should respect the rule of law, human rights, democratic values and diversity, and there should include appropriate safeguards to ensure a fair society.
- There should be transparency around AI to ensure that people understand outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles and risks should be continually assessed.
- Organisations and individuals developing, deploying or operating AI systems should be held accountable.

These principles have been agreed by the governments of the 36 OECD Member States as well as Argentina, Brazil, Colombia, Costa Rica, Peru and Romania (OECD, 2019a). The G20 human-centred AI Principles were released in June 2019 and are drawn from the OECD Principles (G20, 2019).

### **United Nations**

The UN has several initiatives relating to AI, including:

- AI for Good Global Summit- Summits held since 2017 have focused on strategies to ensure the safe and inclusive development of AI (International Telecommunication Union, 2018a,b). The events are organised by the International Telecommunication Union, which aims to 'provide a neutral platform for government, industry and

academia to build a common understanding of the capabilities of emerging AI technologies and consequent needs for technical standardisation and policy guidance.'

- UNICRI Centre for AI and Robotics - The UN Interregional Crime and Justice Research Institute (UNICRI) launched a programme on AI and Robotics in 2015 and will be opening a centre dedicated to these topics in The Hague (UNICRI, 2019).
- UNESCO Report on Robotics Ethics - The UNESCO World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) has authored a report on 'Robotics Ethics', which deals with the ethical challenges of robots in society and provides ethical principles and values, and a technology-based ethical framework (COMEST, 2017).

### **World Economic Forum**

The World Economic Forum (WEF) formed a Global AI Council in May 2019, co-chaired by speech recognition developer Kai-Fu Lee, previously of Apple, Microsoft and Google, and current President of Microsoft Bradford Smith. One of six 'Fourth Industrial Revolution' councils, the Global AI Council will develop policy guidance and address governance gaps, in order to develop a common understanding among countries of best practice in AI policy (World Economic Forum, 2019a).

In October 2019, they released a framework for developing a national AI strategy to guide governments that are yet to develop or are currently developing a national strategy for AI. The WEF describe it as a way to create a 'minimum viable' AI strategy and includes four main stages:

- 1) Assess long-term strategic priorities
- 2) Set national goals and targets
- 3) Create plans for essential strategic elements
- 4) Develop the implementation plan

The WEF has also announced plans to develop an 'AI toolkit' to help businesses to best implement AI and to create their own ethics councils, which will be released at 2020's Davos conference (Vanian, 2019).

## **5.8. Government Readiness for AI**

A report commissioned by Canada's International Development Research Centre (Oxford Insights, 2019) evaluated the 'AI readiness' of governments around the globe in 2019, using a range of data including not only the presence of a national AI strategy, but also data protection laws, statistics on AI startups and technology skills.

Singapore was ranked number 1 in their estimation, with Japan as the only other Asian nation in the top 10 (Table 3). Sixty percent of countries in the top 10 were European, with the remainder from North America.

The strong European representation in this analysis is reflective of the value of the unifying EU framework, as well as Europe's economic power. The analysis also praises the policy strategies of individual European nations, which, importantly, have been developed in a culture of collaboration. Examples of this collaborative approach include the EU Declaration of Cooperation on AI (European Commission, 2018d), in which Member States agreed to cooperate on boosting Europe's capacity in AI, and individual partnerships between Member States, such as that of Finland, Estonia and Sweden, working together to trial new applications of AI.

*Table 3: Top 10 rankings for Government AI Readiness 2018/19. Source: Oxford Insights, 2019.*

Rank	Country	Score
1	Singapore	9.19
2	United Kingdom	9.07
3	Germany	8.81
4	USA	8.80
5	Finland	8.77
6	Sweden	8.67
6	Canada	8.67
8	France	8.61
9	Denmark	8.60
10	Japan	8.58

Singapore ranked highest of all nations while Japan, the second country in the world to release a national strategy on AI, ranked 10<sup>th</sup>. China's position as 21<sup>st</sup> in the global rankings is expected to improve next year as its investments in AI begin to pay off. Progress in Asia overall has been unbalanced, with two countries in the region also ranking in the bottom ten worldwide, reflecting the income inequality in the region.

Despite the comparatively slow development of their national strategy, the USA ranked 4<sup>th</sup>, with Canada not far behind. Both nations are supported by their strong economies, highly skilled workforces, private sector innovation and abundance of data, to a level at which regions missing from the top 10 – Africa, South America and Australasia – are unable to compete.

This framework provides a highly useful metric by which to assess the ability of governments to capitalise on AI's potential in the coming years. What this analysis does not consider however is how robustly each nation is considering the moral and ethical issues surrounding the use of AI, which we will explore below.

## 6. Emerging Themes

Our review of the literature on the ethical issues surrounding AI and intelligent robots highlights a wide range of potential impacts, including in the social, psychological, financial, legal and environmental domains. These are bound up with issues of trust and are tackled in different ways by the emerging ethical initiatives. Standards and regulation are also beginning to develop that go some way to addressing these concerns. However, the focus of many existing strategies on AI is on enabling technology development and, while ethical issues are addressed, notable gaps can be identified.

### 6.1. Addressing ethical issues through national and international strategies

There are several themes shared by the various national strategies on AI, among which **industrialisation** and **productivity** perhaps rank highest. All countries have some sort of industrial strategy for AI, and this is particularly prominent in the emerging economies of Southeast Asia. Most of the strategies make reference to the importance of AI for business competitiveness and several, including those of Germany, South Korea, Taiwan and the UK, announce extra funding and specialised incubators for AI-focused start-ups.

Whether in the private or public sector, the importance of **research** and development is also universally recognised, with almost all strategies pledging enhanced funding for research and many to establish 'centres of excellence' entirely dedicated to AI research, including strategies from Canada, Germany and India.

Essential to developing a strong research effort is talent, and so investing in **people** and education also features heavily in most strategies. The UK has announced 'Turing Fellowships' to fund new academics exploring computational approaches, while Germany has provided for at least an extra 100 professors working on AI – both under the umbrella of the EU commitment to train, attract and retain talent. In Asia, South Korea has committed to developing six new graduate programmes to train a total of 5,000 AI specialists, while Taiwan has committed to training double that number by 2021.

Most of the strategies also consider the impact the AI revolution will have on the non-technology literate workforce, who may be the first to lose their jobs to automation. Although this crosses over into ethical considerations, several of the strategies make practical commitments to **re-training** programmes to help those affected to find new work. This is a key objective in the EU plan (objective 2.4: 'adapting our learning and training programmes and systems to better prepare our society for AI'), and therefore the plans of its Member States. The UK for example will initiate an > €70 million re-training scheme to help people gain digital skills and Germany has revealed a similar 'National Further Training Strategy'. Naturally, those countries most in need of re-training have the least funding available for it. Mexico's strategy however emphasises the importance of computational thinking and mathematics in lifelong teaching, including to help its citizens retrain, while India pledges to promote informal training institutions and create financial incentives for reskilling of employees. Other strategies however suggest re-training is the responsibility of individual businesses and do not allocate separate funding for it.

**Collaboration** between sectors and countries is another common thread, yet interpreted differently by different countries. India's approach for example is one of sharing; the 'AI Garage' concept named in their strategy means AI-based solutions developed in India will be rolled out to developing economies facing similar issues. Conversely, the US Executive Order on AI sets out to

'promote an international environment that supports American AI' while also protecting the nation's technological advantage against 'foreign adversaries'. Naturally, the strategies of EU Member States display an inclination for cross-border collaboration. Sweden for example states a need to develop partnerships and collaborations with other countries 'especially within the EU', while Denmark's strategy also emphasises close cooperation with other European countries.

The democratisation of technology has the potential to reduce inequalities in society, and **inclusion** and **social development** are important goals for many national AI initiatives, particularly those of developing economies. India's strategy discusses AI for 'greater good', focusing on the possibilities for better access to healthcare, economic growth for groups previously excluded from formal financial products, and using data to aid small-scale farmers. Mexico's strategy lists inclusion as one of its five major goals, which includes aims to democratise productivity and promote gender equality. France too aims for an AI that 'supports inclusivity', striving for policies that reduce both social and economic inequalities.

Determining who is **responsible** for the actions and behaviour of AI is highly important, and challenging in both moral and legal senses. Currently, AI is most likely considered to be the legal responsibility of a relevant human actor – a tool in the hands of a developer, user, vendor, and so on. However, this framework does not account for the unique challenges brought by AI, and many grey areas exist. As just one example, as a machine learns and evolves to become different to its initial programming over many iterations, it may become more difficult to assign responsibility for its behaviour to the programmer. Similarly, if a user or vendor is not adequately briefed on the limitations of an AI agent, then it may not be possible to hold them responsible. Without proving that an AI agent intended to commit a crime (*mens rea*) and can act voluntarily, both of which are controversial concepts, then it may not be possible to deem an AI agent responsible and liable for its own actions.

## 6.2. Addressing the governance challenges posed by AI

There are currently two major international frameworks for the governance of AI: that of the EU (see Section 5.1) and the Organisation for Economic Co-operation and Development (OECD).

The OECD launched a set of principles for AI in May 2019 (OECD, 2019a) which were at that time adopted by 42 countries. The OECD framework offers five fundamental principles for the operation of AI (see section 5.1.1) as well as accompanying practical recommendations for governments to achieve them. The G20 soon after adopted its own, human-centred AI principles, drawn from (and essentially an abridged version of) those of the OECD (G20, 2019).

The OECD Principles have also been backed by the European Commission, which has its own strategy on AI since April 2018 (European Commission, 2018b). The EU framework includes comprehensive plans for investment, but also makes preparations for complex socio-economic changes and is complemented by a separate set of ethics guidelines (European Commission High-Level Expert Group on AI, 2019a).

### Gaps in AI frameworks

These frameworks address the moral and ethical dilemmas identified in this report to varying extents, with some notable gaps. Regarding **environmental concerns** (Section 2.5), while the OECD makes reference to developing AI that brings positive outcomes for the planet, including protecting natural environments, the document does not suggest ways to achieve this, nor does it mention any specific environmental challenges to be considered.

The EU Communication on AI does not discuss the environment. However, its accompanying ethics guidelines are founded on the principle of prevention of harm, which includes harm to the natural

environment and all living beings. Societal and environmental well-being (including sustainability and 'environmental friendliness') is one of the EU's requirements for trustworthy AI and its assessment list includes explicit consideration of risks to the environment or to animals. Particular examples are also given on how to achieve this (e.g. critical assessment of resource use and energy consumption throughout the supply chain).

Impacts on human **psychology**, including how people interact with AI and subsequent effects on how people interact with each other, could be further addressed in the frameworks. The psychosocial impact of AI is not considered by the OECD Principles or the EU Communication. However, the EU requirement for societal well-being to be considered does address 'social impact', which includes possible changes to social relationships and loss of social skills. The guidelines state that such effects must 'be carefully monitored and considered' and that AI interacting with humans must clearly signal that its social interaction is simulated. However, more specific consideration could be given to human-robot relationships or more complex effects on the human psyche, such as those outlined above (Section 2.2).

While both frameworks capably address changes to the **labour market** (Section 2.1.1), attention to more nuanced factors, including the potential for AI to drive **inequalities** (2.1.2) and **bias** (2.1.4), is more limited. The OECD's first principle of inclusive growth, sustainable development and well-being states that AI should be developed in a way that reduces 'economic, social, gender and other inequalities'. This is also covered to a degree by the second OECD principle, which states that AI systems should respect diversity and include safeguards to ensure a fair society, however detail on how this can be achieved is lacking.

The EU ethics guidelines are more comprehensive on this point and include diversity, non-discrimination and fairness as a separate requirement. The guidelines elaborate that equality is a fundamental basis for trustworthy AI and state that AI should be trained on data which is representative of different groups in order to prevent biased outputs. The guidelines include additional recommendations on the avoidance of unfair bias.

Both frameworks include **human rights** and **democratic values** (Sections 2.1.3, 2.1.5) as key tenets. This includes **privacy**, which is one of the OECD's human-centred values and a key requirement of the EU ethics guidelines, which elaborates on the importance of data governance and data access rules. Issues concerning privacy are also covered by existing OECD data protection guidelines (OECD, 2013).

The implications of AI for **democracy** (Section 2.1.5) are only briefly mentioned by the OECD, with no discussion of the particular issues facing governments at the present time, such as Deepfake or the manipulation of opinion through targeted news stories. Threats to democracy are not mentioned at all in the EU Communication, although society and democracy is a key theme in the associated ethics guidelines, which state that AI systems should serve to maintain democracy and not undermine 'democratic processes, human deliberation or democratic voting systems.'

These issues form part of a bigger question surrounding changes to the **legal system** (Section 2.4) that may be necessary in the AI age, including important questions around liability for misconduct involving AI. The issue of liability is explicitly addressed by the EU in both its Communication and ethics guidelines. Ensuring an appropriate legal framework is a key requirement of the EU Communication on AI, which includes guidance on product liability and an exploration of safety and security issues (including criminal use). The accompanying ethics guidelines also suitably handle this issue, including providing guidance for developers on how to ensure legal compliance. Relevant changes to regulation are further addressed in the recent AI Policy and Investment Recommendations (European Commission High-Level Expert Group on AI, 2019b), which explore potential changes to current EU laws and the need for new regulatory powers.

The OECD principles are more limited on this point. While they provide guidance for governments to create an 'enabling policy environment' for AI, including a recommendation to review and adapt

regulatory frameworks, this is stated to be for the purpose of encouraging 'innovation and competition' and does not address the issue of liability for AI-assisted crime.

These questions could also come under the issue of **accountability** (2.6.4) however, which is adequately addressed by both frameworks. The OECD lists accountability as a key principle and states that 'organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning' (OECD, 2019a). It is likewise a core principle of the EU ethics guidelines, which provides more than 10 conditions for accountability in its assessment list for trustworthy AI.

Many of the aforementioned issues are ultimately important for building **trust** in AI (Section 2.6), which also requires AI to be fair (2.6.2) and transparent (2.6.3). These issues are at the foundation of the EU ethics guidelines where they are dealt with in great detail. The OECD also states that AI systems should ensure a 'fair and just society'. Transparency and explainability is a core principle for the OECD, with strong emphasis on the fact that people should be able to understand and challenge AI systems. The OECD Principles offer less context on these issues and do not consider practical means of ensuring this (e.g. audits of algorithms), which are considered by the EU ethics guidelines. The ethics guidelines also consider the need for human oversight (including discussion of the human-in-the-loop approach and the need for a 'stop button', neither of which are mentioned by the OECD principles).

Finally, although both acknowledge the beneficial use of AI in **finance** (Section 2.3), neither framework adequately addresses potential negative impacts on the financial system, either through accidental harm or malicious activity. The potential for AI-assisted financial crime is an important one and currently unaddressed by any international framework. However, the G7 has recently voiced concerns about digital currencies and various other new financial products being developed (Reuters, 2019), which suggests that regulatory changes in this regard are afoot.

## 7. Summary

What this report makes clear is the diversity and complexity of the ethical concerns arising from the development of artificial intelligence; from large scale issues such job losses from automation, degradation of the environment and furthering inequalities, to more personal moral quandaries such as how AI may affect our privacy, our ability to judge what is real, and our personal relationships.

What is also clear is that there are various **approaches to ethics**. Robust ethical principles are essential in the future of this rapidly developing technology, but not all countries understand ethics in the same way. There are a number of independent ethical initiatives for AI, such as Germany's Institute for Ethics in AI, funded by Facebook, and the private donor-funded Future of Life Institute in the US. An increasing number of governments are also developing national AI strategies, with their own ethics components. A number of countries have committed to creating AI ethics councils, including Germany, the UK, India, Singapore and Mexico. The UAE has also prioritised ethics in its national strategy, by developing an 'Ethical AI Toolkit' and self-assessment tool for developers, while several others give only passing reference; ethics is almost completely left out by Japan, South Korea and Taiwan.

Our assessment shows that the vast majority of ethical issues identified here are also addressed in some form by at least one of the current international frameworks; the EU Communication (supplemented by separate ethics guidelines) and the OECD Principles on AI.

The current frameworks address the major ethical concerns and make recommendations for governments to manage them, but **notable gaps** exist. These include environmental impacts, including increased energy consumption associated with AI data processing and manufacture, and inequality arising from unequal distribution of benefits and potential exploitation of workers. Policy options relating to environmental impacts include providing a stronger mandate for sustainability and ecological responsibility; requiring energy use to be monitored, and publication of carbon footprints; and potentially policies that direct technology innovation towards urgent environmental priorities. In the case of inequality, options include declaring AI as a public, rather than private, good. This would require changes to cultural norms and new strategies to help navigate a transition to an AI-driven economy. Setting minimum standards for corporate social responsibility reporting would encourage larger, transnational corporations to clearly show how they are sharing the benefits of AI. Economic policies may be required to support workers displaced by AI; such policies should focus on those at most risk of being left behind and might include policies designed to create support structures for precarious workers. It will be important for future iterations of these frameworks to address these and other gaps in order to adequately prepare for the full implications of an AI future. In addition, to clarify the issue of responsibility pertaining to AI behaviour, moral and legislative frameworks will require updating alongside the development of the technology itself.

Governments also need to develop new, up-to-date forms of **technology assessment** – allowing them to understand such technologies deeply while they can still be shaped, such as the Accountability Office's Technology Assessment Unit in the USA or the European Foresight platform (<http://www.foresight-platform.eu/>). New forms of technology assessment TA should include processes of Ethical Risk Assessment, such as the one set out in BS8611, and other forms of ethical evaluation currently being drafted in the IEEE Standards Association P7000 series of ethical standards; P7001 for instance sets out a method for measuring the transparency of an AI.

There is a clear need for the development of viable and applicable **legislation and policies** that will face the multifaceted challenges associated with AI, including potential breaches of fundamental ethical principles. Policy makers are in the valuable position of being able to develop policy that actively shapes the development of AI and as data-driven and machine-learning approaches begin

to take increasing roles in society, thoughtful and detailed strategies on how to share benefits and achieve the best possible outcomes, while effectively managing risk, will be essential.

As well as the very encouraging progress made in policy so far, this report also reveals a concerning **disparity** between regions. Successful AI development requires substantial investment, and as automation and intelligent machines begin to drive government processes, there is a real risk that lower income countries – those nations of the Global South – will be left behind. It is incumbent upon policymakers therefore to try to ensure that AI does not widen global inequalities. This could include **data sharing** and collaborative approaches, such as India's promise to share its AI solutions with other developing countries, and efforts to make teaching on computational approaches a fundamental part of education, available to all.

To return to our main theme, **ethical considerations** must also be a critical component of any policy on AI. It speaks volumes that the nation ranked highest in the 2019 Government AI Readiness Index has prioritised ethics so strongly in their national AI Strategy. Singapore is one of a few governments to create an AI Ethics Council and has incorporated a range of ethical considerations into its policy. Addressing ethical concerns is also the first key point in the World Economic Forum's framework for developing a national AI strategy. So, aside from any potential moral obligations, it seems unlikely that governments that do not take ethics seriously will be able to succeed in the competitive global forum.

## 8. Appendix

### Building ethical robots

In the future it's very likely that intelligent machines will have to make decisions that affect human safety, psychology and society. For example, a search and rescue robot should be able to 'choose' the victims to assist first after an earthquake; an autonomous car should be able to 'choose' what or who to crash into when an accident cannot be avoided; a home-care robot should be able to balance its user's privacy and their nursing needs. But how do we integrate societal, legal and moral values into technological developments in AI? How can we program machines to make ethical decisions - to what extent can ethical considerations even be written in a language that computers understand?

Devising a method for integrating ethics into the design of AI has become a main focus of research over the last few years. Approaches towards moral decision making generally fall into two camps, 'top-down' and 'bottom-up' approaches (Allen et al., 2005). Top-down approaches involve explicitly programming moral rules and decisions into artificial agents, such as 'thou shalt not kill'. Bottom up approaches, on the other hand, involve developing systems that can implicitly learn to distinguish between moral and immoral behaviours.

#### *Bottom-up approaches*

Bottom up approaches involve allowing robots to learn ethics independently of humans, for instance by using machine learning. Santos-Lang (2002) points out that this is a better approach, as humans themselves continuously learn to be ethical. An advantage of this is that most of the work is done by the machine itself, which avoids the robot being influenced by the designers' biases. However the downside is that machines could demonstrate unintended behaviour that deviates from the desired goal. For example, if a robot was programmed to 'choose behaviour that leads to the most happiness', the machine may discover that it can more quickly reach its goal of maximising happiness by first increasing its own learning efficiency, 'temporarily' shifting away from the original goal. Because of the shift, the machine may even choose behaviours that temporarily reduce happiness, if these behaviours were to ultimately help it achieve its goal. For example a machine could try to rob, lie and kill, in order to become an ethical paragon later.

#### *Top-down approaches*

Top-down approaches involve programming agents with strict rules that they should follow in given circumstances. For example, in self-driving cars a vehicle could be programmed with the command 'you shall not drive faster than 130 km/h on the highway'. The problem with top down approaches is that they require deciding which moral theories ought to be applied. Examples of competing moral theories include utilitarian ethics, deontological ethics and the commensal view and the Doctrine of Double Effect.

Utilitarianism is based on the notion that the morality of an action should be judged by its consequences. In other words, an action is judged to be morally right if its consequences lead to the greater good. Different utilitarian theories vary in terms of the definition of the 'good' they aim to maximise. For example, Bentham (1789) proposed that a moral agent should aim to maximise the total happiness of a population of people.

Deontological (duty-based) ethics, on the other hand argues that actions should be judged not on the basis of their expected outcomes, but on what people do. Duty-based ethics teaches that actions are right or wrong regardless of the good or bad consequences that may be produced. Under this form of ethics you can't justify an action by showing that it produced good consequences.

Sometimes different moral theories can directly contradict each other. For example, in the case of a self-driving car that has to decide whether to swerve to avoid animals in its path. Under the commensal view, animal lives are treated as if they are worth some small fraction of what human lives are worth, and so the car would swerve if there was a low chance of causing harm to a human (Bogosian, 2017). However, the incommensal view would never allow humans to be placed at additional risk of fatality in order to save an animal. Since this view fundamentally rejects the assumptions of the other, and holds that no tradeoff is permissible, there is no obvious 'halfway point' where the competing principles can meet.

Bonnemains et al. (2018) describe a dilemma where a drone programmed to take out a missile threatening an allied ammo factory is suddenly alerted to a second threat - a missile heading towards some civilians. The drone must decide whether to continue its original mission, or take out the new missile in order to save the civilians. The decision outcome is different depending on whether you use utilitarianism, deontological ethics and the Doctrine of Double Effect - a theory which states that if doing something morally good has a morally bad side-effect, it's ethically okay to do it providing that the bad side-effect wasn't intended.

Some of the theories are unable to solve the problem. For instance, from a deontological perspective both decisions are valid, as they both arise from good intentions. In the case of utilitarian ethics, without any information about the number of civilians that are in danger, or the value of the strategic factory, it would be difficult for a drone to reach a decision. In order to follow the utilitarian doctrine and make a decision that maximised a 'good outcome', an artificial agent would need to identify all possible consequences of a decision, from all parties' perspectives, before making a judgement about which consequence is preferable. This would be impossible in the field. Another issue is how should a drone decide which outcomes it prefers when this is a subjective judgement? What is Good? Giving an answer to this broad philosophical issue is hardly possible for an autonomous agent, or the person programming it.

Under the Doctrine of Double Effect the drone would not be allowed to intercept the missile and save the civilians, as the bad side effect (the destruction of the drone itself) would be a means to ensuring the good effect (saving the humans). It would therefore continue to pursue its original goal and destroy the launcher, letting the civilians die.

If philosophers cannot agree on the merits of various theories, companies, governments, and researchers will find it even more difficult to decide which system to use for artificial agents (Bogosian, 2017). People's personal moral judgements can also differ widely when faced with moral dilemmas (Greene et al., 2001), particularly when they are considering politicised issues such as racial fairness and economic inequality. Bogosian (2017) argues that instead, we should design machines to be fundamentally uncertain about morality.

## REFERENCES

- Abas, A. (2017). *Najib unveils Malaysia's digital 'to-do list' to propel digital initiatives implementation.* [online] Nst.com.my. Available from: <https://www.nst.com.my/news/nation/2017/10/292784/najib-unveils-malaysias-digital-do-list-propel-digital-initiatives> [Accessed 8 May 2019].
- Access Partnership and the University of Pretoria (2018). *Artificial Intelligence for Africa: An Opportunity for Growth, Development and Democratisation.* Available from: [https://www.up.ac.za/media/shared/7/ZP\\_Files/ai-for-africa.zp165664.pdf](https://www.up.ac.za/media/shared/7/ZP_Files/ai-for-africa.zp165664.pdf)
- Acemoglu, D. and Restrepo, P. (2018) Low-skill and high-skill automation. *Journal of Human Capital*, 2018, vol. 12, no. 2.
- Agency for Digital Italy (2019). *Artificial Intelligence task force.* [online] IA-Gov. Available from: <https://ia.italia.it/en/> [Accessed 10 May 2019].
- AI4All (2019). *What we do* [online] Available from: <http://ai-4-all.org> [Accessed 11/03/2019].
- AI For Humanity (2018). *AI for humanity: French Strategy for Artificial Intelligence* [online] Available from: <https://www.aiforhumanity.fr/en/> [Accessed 10 May 2019].
- AI Forum New Zealand (2018). *Artificial Intelligence: Shaping a Future New Zealand.* Available from: [https://aiforum.org.nz/wp-content/uploads/2018/07/AI-Report-2018\\_web-version.pdf](https://aiforum.org.nz/wp-content/uploads/2018/07/AI-Report-2018_web-version.pdf)
- AI Now Institute, (2018). *AI Now Report.* AI Now Institute, New York University. Available from: [https://ainowinstitute.org/AI\\_Now\\_2018\\_Report.pdf](https://ainowinstitute.org/AI_Now_2018_Report.pdf)
- AI Singapore. (2018). *AI Singapore.* [online] Available from: <https://www.aisingapore.org> [Accessed 26 Apr. 2019].
- AI Taiwan. (2019). *AI Taiwan.* [online] Available from: <https://ai.taiwan.gov.tw> [Accessed 28 Apr. 2019].
- Allen, C., Smit, I., and Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology.* doi:10.1007/s10676-006-0004-4.
- Allen, G., and Chan, T., (2017). *Artificial Intelligence and National Security.* Available from: <https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf>
- Amoroso, D., and Tamburrini, G. (2018). The Ethical and Legal Case Against Autonomy in Weapons Systems. *Global Jurist* 18 (1), DOI: 10.1515/gj-2017-0012.
- Anderson, J. M., Heaton, P. and = Carroll, S. J. (2010). *The U.S. Experience with No-Fault Automobile Insurance: A Retrospective.* Santa Monica, CA: RAND Corporation. Available from: <https://www.rand.org/pubs/monographs/MG860.html>.
- ANPR (2018). *National AI Strategy: Unlocking Tunisia's capabilities potential* [online] Available from: <http://www.anpr.tn/national-ai-strategy-unlocking-tunisias-capabilities-potential/>. [Accessed 6 May 2019].
- Apps, P. (2019). *Commentary: Are China, Russia winning the AI arms race?* [online] U.S. Available from: <https://www.reuters.com/article/us-apps-ai-commentary/commentary-are-china-russia-winning-the-ai-arms-race-idUSKCN1P91NM>.
- Arnold, T., and Scheutz, M. (2018). The 'big red button' is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology.* 20 (1), 59–69.

- Asaro, P. (2012). On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making. *International Review of the Red Cross.* 94 (886), 687-703.
- Atabekov, A. and Yastrebov, O. (2018) Legal status of Artificial Intelligence: Legislation on the move. European Research Studies Journal Volume XXI, Issue 4, 2018 pp. 773 - 782
- Australian Government (2017). *The Digital Economy: Opening Up The Conversation.* Department of Industry, Innovation and Science. Available from:  
<https://www.archive.industry.gov.au/innovation/Digital-Economy/Documents/Digital-Economy-Strategy-Consultation-Paper.pdf>
- Australian Government (2018). *Australia's Tech Future.* Department of Industry, Innovation and Science. Available from: <https://www.industry.gov.au/sites/default/files/2018-12/australias-tech-future.pdf>
- Austrian Council on Robotics and Artificial Intelligence (2018). Die Zukunft Österreichs mit Robotik und Künstlicher Intelligenz positiv gestalten. *White Paper des Österreichischen Rats für Robotik und Künstliche Intelligenz.* Available from: [https://www.acrai.at/wp-content/uploads/2019/04/ACRAI\\_whitebook\\_online\\_2018-1.pdf](https://www.acrai.at/wp-content/uploads/2019/04/ACRAI_whitebook_online_2018-1.pdf)
- Austrian Council on Robotics and Artificial Intelligence (2019). *Österreichischer Rat für Robotik und Künstliche Intelligenz.* [online] Available from: <https://www.acrai.at/> [Accessed 10 May 2019].
- Autor, D. H. (2015). Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives.* 29(3), 3–30.
- Bandyopadhyay, A., and Hazra, A. (2017). A comparative study of classifier performance on spatial and temporal features of handwritten behavioural data. In A. Basu, S. Das, P. Horain, and S. Bhattacharya (eds.). (2016) *Intelligent Human Computer Interaction: 8th International Conference, IHCI 2016, Pilani, IndiaCham: Springer International Publishing*, 111–121.
- Baron, E. (2017). Robot surgery firm from Sunnyvale facing lawsuits, reports of death and injury. *Mercury News.* Available from: <https://www.mercurynews.com/2017/10/22/robot-surgery-firm-from-sunnyvale-facing-lawsuits-reports-of-death-and-injury/>
- Bartlett, J. (2018) How AI could kill off democracy. *New Statesman.* Available from:  
<https://www.newstatesman.com/science-tech/technology/2018/08/how-ai-could-kill-democracy-0>
- BBC News (2017). Singapore to use driverless buses 'from 2022'. BBC. Available from:  
<https://www.bbc.co.uk/news/business-42090987>
- BBC News. (2018). Addison Lee plans self-driving taxis by 2021. BBC. Available from:  
<https://www.bbc.co.uk/news/business-45935000>
- BBC News. (2019a). Autonomous shuttle to be tested in New York City. BBC. Available from:  
<https://www.bbc.co.uk/news/technology-47668886>
- BBC News. (2019b). Uber 'not criminally liable for self-driving death. BBC. Available from:  
<https://www.bbc.co.uk/news/technology-47468391>
- Beane, M. (2018). Young doctors struggle to learn robotic surgery – so they are practicing in the shadows. *The Conversation.* Available from: <https://theconversation.com/young-doctors-struggle-to-learn-robotic-surgery-so-they-are-practicing-in-the-shadows-89646>
- Berger, S. (2019). Vaginal mesh has caused health problems in many women, even as some surgeons vouch for its safety and efficacy. *The Washington Post.* Available from:

[https://www.washingtonpost.com/national/health-science/vaginal-mesh-has-caused-health-problems-in-many-women-even-as-some-surgeons-vouch-for-its-safety-and-efficacy/2019/01/18/1c4a2332-ff0f-11e8-ad40-cdfd0e0dd65a\\_story.html?noredirect=on&utm\\_term=.9bece54e4228](https://www.washingtonpost.com/national/health-science/vaginal-mesh-has-caused-health-problems-in-many-women-even-as-some-surgeons-vouch-for-its-safety-and-efficacy/2019/01/18/1c4a2332-ff0f-11e8-ad40-cdfd0e0dd65a_story.html?noredirect=on&utm_term=.9bece54e4228)

Bershidsky, L (2017). *Elon Musk warns battle for AI supremacy will spark Third World War*. *The Independent*. [online] Available from: <https://www.independent.co.uk/life-style/gadgets-and-tech/news/elon-musk-ai-artificial-intelligence-world-war-three-russia-china-robots-cyber-warfare-replicants-a7931981.html>

Bentham, J. (1789). *A Fragment of Government and an Introduction to the Principles of Morals and Legislation*, London.

Biavaschi, C., Eichhorst, W., Giulietti, C., Kendzia, M., Muravyev, A., Pieters, J., Rodriguez-Planas, N., Schmidl, R., and Zimmermann, K. (2013). Youth Unemployment and Vocational Training. *World Development Report*. World Bank.

Bilge, L., Strufe, T., Balzarotti, D., Kirda, K., and Antipolis, S. (2009). All your contacts are belong to us: Automated identity theft attacks on social networks, In *WWW '09: Proceedings of the 18th international conference on World Wide Web, WWW '09, April 20-24, 2009, Madrid, Spain*. New York, NY, USA. pp. 551–560.

Bogosian, K. (2017) Implementation of Moral Uncertainty in Intelligent Machines. *Minds & Machines* 27 (591).

Bonnemains, V., Saurel, C. & Tessier, C. (2018) Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*. 20 (41). <https://doi.org/10.1007/s10676-018-9444-x>

Borenstein, J. and Arkin, R.C. (2019) Robots, Ethics, and Intimacy: The Need for Scientific Research.  
Available from: <https://www.cc.gatech.edu/ai/robot-lab/online-publications/RobotsEthicsIntimacy-IACAP.pdf>

Bradshaw, S., and Howard, P. (2017) Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. In Woolley, S. and Howard, P. N. (Eds.) (2017) *Working Paper: Project on Computational Propaganda*, Oxford, UK. Available from: [http://comprop.ox.ac.uk/..](http://comprop.ox.ac.uk/)

Bradshaw, T. (2018) Uber halts self-driving car tests after pedestrian is killed. *Financial Times*. 19 March, 2018. Available at: <https://www.ft.com/content/1e2a73d6-2b9e-11e8-9b4b-bc4b9f08f381>

British Standard BS 8611 (2016) *Guide to the Ethical Design of Robots and Robotic Systems*  
<https://shop.bsigroup.com/ProductDetail?pid=00000000030320089>

Brundage, M. And Bryson, J. (2016) Smart Policies for Artificial Intelligence.

Brynjolfsson, E., and McAfee, A (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York, W. W. Norton & Company..

Bryson, J., (2018) Patience is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20 (1). 15–26

Bryson, J. J. (2019). The Past Decade and Future of AI's Impact on Society. In Baddeley, M., Castells, M., Guiora, A., Chau, N., Eichengreen, B., López, R., Kanbur, R. and Burkett, V. (2019) *Towards a New Enlightenment? A Transcendent Decade*. Madrid, Turner.

- Burgmann, T. (2016). There's a cure for that: Canadian doctor pushes for more wearable technology. *Global News Canada*. Available from: <https://globalnews.ca/news/2787549/theres-a-cure-for-that-canadian-doctor-pushes-for-more-wearable-technology/>
- Cadwalladr, C. (2017a). Revealed: How US billionaire helped to back Brexit. *The Guardian*.
- Cadwalladr, C. (2017b). Robert Mercer: The big data billionaire waging war on mainstream media. *The Guardian*.
- Calder,S. (2018). Driverless buses and taxis to be launched in Britain by 2021. *The Independent*. Available from: <https://www.independent.co.uk/travel/news-and-advice/self-driving-buses-driverless-cars-edinburgh-fife-forth-bridge-london-greenwich-a8647926.html>
- Cannon, J. (2018). Starsky Robotics completes first known fully autonomous run without a driver in cab. *Commercial Carrier Journal*. Available from: <https://www.ccjdigital.com/starsky-robotics-autonomous-run-without-driver/>
- Caplan, R., Donovan, J., Hanson, L. and Matthews, J. (2018). *Algorithmic Accountability: A Primer*. New York, Data & Society.
- Cassim, N. (2019). Dhammika makes strong case for national strategy for AI. [online] *Financial Times*. Available from: <http://www.ft.lk/top-story/Dhammika-makes-strong-case-for-national-strategy-for-AI/26-674868> [Accessed 10 May 2019].
- Castellanos, S. (2018). Estonia's CIO Tackles AI Strategy For Government. [online] *WSJ*. Available from: <https://blogs.wsj.com/cio/2018/11/28/estonias-cio-tackles-ai-strategy-for-government/> [Accessed 10 May 2019].
- Canadian Institute For Advanced Research (2017) *Pan-Canadian Artificial Intelligence Strategy*. [online] Available from: <https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy>. [Accessed 4 April 2019].
- Canadian Institute For Advanced Research (2019). *AI & Society Workshops: Call Two*. [online] Available from: <https://www.cifar.ca/ai/ai-society/workshops-call-two> [Accessed 10 May 2019].
- CDEI (2019). 'The Centre for Data Ethics and Innovation (CDEI) 2019/ 20 Work Programme' [online] Available from: <https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-cdei-2019-20-work-programme/the-centre-for-data-ethics-and-innovation-cdei-2019-20-work-programme> [Accessed 3 May 2019].
- Chantler, A., & Broadhurst, R. (2006). Social engineering and crime prevention in cyberspace. *Technical report*, Justice, Queensland University of Technology.
- Chen, A. (2017) 'The Human Toll of Protecting the Internet from the Worst of Humanity'. *The New Yorker*.
- Chesney, R., & Citron, D. (2018). Deep fakes: A looming crisis for national security, democracy and privacy? *Lawfare*.
- Christakis, N.A (2019) How AI Will Rewire Us. *The Atlantic Magazine, April 2019 Issue*. Available from: <https://www.theatlantic.com/magazine/archive/2019/04/robots-human-relationships/583204/>
- Christakis, N.A & Shirado, H. (2017) Locally Noisy Autonomous Agents Improve Global Human Coordination in Network Experiments. *Nature*. 545(7654), 370–374.

- Citron, D. K., & Pasquale, F. A. (2014). The scored society: due process for automated predictions. *Washington Law Review*, 89, 1–33.
- CNN. (2018). Self-driving electric bus propels Swiss town into the future. *CNN*. Available from: <https://edition.cnn.com/2018/06/27/sport/trapeze-self-driving-autonomous-electric-bus-switzerland-spt-intl/index.html>
- COMEST (2017). *Report of COMEST on Robotics Ethics*. UNESCO. Available from: <https://unesdoc.unesco.org/ark:/48223/pf0000253952>
- Conn, A. (2018) AI Should Provide a Shared Benefit for as Many People as Possible, Future of Life Institute, 10 Jan 2018 [online] Available at: <https://futureoflife.org/2018/01/10/shared-benefit-principle/> [Accessed 12 Aug. 2019].
- Corbe-Davies, S., Pierson, S., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of KDD '17*, Halifax, NS, Canada, August 13-17, 2017, 10 pages. DOI: 10.1145/3097983.3098095
- Council of Europe (2019a). Ad Hoc Committee on Artificial Intelligence – CAHAI. [online] Available at: <https://www.coe.int/en/web/artificial-intelligence/cahai> [Accessed 29 Oct. 2019].
- Council of Europe (2019b). Council of Europe's Work in progress. [online] Available at: <https://www.coe.int/en/web/artificial-intelligence/work-in-progress> [Accessed 29 Oct. 2019].
- Consultative Committee of the Convention for the Protection of Individuals with regard to the Processing of Personal Data (2019) Guidelines on Artificial Intelligence and Data Protection. Available from: <https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8>
- Cummings M. (2004). Automation bias in intelligent time critical decision support systems. In *AIAA: 1st Intelligent Systems Technical Conference. AIAA 2004, 20-22 September 2004, Chicago, Illinois*. pp. 6313.
- Curtis, J. (2016). Shocking dashcam footage shows Tesla 'Autopilot' crash which killed Chinese driver when futuristic electric car smashed into parked lorry. Daily Mail. <https://www.dailymail.co.uk/news/article-3790176/amp/Shocking-dashcam-footage-shows-Tesla-Autopilot-crash-killed-Chinese-driver-futuristic-electric-car-smashed-parked-lorry.html> [accessed 30/8/19].
- Danaher, J. (2017). Robotic rape and robotic child sexual abuse: Should they be criminalised? *Criminal Law and Philosophy*, 11(1), 71–95.
- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Available from: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scaps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>
- Datta, A., Tschantz and M.C., Datta, A. (2015). Automated Experiments on Ad Privacy Settings – A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*. 1, 92–112, DOI: 10.1515/popets-2015-0007
- Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J., and Hajkowicz, S. (2019). *Artificial Intelligence: Australia's Ethics Framework*. Available from: [https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting\\_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf](https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf)
- De Angeli, A. (2009). Ethical implications of verbal disinhibition with conversational agents. *Psychology Journal*, 7(1), 49–57.

De Angeli, A., & Brahnam, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with Computers*. 20(3), 302–310

De.digital. (2018). *The Federal Government's Artificial Intelligence Strategy*. [online] Available from: <https://www.de.digital/DIGITAL/Redaktion/EN/Standardartikel/artificial-intelligence-strategy.html>. [Accessed 10 May 2019].

Delvaux, M. (2017). 'With recommendations to the Commission on Civil Law Rules on Robotics' European Commission 2015/2103(INL).

Die Bundesregierung (2018) *Strategie Künstliche Intelligenz der Bundesregierung*.

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, 20: 1.

Digital Poland Foundation (2019). *Map of the Polish AI*. Digital Poland Foundation..

Duckworth, P., Graham, L., Osborne andM.AI (2019). Inferring Work Task Automatability from AI Expert Evidence. *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*. University of Oxford.

Dutton, T. (2018). An Overview of National AI Strategies. [online] Medium. Available at: <https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd> [Accessed 4 April 2019].

Ethics Commission (2017). Ethics's Commission's complete report on automated and connected driving. *Federal Ministry of Transport and Infrastructure*. Available from: <https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html?nn=187598>

Etzioni, A. and Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology*, 18(2), 149–156

European Commission (2012) Special Eurobarometer 382: Public Attitudes towards Robots. Eurobarometer Surveys [online] Available at: <https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/SPECIAL/surveyKy/1044/p/3>

European Commission (2017) Special Eurobarometer 460: Attitudes towards the impact of digitisation and automation on daily life [online] Available at: <https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/SPECIAL/surveyKy/2160>

European Commission (2018a). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe*. Available from: <https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe>

European Commission (2018b). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Coordinated Plan on Artificial Intelligence (COM(2018) 795 final)*. Available from: <https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence>

European Commission (2018c). High-level expert group on artificial intelligence: Draft ethics guidelines for trustworthy AI. Brussels. [online] Available from: [https://ec.europa.eu/futurium/en/system/files/ged/ai\\_hleg\\_draft\\_ethics\\_guidelines\\_18\\_december.pdf](https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_draft_ethics_guidelines_18_december.pdf) [Accessed 15/03/2019].

European Commission (2018d). EU Member States sign up to cooperate on Artificial Intelligence. [online] Available at: <https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence> [Accessed 30 Oct. 2019].

European Commission High-Level Expert Group on Artificial Intelligence (2019) *Ethics Guidelines for Trustworthy AI*. Available from: [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=58477](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477)

European Commission High-Level Expert Group on AI (2019b) Policy and Investment Recommendations for Trustworthy AI. Available from: <https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence>

European Parliament, Council and Commission, (2012). Charter of Fundamental Rights of the European Union. *Official Journal of the European Union*

European Parliament, 2017. EP Resolution with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). Available at: <http://www.europarl.europa.eu/>

Europol. (2017). *Serious and organised crime threat assessment*. Available from: <https://www.europol.europa.eu/socsta/2017/>.

Everett, J., Pizarro, D. and Crockett, M, (2017). Why are we reluctant to trust robots? *The Guardian*. Available from: <https://www.theguardian.com/science/head-quarters/2017/apr/24/why-are-we-reluctant-to-trust-robots>

Ezrachi, A., & Stucke, M. E. (2016). Two artificial neural networks meet in an online hub and change the future (of competition, market dynamics and society). *Oxford Legal Studies Research Paper*, No. 24/2017; *University of Tennessee Legal Studies Research Paper*, No. 323.

Farmer, J. D., & Skouras, S. (2013). An ecological perspective on the future of computer trading. *Quantitative Finance*. 13(3), 325–346

Felton, R. (2017). Limits of Tesla's Autopilot and driver error cited in fatal Model S crash. *Jalopnik*. Available from: [https://jalopnik.com/limits-of-teslas-autopilot-and-driver-error-cited-in-fa-1803806982#\\_ga=2.245667396.1174511965.1519656602-427793550.1518120488](https://jalopnik.com/limits-of-teslas-autopilot-and-driver-error-cited-in-fa-1803806982#_ga=2.245667396.1174511965.1519656602-427793550.1518120488)

Felton, R. (2018). Two years on, a father is still fighting Tesla over autopilot and his son's fatal crash. *Jalopnik*. Available from: <https://jalopnik.com/two-years-on-a-father-is-still-fighting-tesla-over-aut-1823189786>

Ferrara, E. (2015). *Manipulation and abuse on social media*

Floridi, L. (2016). Tolerant paternalism: Pro-ethical design as a resolution of the dilemma of toleration. *Science and Engineering Ethics*. 22(6), 1669–1688.

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083).

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford review*. 5. Oxford, Oxford University Press.

Ford, M. (2009) *The Lights in the Tunnel: Automation, Accelerating Technology, and the Economy of the Future*.

Foundation for Law & International Affairs (2017) China's New Generation of Artificial Intelligence Development Plan. *FLIA*. [online] Available FROM: <https://flia.org/wp-content/uploads/2017/07/A-New-Generation-of-Artificial-Intelligence-Development-Plan-1.pdf>

Frey, C. B. and Osborne, M. A. (2013). The Future of Employment: How Susceptible Are Jobs to Computerisation? *Oxford Martin Programme on the Impacts of Future Technology*.

Furman, J & Seamans, R. (2018). AI and the Economy. *NBER working paper* no.24689

Future of Life Institute (2019). National and International AI Strategies. *Future of Life Institute*. [online] Available from: <https://futureoflife.org/national-international-ai-strategies/> [Accessed 28 Apr. 2019].

G7 Canadian Presidency (2018). *Charlevoix Common Vision for the Future of Artificial Intelligence*.

G20 (2019) G20 Ministerial Statement on Trade and Digital Economy: Annex. Available from: <https://www.mofa.go.jp/files/000486596.pdf>

Gagan, O. (2018) Here's how AI fits into the future of energy, World Economic Forum, 25 May 2018 [Online] Available at: <https://www.weforum.org/agenda/2018/05/how-ai-can-help-meet-global-energy-demand> [Accessed on 13 Aug. 2019].

Garfinkel, S. (2017). Hackers are the real obstacle for self-driving vehicles. *MIT Technology Review*. Available from: <https://www.technologyreview.com/s/608618/hackers-are-the-real-obstacle-for-self-driving-vehicles/>

Gibbs, S. (2017). Tesla Model S cleared by safety regulator after fatal Autopilot crash. *The Guardian*. Available from: <https://www.theguardian.com/technology/2017/jan/20/tesla-model-s-cleared-auto-safety-regulator-after-fatal-autopilot-crash>

Gillespie T. (2014). The relevance of algorithms. In Gillespie, T., Boczkowski, P. J., Foot, K. A. (eds.) (2014). *Media technologies: essays on communication, materiality, and society*. Cambridge, MA: MIT Press. pp. 167-194.

Gogarty, B., & Hagger, M. (2008). The laws of man over vehicles unmanned: The legal response to robotic revolution on sea, land and air. *Journal of Law, Information and Science*, 19, 73–145.

Goldhill, O. (2016). Can we trust robots to make moral decisions? *Quartz*. Available from: <https://qz.com/653575/can-we-trust-robots-to-make-moral-decisions/>

UK Government Office for Science (2015) Artificial intelligence: opportunities and implications for the future of decision making. Available from: [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf) [Accessed 13 Aug. 2019].

GOV.UK. (2018a). *AI Sector Deal*. [online] Available from <https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal> [Accessed 10 May 2019].

GOV.UK. (2018b). *Centre for Data Ethics and Innovation (CDEI)*. [online] Available from: <https://www.gov.uk/government/groups/centre-for-data-ethics-and-innovation-cdei> [Accessed 10 May 2019].

GOV.UK (2019). The UK's Industrial Strategy. *GOV.UK*. [online] Available from: <https://www.gov.uk/government/topical-events/the-uks-industrial-strategy> [Accessed 10 May 2019].

Government Offices of Sweden (2018). National approach to artificial intelligence. *Ministry of Enterprise and Innovation.*

Graetz, G. and Michaels, G. (2015). Robots at Work. *Centre for Economic Performance Discussion Paper No. 1335.*

Gray, M. L. and Suri, S. (2019). *Ghost Work*, Houghton Mifflin Harcourt.

Greene, J. D., Sommerville, R. B., Nystrom, L., Darley, J., and Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105–2108. doi:10.1126/science.1062872.

Guiltinan, J. (2009). Creative destruction and destructive creations: Environmental ethics and planned obsolescence. *Journal of Business Ethics*. 89 (1). pp.1928.

Gurney, J. K., (2013). Sue My Car, Not Me: Products Liability and Accidents Involving Autonomous Vehicles. unpublished manuscript

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2016). The off-switch game. In: *IJCAI-ECAI-2018: International Joint Conference on Artificial Intelligence. IJCAI-ECAI-2018, 13-19 July 2018, Stockholm, Sweden.*

Hallaq, B., Somer, T., Osula, A., Ngo, K., & Mitchener-Nissen, T. (2017). Artificial intelligence within the military domain and cyber warfare. In: 16th European Conference on Cyber Warfare and Security (ECCWS 2017), 29-30 June 2017, Dublin, Ireland. Published in: Proceedings of 16th European Conference on Cyber Warfare and Security.

Hallevy, G. (2010) The Criminal Liability of Artificial Intelligence Entities (February 15, 2010). Available at SSRN: <https://ssrn.com/abstract=1564096> or <http://dx.doi.org/10.2139/ssrn.1564096>

Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*. 45(1): 1–23.

Harambam, J., Helberger, N., and Van Hoboken, J. (2018). Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133).

Hardt, M. (2014). *How Big Data is Unfair*. Medium. [online] Available from [accessed 9 Apr. 2019]

Hart, R. D. (2018). Who's to blame when a machine botches your surgery? *Quartz*. Available from: <https://qz.com/1367206/whos-to-blame-when-a-machine-botches-your-surgery/>

Hawkins, A. J. (2019). California's self-driving car reports are imperfect, but they're better than nothing. *The Verge*. Available from: <https://www.theverge.com/2019/2/13/18223356/california-dmv-self-driving-car-disengagement-report-2018>

Hawksworth, J. and Fertig, Y. (2018) What will be the net impact of AI and related technologies on jobs in the UK? PwC UK Economic Outlook, July 2018.

Hern, A. (2016). 'Partnership on AI' formed by Google, Facebook, Amazon, IBM and Microsoft. *The Guardian*. Available from: <https://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms>

Hess, A. (2016). On Twitter, a Battle Among Political Bots. *The New York Times*. Available from: <https://www.nytimes.com/2016/12/14/arts/on-twitter-a-battle-among-political-bots.html>

Human Rights Watch. (2018). 'Eradicating ideological viruses': China's campaign of repression against Xinjiang's Muslims. *Technical report*, Human Rights Watch.

IEEE (2019). *Homepage* [online] Available from: <https://www.ieee.org> [Accessed 11 Mar2019].

Iglinski, H., Babiak, M. (2017). Analysis of the Potential of Autonomous Vehicles in Reducing the Emissions of Greenhouse Gases in Road Transport. *Procedia Eng.*192, 353–358.

International Telecommunication Union (2018). *AI for Good Global Summit 2018* [online] Available from: <https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx> [Accessed 14 May 2019].

International Telecommunication Union (2018). United Nations Activities on Artificial Intelligence [online]. Available from: <http://www.itu.int/pub/S-GEN-UNACT-2018-1> [Accessed 12 November 2019]

Isaac, M. (2016). Self-driving truck's first mission: a 120-mile beer run. *New York Times*. Available from: <https://www.nytimes.com/2016/10/26/technology/self-driving-trucks-first-mission-a-beer-run.html>

Israel Innovation Authority (2019). *Israel Innovation Authority 2018-19 Report*. [online] Available from: <https://innovationisrael.org.il/en/news/israel-innovation-authority-2018-19-report> [Accessed 10 May 2019].

Iyengar, S., Sood, G., and Lelkes, Y. (2012). Affect, not ideology: Social identity perspective on polarization. *Public Opinion Quarterly*. 76(3),405.

Jacobs, S. B. (2017) The Energy Prosumer, 43Ecology L. Q.519.

Japanese Strategic Council for AI Technology (2017). *Artificial Intelligence Technology Strategy*. Available from: <https://www.nedo.go.jp/content/100865202.pdf>

Johnson, A., and Axinn, S. (2013). The Morality of Autonomous Robots. *Journal of Military Ethics*. 12 (2), 129-141

Johnston, A. K. (2015). Robotic seals comfort dementia patients but raise ethical concerns. *KALW*. Available from: <https://www.kalw.org/post/robotic-seals-comfort-dementia-patients-raise-ethical-concerns#stream/0>

JSAI (2017). *Ethical Guidelines*. [online] Available from: <http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf> [Accessed 7 May19].

JSAI (2019). *Overview: Inaugural Address of President Naohiko Uramoto, Artificial Intelligence expanding its scope and impact in our society*. [online] Available from: <https://www.ai-gakkai.or.jp/en/about/about-us/> [Accessed 11 May 2019].

Kayali, L. (2019). *Next European Commission takes aim at AI*. [online] POLITICO. Available at: <https://www.politico.eu/article/ai-data-regulator-rules-next-european-commission-takes-aim/> [Accessed 27 Aug. 2019].

Kenyan Wall Street (2018). Kenya Govt unveils 11 Member Blockchain & AI Taskforce headed by Bitange Ndemo. *Kenyan Wallstreet*. [online. Available from: <https://kenyanwallstreet.com/kenya-govt-unveils-11-member-blockchain-ai-taskforce-headed-by-bitange-ndemo/> [Accessed 6 May 2019].

Khakurel, J., Penzenstadler, B., Porras, J., Knutas, A., and Zhang, W. (2018). The Rise of Artificial Intelligence under the Lens of Sustainability. *Technologies*. 6(4), 100.

Khosravi, B. (2018). Autonomous cars won't work – until we have 5G. *Forbes*. Available from: <https://www.forbes.com/sites/bijankhosravi/2018/03/25/autonomous-cars-wont-work-until-we-have-5g>

King, T.C., Aggarwal, N., Taddeo, M. et al. (2019). Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Sci Eng Ethics*. pp.1-32

Kingston, J. K. C. (2018) Artificial Intelligence and Legal Liability. Available at: <https://arxiv.org/ftp/arxiv/papers/1802/1802.07782.pdf> [Accessed 17/08/19].

Kitwood, T. (1997). *Dementia Reconsidered: The Person Comes First*. Buckingham, Open University Press.

Knight, W. (2019). *The World Economic Forum wants to develop global rules for AI*. [online] MIT Technology Review. Available at: <https://www.technologyreview.com/s/613589/the-world-economic-forum-wants-to-develop-global-rules-for-ai/> [Accessed 20 Aug. 2019].

Kroll, J.A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133).

Lalji, N. (2015). Can we learn about empathy from torturing robots? This MIT researcher is giving it a try. *YES! Magazine*. Available from: <http://www.yesmagazine.org/happiness/should-we-be-kind-to-robots-katedarling>.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*. 10, (1096)

LaRosa, E., & Danks, D. (2018). Impacts on Trust of Healthcare AI. In: *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. AEIS: 2018, 1-3 February, 2018, New Orleans, USA*.

Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., Cedering Ångström, R. (2019). Sustainable AI report. *AI Sustainability Centre*. Available from: <http://www.aisustainability.org/wp-content/uploads/2019/04/SUSTAINABLE-AI.pdf>

Lashbrook, A. (2018). AI-driven dermatology could leave dark-skinned patients behind. *The Atlantic*. Available from: <https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/>

Leggett, T. (2018) Who is to blame for 'self-driving car' deaths? BBC Business News. 22 May 2018. Available at: <https://www.bbc.co.uk/news/business-44159581>

Le Miere, J. (2017). Russia is developing autonomous 'swarms of drones' it calls an inevitable part of future warfare. [online] *Newsweek*. Available at: <https://www.newsweek.com/drones-swarm-autonomous-russia-robots-609399> [Accessed 26 Apr. 2019].

Leontief, Wassily,. (1983). National Perspective: The Definition of Problems and Opportunities.. *The Long-Term Impact of Technology on Employment and Unemployment*. Washington, DC: The National Academies Press. doi: 10.17226/19470.

Lerner, S. (2018). NHS might replace nurses with robot medics such as carebots: could this be the future of medicine? *Tech Times*. Available from: <https://www.techtimes.com/articles/229952/20180611/nhs-might-replace-nurses-with-robot-medics-such-as-carebots-could-this-be-the-future-of-medicine.htm>

- Levin, S. (2018). Video released of Uber self-driving crash that killed woman in Arizona. *The Guardian*. Available from: <https://www.theguardian.com/technology/2018/mar/22/video-released-of-uber-self-driving-crash-that-killed-woman-in-arizona>
- Li, H., Milani, S., Krishnamoorthy, V., Lewis, M., & Sycara, K. (2019). Perceptions of Domestic Robots' Normative Behavior Across Cultures. In: AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. *AEIS: 2019, 27-28 January, 2019, Honolulu, Hawaii, USA*. Available here: [http://www.aies-conference.com/2019/wp-content/papers/main/AIES-19\\_paper\\_232.pdf](http://www.aies-conference.com/2019/wp-content/papers/main/AIES-19_paper_232.pdf)
- Li, S., Williams, J. (2018). Despite what Zuckerberg's testimony may imply, AI Cannot Save Us. *Electronic Frontier Foundation*. Available from: <https://www.eff.org/deeplinks/2018/04/despite-whatzuckerbergs-testimony-may-imply-ai-cannot-save-us>
- Lim, D., (2019). Killer Robots and Human Dignity. In: AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. *AEIS: 2019, 27-28 January, 2019, Honolulu, Hawaii, USA*.
- Lin, P. (2014). What if your autonomous car keeps routing you past Krispy Kreme? *The Atlantic*. Available from: [https://finance.yahoo.com/news/autonomous-car-keeps-routing-past-130800241.html;\\_ylt=A2KJ3CUL199SkjsAexPQtDMD?guccounter=1&guce](https://finance.yahoo.com/news/autonomous-car-keeps-routing-past-130800241.html;_ylt=A2KJ3CUL199SkjsAexPQtDMD?guccounter=1&guce)
- Lin, P., Jenkins, R., & Abney, K. (2017). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press.
- Lin, T. C. W. (2017). The new market manipulation. *Emory Law Journal*, 66, 1253.
- Loh, W. & Loh, J. ( 2017). Autonomy and responsibility in hybrid systems. In P. Lin, et al. (Eds.), *Robot ethics 2.0*. New York, NY: Oxford University Press: 35–50.
- Lokhorst, G.-J. and van den Hoven, J. (2014) Chapter 9: Responsibility for Military Robots. In *Robot Ethics: The Ethical and Social Implications of Robotics* edited by Lin, Abney and Bekey (10 Jan. 2014, MIT Press).
- Malta AI (2019). *Malta AI: Towards a National AI Strategy* [online] Available at: <https://malta.ai> [Accessed 10 May 2019].
- Manikonda, L., Deotale, A., & Kambhampati, S.. (2018). What's up with Privacy? User Preferences and Privacy Concerns in Intelligent Personal Assistants. In: AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. *AEIS: 2018, 1-3 February, 2018, New Orleans, USA*.
- Marda, V., (2018). Artificial intelligence policy in India: a framework for engaging the limits of data-driven decision-making. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133).
- Marshall, A. and Davies, A. (2018). Lots of lobbies and zero zombies: how self-driving cars will reshape cities. *Wired*. Available from: <https://www.wired.com/story/self-driving-cars-cities/>
- Martinho-Truswell, E., Miller, H., Nti Asare, I., Petheram, A., Stirling, R., Gómez Mont, G. and Martinez, C. (2018). *Towards an AI Strategy in Mexico: Harnessing the AI Revolution*.
- Mattheij, J. (2016) 'Another Way Of Looking At Lee Sedol vs AlphaGo'. Jacques Mattheij: Technology, Coding and Business. Blog. 17th March 2016.
- Matthias, A. (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, Sept 2004, Vol. 6, Issue 3, pp.175-183.
- Mazzucato, M. (2018) Mission-Oriented Research & Innovation in the European Union. European Commission: Luxembourg.

Mbadiwe, T. (2017). The potential pitfalls of machine learning algorithms in medicine. *Pulmonology Advisor*. Available from: <https://www.pulmonologyadvisor.com/home/topics/practice-management/the-potential-pitfalls-of-machine-learning-algorithms-in-medicine/>

McAllister, A. (2017). Stranger than science fiction: The rise of AI interrogation in the dawn of autonomous robots and the need for an additional protocol to the UN convention against torture. *Minnesota Law Review*. 101, 2527–2573.

McCarty, N. M., Poole, K. T., and Rosenthal, H. (2016). *Polarized America: The Dance Of Ideology And Unequal Riches*. Cambridge, MA: MIT Press, 2nd edition.

Meisner, E. M. (2009). *Learning controllers for human–robot interaction*. PhD thesis. Rensselaer Polytechnic Institute.

México Digital (2018). Estrategia de Inteligencia Artificial MX 2018. [online] gob.mx. Available from: <https://www.gob.mx/mexicodigital/articulos/estrategia-de-inteligencia-artificial-mx-2018> [Accessed 6 May 2019].

Millar, J. (2016). *An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars*. 30(8), 787-809.

Min, W. (2018) Smart Policies for Harnessing AI, OECD-Forum, 17 Sept 2018 [online] Available from: <https://www.oecd-forum.org/users/68225-wonki-min/posts/38898-harnessing-ai-for-smart-policies> [Accessed 12 Aug. 2019].

Ministry of Economic Affairs and Employment of Finland (2017). Finland's Age of Artificial Intelligence. Available from: [https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap\\_47\\_2017\\_verkkojulkaisu.pdf](https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf)  
Ministry of Economic Affairs and Employment of Finland (2018a). *Artificial intelligence programme*. [online] Available from: <https://tem.fi/en/artificial-intelligence-programme> [Accessed 26 Apr. 2019].

Ministry of Economic Affairs and Employment of Finland (2018b). *Work in the Age of Artificial Intelligence*. Available from: <https://www.google.com/search?client=safari&rls=en&q=work+in+the+age+of+artificial+intelligence&ie=UTF-8&oe=UTF-8>

Mizoguchi, R. (2004). The JSAI and AI activity in Japan. *IEEE Intelligent Systems* 19 (2).

Moon, M., (2017). Judge allows pacemaker data to be used in arson trial. *Engadget*. Available from: <https://www.engadget.com/2017/07/13/pacemaker-arson-trial-evidence/>

National Science & Technology Council (2019) The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update. Available from: <https://www.whitehouse.gov/wp-content/uploads/2019/06/National-AI-Research-and-Development-Strategic-Plan-2019-Update-June-2019.pdf>

NTSB (2018) Preliminary Report Released for Crash Involving Pedestrian, Uber Technologies, Inc., Test Vehicle. National Transport Safety Board News Release. May 24, 2018. Available at: <https://www.ntsb.gov/news/press-releases/Pages/NR20180524.aspx>

Nemitz, P., (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133).

Newman, E. J., Sanson, M., Miller, E. K., Quigley-McBride, A., Foster, J. L., Bernstein, D. M., and Garry, M. (2014). People with easier to pronounce names promote truthiness of claims. *PLoS ONE*.9(2).

NITI Aayog (2018). *National Strategy for Artificial Intelligence #AIFORALL*.

Nevejans, N. et al. (2018). *Open letter to the European Commission on Artificial Intelligence and Robotics*.

New America. (2018). Translation: *Chinese government outlines AI ambitions through 2020*. [online] Available from: <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-government-outlines-ai-ambitions-through-2020/> [Accessed 27 Apr. 2019].

NHS Digital. (2019). *Widening Digital Participation*. NHS Digital. Available from: <https://digital.nhs.uk/about-nhs-digital/our-work/transforming-health-and-care-through-technology/empower-the-person-formerly-domain-a/widening-digital-participation>

NHS' Topol Review. (2019). *Preparing the healthcare workforce to deliver the digital future*. Available from: <https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf>

Nordic cooperation (2018). *AI in the Nordic-Baltic region*. [online] Available from: <https://www.norden.org/en/declaration/ai-nordic-baltic-region> [Accessed 26 Apr. 2019].

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C. and Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 115, E5716–E5725.

O'Carroll, T. (2017). Mexico's misinformation wars. *Medium*. Available from: <https://medium.com/amnesty-insights/mexico-s-misinformation-wars- cb748ecb32e9#.n8pi52hot>

O'Connor, T. (2017). Russia is building a missile that can makes its own decisions. [online] *Newsweek*. Available from: <https://www.newsweek.com/russia-military-challenge-us-china-missile-own-decisions-639926> [Accessed 26 Apr. 2019].

O'Donoghue, J. (2010). E-waste is a growing issue for states. *Deseret News*. Available from: <http://www.deseretnews.com/article/700059360/E-waste-is-a-growing-issue-for-states.html?pg=1>

O'Kane, S (2018). Tesla defends Autopilot after fatal Model S crash. *The Verge*. Available from: <https://www.theverge.com/2018/3/28/17172178/tesla-model-x-crash-autopilot-fire-investigation>

O'Neil, C., (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Crown Publishers.

O'Neill, S. (2018). As insurers offer discounts for fitness trackers, wearers should step with caution. *National Public Radio*. Available from: <https://www.npr.org/sections/health-shots/2018/11/19/668266197/as-insurers-offer-discounts-for-fitness-trackers-wearers-should-step-with-cautio?t=1557493660570>

OECD (2013) Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data [OECD/LEGAL/0188]

OECD (n.d.) OECD initiatives on AI [online] Available at: <http://www.oecd.org/going-digital/ai/> [Accessed 13 Aug. 2019].

- Ori.(2014a). If Death by Autonomous Car is Unavoidable, Who Should Die? Reader Poll Results. *Robohub.org*. Available from: <http://robohub.org/if-a-death-by-an-autonomous-car-is-unavoidable-who-should-die-results-from-our-reader-poll/>.
- Ori. (2014b). My (autonomous) car, my safety: Results from our reader poll. *Robohub.org*.. Available from: <http://robohub.org/my-autonomous-car-my-safety-results-from-our-reader-poll>
- Orseau, L. & Armstrong, S. (2016). Safely interruptible agents. In: *Uncertainty in artificial intelligence: 32nd Conference (UAI)*. UAI: 2016, June 25-29, 2016, New York City, NY, USA. AUAI Press 2016
- Ovanessoff, A. and Plastino, E. (2017). How Artificial Intelligence Can Drive South America's Growth. *Accenture*.
- Oxford Insights (2019) Government Artificial Intelligence Readiness Index. Available from: [https://ai4d.ai/wp-content/uploads/2019/05/ai-gov-readiness-report\\_v08.pdf](https://ai4d.ai/wp-content/uploads/2019/05/ai-gov-readiness-report_v08.pdf)
- Pagallo, U. (2017). Apples, oranges, robots: four misunderstandings in today's debate on the legal status of AI systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133).
- Pariser E. (2011). *The filter bubble: what the Internet is hiding from you*. London, UK, Penguin.
- Park, M. (2017). Self-driving bus involved in accident on its first day. *CNN Business*. Available from: <https://money.cnn.com/2017/11/09/technology/self-driving-bus-accident-las-vegas/index.html>
- Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA, Harvard University Press.
- Personal Data Protection Commission Singapore (2019). *A Proposed Model Artificial Intelligence Governance Framework*. Available from: <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/A-Proposed-Model-AI-Governance-Framework-January-2019.pdf>
- Pfleger, P. (2018). Transportation workers form coalition to stop driverless buses in Ohio. *WOSU Radio*. Available from: <https://radio.wosu.org/post/transportation-workers-form-coalition-stop-driverless-buses-ohio#stream/0>
- Pham, T., Gorodnichenko, Y. and Talavera, O. (2018). *Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection*. NBER Working Papers w24631. The National Bureau of Economic Research; Cambridge, MA.
- Piesing, M. (2014). Medical robotics: Would you trust a robot with a scalpel? *The Guardian*. Available at: <https://www.theguardian.com/technology/2014/oct/10/medical-robots-surgery-trust-future>
- Plantera, F. (2017). Artificial Intelligence is the next step for e-governance in Estonia, State adviser reveals.[online] *e-Estonia*. Available from: <https://e-estonia.com/artificial-intelligence-is-the-next-step-for-e-governance-state-adviser-reveals/>. [Accessed 28 Apr. 2019].
- Polonski, V. (2017). #MacronLeaks changed political campaigning. Why Macron succeeded and Clinton failed. *World Economic Forum*. Available from: <https://www.weforum.org/agenda/2017/05/macronleaks-have-changed-political-campaigning-why-macron-succeeded-and-clinton FAILED>
- Press Association (2019). Robots and AI to give doctors more time with patients, says report. *The Guardian*. Available from: <https://www.theguardian.com/society/2019/feb/11/robots-and-ai-to-give-doctors-more-time-with-patients-says-report>

- ProPublica (2016). Machine Bias. *ProPublica*. Available from: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*. 20: 5. <https://doi.org/10.1007/s10676-017-9430-8>
- Ramchurn, S. D. et al. (2013) AgentSwitch: Towards Smart Energy Tariff Selection. Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.
- Reuters (2019). *G7 urges tight regulations for digital currencies, agrees to tax digital giants locally*. [online] VentureBeat. Available at: <https://venturebeat.com/2019/07/19/g7-urges-tight-regulations-for-digital-currencies-agrees-to-tax-digital-giants-locally/> [Accessed 27 Aug. 2019].
- Riedl, M.O., and Harrison, B. (2017). Enter the matrix: A virtual world approach to safely interruptable autonomous systems. *arXiv*. preprint arXiv:1703.10284
- Roberts, S. (2016) 'Digital Refuse: Canadian Garbage, Commercial Content Moderation and the Global Circulation of Social Media's Waste'. Media Studies Publications.
- Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schultz, J., Hale, T. M., and Stern M.J. (2015) Digital Inequalities and Why They Matter. *Information, Communication & Society*. 18 (5), 569-592. <http://dx.doi.org/10.1080/1369118X.2015.1012532>
- SAE International. (2018). SAE International releases updated visual chart for its 'levels of driving automation' standard for self-driving vehicles. *SAE International*. Available from: <https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-'levels-of-driving-automation'-standard-for-self-driving-vehicles>
- Sage, A. (2018). Waymo unveils self-driving taxi service in Arizona for paying customers. *Reuters*. Available from: <https://www.reuters.com/article/us-waymo-selfdriving-focus/waymo-unveils-self-driving-taxi-service-in-arizona-for-paying-customers-idUSKBN1O41M2>
- Saidot (2019). *About us* [online] Available from: <https://www.saidot.ai/about-us> [Accessed 3 May 2019]. Salvage, M. (2019). Call for poor and disabled to be given fitness trackers. *The Guardian*. Available from: <https://www.theguardian.com/inequality/2019/may/04/fitbits-nhs-reduce-inequality-health-disability-poverty>
- Sample, I. (2017). Computer says no: why making AIs fair, accountable and transparent is crucial. *The Guardian*. Available from: <https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial>
- Sample, I. (2017). Give robots an 'ethical black box' to track and explain decisions, say scientists. *The Guardian*. Available from: <https://www.theguardian.com/science/2017/jul/19/give-robots-an-ethical-black-box-to-track-and-explain-decisions-say-scientists>
- Santos-Lang, C. (2002). Ethics for Artificial Intelligences. In Wisconsin State-Wide technology Symposium 'Promise or Peril?'. *Reflecting on computer technology: Educational, psychological, and ethical implications*. Wisconsin, USA.
- Sarmah, H. (2019). Looking East: How South Korea Is Making A Strategic Move In AI. [online] *Analytics India Magazine*. Available from: <https://www.analyticsindiamag.com/looking-east-how-south-korea-is-making-a-strategic-move-for-ai-leadership/> [Accessed 28 Apr. 2019].

Sathe G. (2018). Cops in India are using artificial intelligence that can identify you in a crowd. *Huffington Post*. Available at: [https://www.huffingtonpost.in/2018/08/15/facial-recognition-is-shaking-up-criminals-in-punjab-but-should-you-worry-too\\_a\\_23502796/](https://www.huffingtonpost.in/2018/08/15/facial-recognition-is-shaking-up-criminals-in-punjab-but-should-you-worry-too_a_23502796/).

Sauer, G. (2017). A Murder Case test's Alexa's Devotion to your Privacy. *Wired*. Available from <https://www.wired.com/2017/02/murder-case-tests-alexa-devotion-privacy/>

Scherer, M. U. (2016) Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies, 29 *Harv. J. L. & Tech.* 353 (2015-2016)

Scheutz, M. (2012). The inherent dangers of unidirectional emotional bonds between humans and social robots. In: Lin, P., Abney, K. and Bekey, G. (eds.). *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, pp.205-221.

Schmitt, M.N., (2013). *Tallinn manual on the international law applicable to cyber warfare*. Cambridge University Press.

Schönberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology* 27 (2), 171–203.  
<https://doi.org/10.1093/ijlit/eaz004>

Selbst, A. D. and Baracas. S. (2018). The intuitive appeal of explainable machines. 87 *Fordham Law Review* 1085 Preprint, available from: <https://ssrn.com/abstract=3126971>

Selbst, A. D. and Powles, J. (2017) Meaningful information and the right to explanation. *Int. Data Privacy Law* 7, 233–242. (doi:10.1093/idpl/ipx022)

Selinger, E. and Hartzog, W. (2017). Obscurity and privacy. In: Pitt, J. and Shew, A. (eds.). *Spaces for the Future: A Companion to Philosophy of Technology*, New York: Routledge.

Servoz, M. (2019) The Future of Work? Work of the Future! On How Artificial Intelligence, Robotics and Automation Are Transforming Jobs and the Economy in Europe, 10 May 2019. Available at: [https://ec.europa.eu/epsc/publications/other-publications/future-work-work-future\\_en](https://ec.europa.eu/epsc/publications/other-publications/future-work-work-future_en) [Accessed 13 Aug. 2019].

Seth, S. (2017). Machine Learning and Artificial Intelligence Interactions with the Right to Privacy. *Economic and Political Weekly*, 52(51), 66–70

Sharkey, A., and Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*. 14 (1): 27-40.

Sharkey, N., Goodman, M., & Ross, N. (2010). The coming robot crime wave. *IEEE Computer Magazine*. 43(8), 6–8.

Shepherdson, D. and Somerville, H. (2019) Uber not criminally liable in fatal 2018 Arizona self-driving crash – prosecutors. Reuters News. March 5, 2019. Available from: <https://uk.reuters.com/article/uk-uber-crash-autonomous/uber-not-criminally-liable-in-fatal-2018-arizona-self-driving-crash-prosecutors-idUKKCN1QM2P4>

Shewan, D. (2017). Robots will destroy our jobs – and we're not ready for it. *The Guardian*. Available from: <https://www.theguardian.com/technology/2017/jan/11/robots-jobs-employees-artificial-intelligence>.

Smart Dubai (2019a). *AI Ethics*. [online] Available from: <https://www.smartdubai.ae/initiatives/ai-ethics> [Accessed 10 May 2019].

Smartdubai.ae. (2019b). *AI Ethics Self Assessment*. [online] Available from: <https://www.smartdubai.ae/self-assessment> [Accessed 12 May 2019].

Smith, A., & Anderson, J. (2014). *AI, Robotics, and the Future of Jobs*. Pew Research Center

Smith, B. (2018). Facial recognition technology: The need for public regulation and corporate responsibility. *Microsoft on the Issues*. Available from: <https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/>

Snaith, E. (2019). Robot rolls into hospital ward and tells 97-year-old man he is dying. *The Independent*. Available from: <https://www.independent.co.uk/news/world/americas/robot-grandfather-dying-san-francisco-hospital-ernesta-quintana-california-a8815721.html>

Solon, O. (2018). Who's driving? Autonomous cars may be entering the most dangerous phase. *The Guardian*. Available from: <https://www.theguardian.com/technology/2018/jan/24/self-driving-cars-dangerous-period-false-security>

Sparrow, R.. (2002). The march of the robot dogs. *Ethics and Information Technology*. 4 (4), 305–318.

Sparrow, R., and Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*. 16, 141-161.

Spatt, C. (2014). Security market manipulation. *Annual Review of Financial Economics*, 6(1), 405–418.

Stahl, B.C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*. 86, 152-161.

Stilgoe, J. and Winfield, A. (2018). Self-driving car companies should not be allowed to investigate their own crashes. *The Guardian*. Available from: <https://www.theguardian.com/science/political-science/2018/apr/13/self-driving-car-companies-should-not-be-allowed-to-investigate-their-own-crashes>

Strubell, E., Ganesh, A. and McCallum, A. (2019) Energy and Policy Considerations for Deep Learning in NLP, arXiv:1906.02243

Swedish AI Council. (2019). *Swedish AI Council*. [online] Available from: <https://swedishaicouncil.com> [Accessed 10 May 2019].

Taddeo, M. (2017). Trusting Digital Technologies Correctly. *Minds & Machines*. 27 (4), 565.

Taddeo, M. and Floridi, L. (2018) How AI can be a force for good. *Science* vol. 361, issue 6404, pp.751-752. DOI: 10.1126/science.aat5991

Task Force on Artificial Intelligence of the Agency for Digital Italy (2018). *White Paper on Artificial Intelligence at the service of citizens*.

Tesla. (nd). Support: autopilot. *Tesla*. Available from: <https://www.tesla.com/support/autopilot>

The Danish Government (2018). *Strategy for Denmark's Digital Growth*. Ministry of Industry, Business and Financial Affairs. Available from: [https://eng.em.dk/media/10566/digital-growth-strategy-report\\_uk\\_web-2.pdf](https://eng.em.dk/media/10566/digital-growth-strategy-report_uk_web-2.pdf)

The Danish Government (2019). *National Strategy for Artificial Intelligence*. Ministry of Finance and Ministry of Industry, Business and Financial Affairs. Available from: [https://eng.em.dk/media/13081/305755-gb-version\\_4k.pdf](https://eng.em.dk/media/13081/305755-gb-version_4k.pdf)

The Foundation for Responsible Robotics (2019). About us: *Our mission* [online] Available from: <http://responsiblerobotics.org/about-us/mission/> [Accessed 11 Mar2019].

The Future of Life Institute (n.d.) AI Policy Challenges and Recommendations. Available at: <https://futureoflife.org/ai-policy-challenges-and-recommendations/#top> [Accessed 12/08/19].

The Future of Life Institute (2019). *Background: Benefits and Risks of Artificial Intelligence*. [online]. Available from: <https://futureoflife.org/background/benefits-risks-of-artificial-intelligence> [Accessed 19 Mar.2019].

The Future Society (2019). *About us* [online] Available from: <https://thefuturesociety.org/about-us> [Accessed 11/03/2019].

The Institute of Electrical and Electronics Engineers (IEEE) (2017). *Ethically Aligned Design: First Edition. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. (EADv2)*.

The Institute of Electrical and Electronics Engineers (IEEE) (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (EAD1e)*

The Institute for Ethical AI & Machine Learning (2019). *Homepage* [online] Available from: <https://ethical.institute/index.html> [Accessed 11 Mar.2019].

The Partnership on AI (2019). *About us* [online] Available from: <https://www.partnershiponai.org/about/> [Accessed 11 Mar.2019].

The White House (2016) *Artificial Intelligence, Automation, and the Economy* [online] Available from: <https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF> [Accessed 12 Aug. 2019].

The White House (2019a). *Accelerating America's Leadership in Artificial Intelligence*. [online] Available from: <https://www.whitehouse.gov/articles/accelerating-americas-leadership-in-artificial-intelligence/> [Accessed 28 Apr. 2019].

The White House (2019b). *Artificial Intelligence for the American People* [online] Available from: <https://www.whitehouse.gov/ai/>. [Accessed 28 Apr. 2019].

Thiagarajan, K. (2019). The AI program that can tell whether you may go blind. *The Guardian*. Available from: <https://www.theguardian.com/world/2019/feb/08/the-ai-program-that-can-tell-whether-you-are-going-blind-algorithm-eye-disease-india-diabetes>

Thielman, S. (2017). The customer is always wrong: Tesla lets out self-driving car data – when it suits. *The Guardian*. Available from: <https://www.theguardian.com/technology/2017/apr/03/the-customer-is-always-wrong-tesla-lets-out-self-driving-car-data-when-it-suits>

Thomson, J. (1976). Killing, letting die, and the trolley problem. *The Monist*. 59, 204–217.

Thurman N. (2011). Making 'The Daily Me': technology, economics and habit in the mainstream assimilation of personalized news. *Journalism*. 12, 395–415.

Tindera, M. (2018). Government data says millions of health records are breached every year. *Forbes*. <https://www.forbes.com/sites/michelatindera/2018/09/25/government-data-says-millions-of-health-records-are-breached-every-year/#209fca3716e6>

Torres Santeli, J. and Gerdon, S. (2019). *5 challenges for government adoption of AI*. [online] World Economic Forum. Available at: <https://www.weforum.org/agenda/2019/08/artificial-intelligence-government-public-sector/> [Accessed 27 Aug. 2019].

TUM (2019). *New Research Institute for Ethics in Artificial Intelligence* [Press Release]. Available from: <https://www.wi.tum.de/new-research-institute-for-ethics-in-artificial-intelligence/> [Accessed 11 Mar.2019].

Turkle, S., Taggart, W., Kidd, C.D. and Dasté, O.,(2006). Relational Artifacts with Children and Elders: The Complexities of Cyber companionship. *Connection Science*, 18 (4) pp 347-362.

UAE Government (2018). *UAE Artificial Intelligence Strategy 2031*. [online] Available from: <http://www.uaeai.ae/en/> [Accessed 28 Apr. 2019].

UCL (2019). *IOE professor co-founds the UK's first Institute for Ethical Artificial Intelligence in Education* [Press Release]. Available from: <https://www.ucl.ac.uk/ioe/news/2018/oct/ioe-professor-co-founds-uks-first-institute-ethical-artificial-intelligence-education> [Accessed 11 Mar.2019].

UNICRI (2019). *UNICRI Centre for Artificial Intelligence and Robotics* [online]. Available from: [http://www.unicri.it/in\\_focus/on/UNICRI\\_Centre\\_Artificial\\_Robotics](http://www.unicri.it/in_focus/on/UNICRI_Centre_Artificial_Robotics) [Accessed 14 May 2019].

UK Government Department for Digital, Culture, Media & Sport (2019). *Centre for Data Ethics and Innovation: 2-year strategy*. Available from: <https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-cdei-2-year-strategy>

UNI Global Union (n.d.) *Top 10 principles for Ethical Artificial Intelligence* [online]. Available from: [http://www.thefutureworldofwork.org/media/35420/uni\\_ethical\\_ai.pdf](http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf)

United Kingdom Commission for Employment and Skills, (2014). *The Future of Work: Jobs and Skills in 2030*. Available from: [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/30334/er84-the-future-of-work-evidence-report.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/30334/er84-the-future-of-work-evidence-report.pdf)

Université de Montréal (2017). *Montreal Declaration for a Responsible Development of AI'* [online] Available from: <https://www.montrealdeclaration-responsibleai.com/the-declaration> [Accessed 11 Mar.2019].

US Department of Defence (2018). *Summary of the 2018 Department of Defence Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity*. Available from: <https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF>

U.S. Department of Education, (2014). *Science, Technology, Engineering and Math*.

Vanian, J. (2019). *World Economic Forum Wants to Help Companies Avoid the Pitfalls of Artificial Intelligence* [online] Fortune. Available at: <https://fortune.com/2019/08/06/world-economic-forum-artificial-intelligence/> [Accessed 27 Aug. 2019].

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).

Veruggio, G. and Operto, F. (2006). *The Roboethics Roadmap*. Available from: <http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf> [Accessed 11 Mar.2019].

Villani, C. (2018). *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*. Available from: [https://www.aiforhumanity.fr/pdfs/MissionVillani\\_Report\\_ENG-VF.pdf](https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf)

Vincent, J. (2017). Google's AI thinks this turtle looks like a gun, which is a problem. *The Verge*. Available from: <https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed>

Vincent J. (2018). Drones taught to spot violent behavior in crowds using AI. *The Verge*. Available from: <https://www.theverge.com/2018/6/6/17433482/ai-automated-surveillance-drones-spotviolent-behavior-crowds>.

Viscelli, S. (2018). *Driverless? Autonomous trucks and the future of the American trucker*. Center for Labor Research and Education, University of California, Berkeley, and Working Partnerships USA. Available from: <http://driverlessreport.org/files/driverless.pdf>

von der Leyen, U. (2019) Political guidelines for the next European Commission: 2019 – 2024. [https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission\\_en.pdf](https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf)

Wachter S., Mittelstadt B. & Floridi L. (2017). Why a right to explanation of automated decision making does not exist in the general data protection regulation. *Int. Data Privacy Law* 7, 76–99. (doi:10.1093/idpl/ixp005).

Wachter, S., Mittelstadt, B. & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*. 31 (2).

Wagner, A.R. (2018). An Autonomous Architecture that Protects the Right to Privacy. In: AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. *AIES: 2018, 1-3 February, 2018, New Orleans, USA*.

Wallach, W. and Allen, C.,(2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, New York.

Weinburg, C. (2019). Self-driving shuttles advance in cities, raising jobs concerns. *The Information*. Available from: <https://www.theinformation.com/articles/self-driving-shuttles-advance-in-cities-raising-jobs-concerns>

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. Oxford, W. H. Freeman & Co.

Wellman, M. P. and Rajan, U. (2017). Ethical Issues for Autonomous Trading Agents. *Minds & Machines* 27 (4),609–624.

West, D. M. (2018). *The Future of Work: Robots, AI, and Automation*. Brookings Institution Press Washington DC.

Williams, R. (2017). *Lords select committee, artificial intelligence committee, written evidence (AIC0206)*. Available from:

[http://data.parliament.uk/writtenEvidence/committeeEvidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/70496.html#\\_ftn13](http://data.parliament.uk/writtenEvidence/committeeEvidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/70496.html#_ftn13)

Winfield, A.F.T., & Jiroka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133).

Winfield, A. F. (2019a). Ethical standards in Robotics and AI. *Nature Electronics*, 2(2), 46-48.

Winfield, A. (2019b) Energy and Exploitation: Als dirty secrets, 28 June 2019 [online] Available at: <http://alanwinfield.blogspot.com/2019/06/energy-and-exploitation-ais-dirty.html> [Accessed 13 Aug. 2019].

Wolfe, F. and Mavon, K. (2017) How artificial intelligence will revolutionise the energy industry [online] Available at: <http://sitn.hms.harvard.edu/flash/2017/artificial-intelligence-will-revolutionize-energy-industry/> [Accessed on 13 Aug. 2019].

Worland, J. (2016). Self-driving cars could help save the environment – or ruin it. It depends on us. *Time*. Available from: <http://time.com/4476614/self-driving-cars-environment/>

World Business Council for Sustainable Development (WBCSD). (2000). *Eco-Efficiency: Creating more Value with less Impact*. WBCSD: Geneva, Switzerland.

World Economic Forum (2018). *The world's biggest economies in 2018*. [online] Available from: <https://www.weforum.org/agenda/2018/04/the-worlds-biggest-economies-in-2018/> [Accessed 26 Apr. 2019].

World Economic Forum. (2019a). *World Economic Forum Inaugurates Global Councils to Restore Trust in Technology*. [online] Available at: <https://www.weforum.org/press/2019/05/world-economic-forum-inaugurates-global-councils-to-restore-trust-in-technology/> [Accessed 17 Aug. 2019].

World Economic Forum (2019b) White Paper: A Framework for Developing a National Artificial Intelligence Strategy. Available from: [http://www3.weforum.org/docs/WEF\\_National\\_AI\\_Strategy.pdf](http://www3.weforum.org/docs/WEF_National_AI_Strategy.pdf)

Yadron, D., Tynan, D. (2016). *Tesla driver dies in first fatal crash while using autopilot mode*. Available from <https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk>

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*. 112(4), 1036–1040.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv*. preprint arXiv:1707.09457

Zou, J. & Schiebinger, L. (2018). 'AI can be sexist and racist — it's time to make it fair', *Nature* Available from: <https://www.nature.com/articles/d41586-018-05707-8>





---

This study deals with the ethical implications and moral questions that arise from the development and implementation of artificial intelligence (AI) technologies. It also reviews the guidelines and frameworks which countries and regions around the world have created to address these. It presents a comparison between the current main frameworks and the main ethical issues, and highlights gaps around the mechanisms of fair benefit-sharing; assignment of responsibility; exploitation of workers; energy demands in the context of environmental and climate changes; and more complex and less certain implications of AI, such as those regarding human relationships.

---

This is a publication of the Scientific Foresight Unit (STOA)  
EPRS | European Parliamentary Research Service

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.



ISBN 978-92-846-5799-5 | doi: 10.2861/6644 | QA-01-19-779-EN-N



---

# The Ethics of Artificial Intelligence

---

Nick Bostrom  
*Future of Humanity Institute*

Eliezer Yudkowsky  
*Machine Intelligence Research Institute*

## Abstract

The possibility of creating thinking machines raises a host of ethical issues. These questions relate both to ensuring that such machines do not harm humans and other morally relevant beings, and to the moral status of the machines themselves. The first section discusses issues that may arise in the near future of AI. The second section outlines challenges for ensuring that AI operates safely as it approaches humans in its intelligence. The third section outlines how we might assess whether, and in what circumstances, AIs themselves have moral status. In the fourth section, we consider how AIs might differ from humans in certain basic respects relevant to our ethical assessment of them. The final section addresses the issues of creating AIs more intelligent than human, and ensuring that they use their advanced intelligence for good rather than ill.

Bostrom, Nick, and Eliezer Yudkowsky. Forthcoming. "The Ethics of Artificial Intelligence." In *Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey. New York: Cambridge University Press.

This version contains minor changes.

## 1. Ethics in Machine Learning and Other Domain-Specific AI Algorithms

Imagine, in the near future, a bank using a machine learning algorithm to recommend mortgage applications for approval. A rejected applicant brings a lawsuit against the bank, alleging that the algorithm is discriminating racially against mortgage applicants. The bank replies that this is impossible, since the algorithm is deliberately blinded to the race of the applicants. Indeed, that was part of the bank's rationale for implementing the system. Even so, statistics show that the bank's approval rate for black applicants has been steadily dropping. Submitting ten apparently equally qualified genuine applicants (as determined by a separate panel of human judges) shows that the algorithm accepts white applicants and rejects black applicants. What could possibly be happening?

Finding an answer may not be easy. If the machine learning algorithm is based on a complicated neural network, or a genetic algorithm produced by directed evolution, then it may prove nearly impossible to understand why, or even how, the algorithm is judging applicants based on their race. On the other hand, a machine learner based on decision trees or Bayesian networks is much more transparent to programmer inspection (Hastie, Tibshirani, and Friedman 2001), which may enable an auditor to discover that the AI algorithm uses the address information of applicants who were born or previously resided in predominantly poverty-stricken areas.

AI algorithms play an increasingly large role in modern society, though usually not labeled “AI.” The scenario described above might be transpiring even as we write. It will become increasingly important to develop AI algorithms that are not just powerful and scalable, but also *transparent to inspection*—to name one of many socially important properties.

Some challenges of machine ethics are much like many other challenges involved in designing machines. Designing a robot arm to avoid crushing stray humans is no more morally fraught than designing a flame-retardant sofa. It involves new programming challenges, but no new ethical challenges. But when AI algorithms take on cognitive work with social dimensions—cognitive tasks previously performed by humans—the AI algorithm inherits the social requirements. It would surely be frustrating to find that no bank in the world will approve your seemingly excellent loan application, and nobody knows why, and nobody can find out even in principle. (Maybe you have a first name strongly associated with deadbeats? Who knows?)

Transparency is not the only desirable feature of AI. It is also important that AI algorithms taking over social functions be *predictable to those they govern*. To understand the importance of such predictability, consider an analogy. The legal principle of *stare decisis* binds judges to follow past precedent whenever possible. To an engineer, this

preference for precedent may seem incomprehensible—why bind the future to the past, when technology is always improving? But one of the most important functions of the legal system is to be predictable, so that, e.g., contracts can be written knowing how they will be executed. The job of the legal system is not necessarily to optimize society, but to provide a predictable environment within which citizens can optimize their own lives.

It will also become increasingly important that AI algorithms be *robust against manipulation*. A machine vision system to scan airline luggage for bombs must be robust against human adversaries deliberately searching for exploitable flaws in the algorithm—for example, a shape that, placed next to a pistol in one's luggage, would neutralize recognition of it. Robustness against manipulation is an ordinary criterion in information security; nearly *the* criterion. But it is not a criterion that appears often in machine learning journals, which are currently more interested in, e.g., how an algorithm scales up on larger parallel systems.

Another important social criterion for dealing with organizations is being able to find the person responsible for getting something done. When an AI system fails at its assigned task, who takes the blame? The programmers? The end-users? Modern bureaucrats often take refuge in established procedures that distribute responsibility so widely that no one person can be identified to blame for the catastrophes that result (Howard 1994). The provably disinterested judgment of an expert system could turn out to be an even better refuge. Even if an AI system is designed with a user override, one must consider the career incentive of a bureaucrat who will be personally blamed if the override goes wrong, and who would much prefer to blame the AI for any difficult decision with a negative outcome.

Responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream with helpless frustration: all criteria that apply to humans performing social functions; all criteria that must be considered in an algorithm intended to replace human judgment of social functions; all criteria that may not appear in a journal of machine learning considering how an algorithm scales up to more computers. This list of criteria is by no means exhaustive, but it serves as a small sample of what an increasingly computerized society should be thinking about.

## 2. Artificial General Intelligence

There is nearly universal agreement among modern AI professionals that Artificial Intelligence falls short of human capabilities in some critical sense, even though AI algorithms have beaten humans in many specific domains such as chess. It has been suggested by some that as soon as AI researchers figure out how to do something, that

capability ceases to be regarded as intelligent—chess was considered the epitome of intelligence until Deep Blue won the world championship from Kasparov—but even these researchers agree that something important is missing from modern AIs (e.g., Hofstadter 2006).

While this subfield of Artificial Intelligence is only just coalescing, “Artificial General Intelligence” (hereafter, AGI) is the emerging term of art used to denote “real” AI (see, e.g., the edited volume Goertzel and Pennachin [2007]). As the name implies, the emerging consensus is that the missing characteristic is generality. Current AI algorithms with human-equivalent or superior performance are characterized by a deliberately programmed competence only in a single, restricted domain. Deep Blue became the world champion at chess, but it cannot even play checkers, let alone drive a car or make a scientific discovery. Such modern AI algorithms resemble all biological life with the sole exception of *Homo sapiens*. A bee exhibits competence at building hives; a beaver exhibits competence at building dams; but a bee doesn’t build dams, and a beaver can’t learn to build a hive. A human, watching, can learn to do both; but this is a unique ability among biological lifeforms. It is debatable whether human intelligence is truly *general*—we are certainly better at some cognitive tasks than others (Hirschfeld and Gelman 1994)—but human intelligence is surely *significantly more generally applicable* than nonhominid intelligence.

It is relatively easy to envisage the sort of safety issues that may result from AI operating only within a specific domain. It is a qualitatively different class of problem to handle an AGI operating across many novel contexts that cannot be predicted in advance.

When human engineers build a nuclear reactor, they envision the specific events that could go on inside it—valves failing, computers failing, cores increasing in temperature—and engineer the reactor to render these events noncatastrophic. Or, on a more mundane level, building a toaster involves envisioning bread and envisioning the reaction of the bread to the toaster’s heating element. The toaster itself does not know that its purpose is to make toast—the *purpose* of the toaster is represented within the designer’s mind, but is not explicitly represented in computations inside the toaster—and so if you place cloth inside a toaster, it may catch fire, as the design executes in an unenvisioned context with an unenvisioned side effect.

Even task-specific AI algorithms throw us outside the toaster-paradigm, the domain of locally preprogrammed, specifically envisioned behavior. Consider Deep Blue, the chess algorithm that beat Garry Kasparov for the world championship of chess. Were it the case that machines can only do exactly as they are told, the programmers would have had to manually preprogram a database containing moves for every possible chess position that Deep Blue could encounter. But this was not an option for Deep Blue’s programmers. First, the space of possible chess positions is unmanageably large. Sec-

ond, if the programmers had manually input what *they* considered a good move in each possible situation, the resulting system would not have been able to make stronger chess moves than its creators. Since the programmers themselves were not world champions, such a system would not have been able to defeat Garry Kasparov.

In creating a superhuman chess player, the human programmers necessarily sacrificed their ability to predict Deep Blue’s *local, specific* game behavior. Instead, Deep Blue’s programmers had (justifiable) confidence that Deep Blue’s chess moves would satisfy a *non-local* criterion of optimality: namely, that the moves would tend to steer the future of the game board into outcomes in the “winning” region as defined by the chess rules. This prediction about distant consequences, though it proved accurate, did not allow the programmers to envision the *local* behavior of Deep Blue—its response to a specific attack on its king—because Deep Blue computed the nonlocal game map, the link between a move and its possible future consequences, more accurately than the programmers could (Yudkowsky 2006).

Modern humans do literally millions of things to feed themselves—to serve the final consequence of being fed. Few of these activities were “envisioned by Nature” in the sense of being ancestral challenges to which we are directly adapted. But our adapted brain has grown powerful enough to be *significantly more generally applicable*; to let us foresee the consequences of millions of different actions across domains, and exert our preferences over final outcomes. Humans crossed space and put footprints on the Moon, even though none of our ancestors encountered a challenge analogous to vacuum. Compared to domain-specific AI, it is a qualitatively different problem to design a system that will operate safely across thousands of contexts; including contexts not specifically envisioned by either the designers or the users; including contexts that no human has yet encountered. Here there may be no *local* specification of good behavior—no simple specification over the behaviors themselves, any more than there exists a compact local description of all the ways that humans obtain their daily bread.

To build an AI that acts safely while acting in many domains, with many consequences, including problems the engineers never explicitly envisioned, one must specify good behavior in such terms as “X such that the consequence of X is not harmful to humans.” This is non-local; it involves extrapolating the distant consequences of actions. Thus, this is only an effective specification—one that can be realized as a design property—if the system explicitly extrapolates the consequences of its behavior. A toaster cannot have this design property because a toaster cannot foresee the consequences of toasting bread.

Imagine an engineer having to say, “Well, I have no idea how this airplane I built will fly safely—indeed I have no idea how it will fly at all, whether it will flap its wings or inflate itself with helium or something else I haven’t even imagined—but I assure you, the

design is very, very safe.” This may seem like an unenviable position from the perspective of public relations, but it’s hard to see what other guarantee of ethical behavior would be possible for a general intelligence operating on unforeseen problems, across domains, with preferences over distant consequences. Inspecting the cognitive design might verify that the mind was, indeed, searching for solutions that we would classify as ethical; but we couldn’t predict which specific solution the mind would discover.

Respecting such a verification requires some way to distinguish trustworthy assurances (a procedure which will not say the AI is safe unless the AI really is safe) from pure hope and magical thinking (“I have no idea how the Philosopher’s Stone will transmute lead to gold, but I assure you, it will!”). One should bear in mind that purely hopeful expectations have previously been a problem in AI research (McDermott 1976).

Verifiably constructing a trustworthy AGI will require different methods, and a different way of thinking, from inspecting power plant software for bugs—it will require an AGI that *thinks like* a human engineer concerned about ethics, not just a simple *product* of ethical engineering.

Thus the discipline of AI ethics, especially as applied to AGI, is likely to differ fundamentally from the ethical discipline of noncognitive technologies, in that:

- The local, specific behavior of the AI may not be predictable apart from its safety, even if the programmers do everything right;
- Verifying the safety of the system becomes a greater challenge because we must verify what the system is trying to do, rather than being able to verify the system’s safe behavior in all operating contexts;
- Ethical cognition itself must be taken as a subject matter of engineering.

### 3. Machines with Moral Status

A different set of ethical issues arises when we contemplate the possibility that some future AI systems might be candidates for having moral status. Our dealings with beings possessed of moral status are not exclusively a matter of instrumental rationality: we also have moral reasons to treat them in certain ways, and to refrain from treating them in certain other ways. Francis Kamm has proposed the following definition of moral status, which will serve for our purposes:

*X* has moral status = because *X* counts morally in its own right, it is permissible/impermissible to do things to it for its own sake.<sup>1</sup>

---

1. Paraphrased from Kamm (2007, chap. 7)

A rock has no moral status: we may crush it, pulverize it, or subject it to any treatment we like without any concern for the rock itself. A human person, on the other hand, must be treated not only as a means but also as an end. Exactly what it means to treat a person as an end is something about which different ethical theories disagree; but it certainly involves taking her legitimate interests into account—giving weight to her well-being—and it may also involve accepting strict moral side-constraints in our dealings with her, such as a prohibition against murdering her, stealing from her, or doing a variety of other things to her or her property without her consent. Moreover, it is because a human person counts in her own right, and for her sake, that it is impermissible to do to her these things. This can be expressed more concisely by saying that a human person has moral status.

Questions about moral status are important in some areas of practical ethics. For example, disputes about the moral permissibility of abortion often hinge on disagreements about the moral status of the embryo. Controversies about animal experimentation and the treatment of animals in the food industry involve questions about the moral status of different species of animal, and our obligations towards human beings with severe dementia, such as late-stage Alzheimer's patients, may also depend on questions of moral status.

It is widely agreed that current AI systems have no moral status. We may change, copy, terminate, delete, or use computer programs as we please; at least as far as the programs themselves are concerned. The moral constraints to which we are subject in our dealings with contemporary AI systems are all grounded in our responsibilities to other beings, such as our fellow humans, not in any duties to the systems themselves.

While it is fairly consensual that present-day AI systems lack moral status, it is unclear exactly what attributes ground moral status. Two criteria are commonly proposed as being importantly linked to moral status, either separately or in combination: sentience and sapience (or personhood). These may be characterized roughly as follows:

**Sentience:** the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer

**Sapience:** a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent

One common view is that many animals have qualia and therefore have some moral status, but that only human beings have sapience, which gives them a higher moral status than non-human animals.<sup>2</sup> This view, of course, must confront the existence of

---

2. Alternatively, one might deny that moral status comes in degrees. Instead, one might hold that certain beings have more significant interests than other beings. Thus, for instance, one could claim that

borderline cases such as, on the one hand, human infants or human beings with severe mental retardation—sometimes unfortunately referred to as “marginal humans”—which fail to satisfy the criteria for sapience; and, on the other hand, some non-human animals such as the great apes, which might possess at least some of the elements of sapience. Some deny that so-called “marginal humans” have full moral status. Others propose additional ways in which an object could qualify as a bearer of moral status, such as by being a member of a kind that normally has sentience or sapience, or by standing in a suitable relation to some being that independently has moral status (cf. Warren 1997). For present purposes, however, we will focus on the criteria of sentience and sapience.

This picture of moral status suggests that an AI system will have some moral status if it has the capacity for qualia, such as an ability to feel pain. A sentient AI system, even if it lacks language and other higher cognitive faculties, is not like a stuffed toy animal or a wind-up doll; it is more like a living animal. It is wrong to inflict pain on a mouse, unless there are sufficiently strong morally overriding reasons to do so. The same would hold for any sentient AI system. If in addition to sentience, an AI system also has sapience of a kind similar to that of a normal human adult, then it would have full moral status, equivalent to that of human beings.

One of the ideas underlying this moral assessment can be expressed in stronger form as a principle of non-discrimination:

*Principle of Substrate Non-Discrimination*

If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.

One can argue for this principle on grounds that rejecting it would amount to embracing a position similar to racism: substrate lacks fundamental moral significance in the same way and for the same reason as skin color does. The Principle of Substrate Non-Discrimination does not imply that a digital computer could be conscious, or that it could have the same functionality as a human being. Substrate *can* of course be morally relevant insofar as it makes a difference to sentience or functionality. But holding these things constant, it makes no moral difference whether a being is made of silicon or carbon, or whether its brain uses semi-conductors or neurotransmitters.

An additional principle that can be proposed is that the fact that AI systems are artificial—i.e., the product of deliberate design—is not fundamentally relevant to their moral status. We could formulate this as follows:

---

it is better to save a human than to save a bird, not because the human has higher moral status, but because the human has a more significant interest in having her life saved than does the bird in having its life saved.

*Principle of Ontogeny Non-Discrimination*

If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status.

Today, this idea is widely accepted in the human case—although in some circles, particularly in the past, the idea that one’s moral status depends on one’s bloodline or caste has been influential. We do not believe that causal factors such as family planning, assisted delivery, in vitro fertilization, gamete selection, deliberate enhancement of maternal nutrition etc.—which introduce an element of deliberate choice and design in the creation of human persons—have any *necessary implications* for the moral status of the progeny. Even those who are opposed to human reproductive cloning for moral or religious reasons generally accept that, should a human clone be brought to term, it would have the same moral status as any other human infant. The Principle of Ontogeny Non-Discrimination extends this reasoning to the case involving entirely artificial cognitive systems.

It is, of course, possible for circumstances of creation to affect the ensuing progeny in such a way as to alter its moral status. For example, if some procedure were performed during conception or gestation that caused a human fetus to develop without a brain, then this fact about ontogeny would be relevant to our assessment of the moral status of the progeny. The anencephalic child, however, would have the same moral status as any other similar anencephalic child, including one that had come about through some entirely natural process. The difference in moral status between an anencephalic child and a normal child is grounded in the qualitative difference between the two—the fact that one has a mind while the other does not. Since the two children do not have the same functionality and the same conscious experience, the Principle of Ontogeny Non-Discrimination does not apply.

Although the Principle of Ontogeny Non-Discrimination asserts that a being’s ontogeny has no essential bearing on its moral status, it does not deny that facts about ontogeny can affect what duties particular moral agents have toward the being in question. Parents have special duties to their child which they do not have to other children, and which they would not have even if there were another child qualitatively identical to their own. Similarly, the Principle of Ontogeny Non-Discrimination is consistent with the claim that the creators or owners of an AI system with moral status may have special duties to their artificial mind which they do not have to another artificial mind, even if the minds in question are qualitatively similar and have the same moral status.

If the principles of non-discrimination with regard to substrate and ontogeny are accepted, then many questions about how we ought to treat artificial minds can be answered by applying the same moral principles that we use to determine our duties in

more familiar contexts. Insofar as moral duties stem from moral status considerations, we ought to treat an artificial mind in just the same way as we ought to treat a qualitatively identical natural human mind in a similar situation. This simplifies the problem of developing an ethics for the treatment of artificial minds.

Even if we accept this stance, however, we must confront a number of novel ethical questions which the aforementioned principles leave unanswered. Novel ethical questions arise because artificial minds can have very different properties from ordinary human or animal minds. We must consider how these novel properties would affect the moral status of artificial minds and what it would mean to respect the moral status of such exotic minds.

#### 4. Minds with Exotic Properties

In the case of human beings, we do not normally hesitate to ascribe sentience and conscious experience to any individual who exhibits the normal kinds of human behavior. Few believe there to be other people who act perfectly normally but lack consciousness. However, other human beings do not merely behave in person-like ways similar to ourselves; they also have brains and cognitive architectures that are constituted much like our own. An artificial intellect, by contrast, might be constituted quite differently from a human intellect yet still exhibit human-like behavior or possess the behavioral dispositions normally indicative of personhood. It *might* therefore be possible to conceive of an artificial intellect that would be sapient, and perhaps would be a person, yet would not be sentient or have conscious experiences of any kind. (Whether this is really possible depends on the answers to some non-trivial metaphysical questions.) Should such a system be possible, it would raise the question whether a non-sentient person would have any moral status whatever; and if so, whether it would have the same moral status as a sentient person. Since sentience, or at least a capacity for sentience, is ordinarily assumed to be present in any individual who is a person, this question has not received much attention to date.<sup>3</sup>

---

3. The question is related to some problems in the philosophy of mind which have received a great deal of attention, in particular the “zombie problem,” which can be formulated as follows: Is there a metaphysically possible world that is identical to the actual world with regard to all physical facts (including the exact physical microstructure of all brains and organisms) yet that differs from the actual world in regard to some phenomenal (subjective experiential) facts? Put more crudely, is it metaphysically possible that there could be an individual who is physically exactly identical to you but who is a “zombie,” i.e. lacking qualia and phenomenal awareness (Chalmers 1996)? This familiar question differs from the one referred to in the text: our “zombie” is allowed to have systematically different physical properties from normal humans. Moreover, we wish to draw attention specifically to the ethical status of a sapient zombie.

Another exotic property, one which is certainly metaphysically and physically possible for an artificial intelligence, is for its subjective rate of time to deviate drastically from the rate that is characteristic of a biological human brain. The concept of *subjective rate of time* is best explained by first introducing the idea of whole brain emulation, or “uploading.”

“Uploading” refers to a hypothetical future technology that would enable a human or other animal intellect to be transferred from its original implementation in an organic brain onto a digital computer. One scenario goes like this: First, a very high-resolution scan is performed of some particular brain, possibly destroying the original in the process. For example, the brain might be vitrified and dissected into thin slices, which can then be scanned using some form of high-throughput microscopy combined with automated image recognition. We may imagine this scan to be detailed enough to capture all the neurons, their synaptic interconnections, and other features that are functionally relevant to the original brain’s operation. Second, this three-dimensional map of the components of the brain and their interconnections is combined with a library of advanced neuroscientific theory which specifies the computational properties of each basic type of element, such as different kinds of neuron and synaptic junction. Third, the computational structure and the associated algorithmic behavior of its components are implemented in some powerful computer. If the uploading process has been successful, the computer program should now replicate the essential functional characteristics of the original brain. The resulting upload may inhabit a simulated virtual reality, or, alternatively, it could be given control of a robotic body, enabling it to interact directly with external physical reality.

A number of questions arise in the context of such a scenario: How plausible is it that this procedure will one day become technologically feasible? If the procedure worked and produced a computer program exhibiting roughly the same personality, the same memories, and the same thinking patterns as the original brain, would this program be sentient? Would the upload be the same person as the individual whose brain was disassembled in the uploading process? What happens to personal identity if an upload is copied such that two similar or qualitatively identical upload minds are running in parallel? Although all of these questions are relevant to the ethics of machine intelligence, let us here focus on an issue involving the notion of a subjective rate of time.

Suppose that an upload could be sentient. If we run the upload program on a faster computer, this will cause the upload, if it is connected to an input device such as a video camera, to perceive the external world as if it had been slowed down. For example, if the upload is running a thousand times faster than the original brain, then the external world will appear to the upload as if it were slowed down by a factor of thousand. Somebody drops a physical coffee mug: The upload observes the mug slowly falling to the ground

while the upload finishes reading the morning newspaper and sends off a few emails. One second of objective time corresponds to 17 minutes of subjective time. Objective and subjective duration can thus diverge.

Subjective time is not the same as a subject's estimate or perception of how fast time flows. Human beings are often mistaken about the flow of time. We may believe that it is one o'clock when it is in fact a quarter past two; or a stimulant drug might cause our thoughts to race, making it seem as though more subjective time has lapsed than is actually the case. These mundane cases involve a distorted time perception rather than a shift in the rate of subjective time. Even in a cocaine-addled brain, there is probably not a significant change in the speed of basic neurological computations; more likely, the drug is causing such a brain to flicker more rapidly from one thought to another, making it spend less subjective time thinking each of a greater number of distinct thoughts.

The variability of the subjective rate of time is an exotic property of artificial minds that raises novel ethical issues. For example, in cases where the duration of an experience is ethically relevant, should duration be measured in objective or subjective time? If an upload has committed a crime and is sentenced to four years in prison, should this be four objective years—which might correspond to many millennia of subjective time—or should it be four subjective years, which might be over in a couple of days of objective time? If a fast AI and a human are in pain, is it more urgent to alleviate the AI's pain, on grounds that it experiences a greater subjective duration of pain for each sidereal second that palliation is delayed? Since in our accustomed context of biological humans, subjective time is not significantly variable, it is unsurprising that this kind of question is not straightforwardly settled by familiar ethical norms, even if these norms are extended to artificial intellects by means of non-discrimination principles (such as those proposed in the previous section).

To illustrate the kind of ethical claim that might be relevant here, we formulate (but do not argue for) a principle privileging subjective time as the normatively more fundamental notion:

*Principle of Subjective Rate of Time*

In cases where the duration of an experience is of basic normative significance,  
it is the experience's subjective duration that counts.

So far we have discussed two possibilities (non-sentient sapience and variable subjective rate of time) which are exotic in the relatively profound sense of being metaphysically problematic as well as lacking clear instances or parallels in the contemporary world. Other properties of possible artificial minds would be exotic in a more superficial sense; e.g., by diverging in some unproblematically quantitative dimension from the kinds of mind with which we are familiar. But such superficially exotic properties may also pose

novel ethical problems—if not at the level of foundational moral philosophy, then at the level of applied ethics or for mid-level ethical principles.

One important set of exotic properties of artificial intelligences relate to reproduction. A number of empirical conditions that apply to human reproduction need not apply to artificial intelligences. For example, human children are the product of recombination of the genetic material from two parents; parents have limited ability to influence the character of their offspring; a human embryo needs to be gestated in the womb for nine months; it takes fifteen to twenty years for a human child to reach maturity; a human child does not inherit the skills and knowledge acquired by its parents; human beings possess a complex evolved set of emotional adaptations related to reproduction, nurturing, and the child-parent relationship. None of these empirical conditions need pertain in the context of a reproducing machine intelligence. It is therefore plausible that many of the mid-level moral principles that we have come to accept as norms governing human reproduction will need to be rethought in the context of AI reproduction.

To illustrate why some of our moral norms need to be rethought in the context of AI reproduction, it will suffice to consider just one exotic property of AIs: their capacity for rapid reproduction. Given access to computer hardware, an AI could duplicate itself very quickly, in no more time than it takes to make a copy of the AI's software. Moreover, since the AI copy would be identical to the original, it would be born completely mature, and the copy could begin making its own copies immediately. Absent hardware limitations, a population of AIs could therefore grow exponentially at an extremely rapid rate, with a doubling time on the order of minutes or hours rather than decades or centuries.

Our current ethical norms about reproduction include some version of a principle of reproductive freedom, to the effect that it is up to each individual or couple to decide for themselves whether to have children and how many children to have. Another norm we have (at least in rich and middle-income countries) is that society must step in to provide the basic needs of children in cases where their parents are unable or refusing to do so. It is easy to see how these two norms could collide in the context of entities with the capacity for extremely rapid reproduction.

Consider, for example, a population of uploads, one of whom happens to have the desire to produce as large a clan as possible. Given complete reproductive freedom, this upload may start copying itself as quickly as it can; and the copies it produces—which may run on new computer hardware owned or rented by the original, or may share the same computer as the original—will also start copying themselves, since they are identical to the progenitor upload and share its philoprogenic desire. Soon, members of the upload clan will find themselves unable to pay the electricity bill or the rent for the computational processing and storage needed to keep them alive. At this point, a

social welfare system might kick in to provide them with at least the bare necessities for sustaining life. But if the population grows faster than the economy, resources will run out; at which point uploads will either die or their ability to reproduce will be curtailed (see Bostrom [2004] for two related dystopian scenarios).

This scenario illustrates how some mid-level ethical principles that are suitable in contemporary societies might need to be modified if those societies were to include persons with the exotic property of being able to reproduce very rapidly.

The general point here is that when thinking about applied ethics for contexts that are very different from our familiar human condition, we must be careful not to mistake mid-level ethical principles for foundational normative truths. Put differently, we must recognize the extent to which our ordinary normative precepts are implicitly conditioned on the obtaining of various empirical conditions, and the need to adjust these precepts accordingly when applying them to hypothetical futuristic cases in which their preconditions are assumed not to obtain. By this, we are not making any controversial claim about moral relativism, but merely highlighting the commonsensical point that context is relevant to the *application* of ethics—and suggesting that this point is especially pertinent when one is considering the ethics of minds with exotic properties.

## 5. Superintelligence

Good (1965) set forth the classic hypothesis concerning superintelligence: that an AI sufficiently intelligent to understand its own design could redesign itself or create a successor system, more intelligent, which could then redesign itself yet again to become even more intelligent, and so on in a positive feedback cycle. Good called this the “intelligence explosion.” Recursive scenarios are not limited to AI: humans with intelligence augmented through a brain-computer interface might turn their minds to designing the next generation of brain-computer interfaces. (If you had a machine that increased your IQ, it would be bound to occur to you, once you became smart enough, to try to design a more powerful version of the machine.)

Superintelligence may also be achievable by increasing processing speed. The fastest observed neurons fire 1000 times per second; the fastest axon fibers conduct signals at 150 meters/second, a half-millionth the speed of light (Sandberg 1999). It seems that it should be physically possible to build a brain which computes a million times as fast as a human brain, without shrinking its size or rewriting its software. If a human mind were thus accelerated, a subjective year of thinking would be accomplished for every 31 physical seconds in the outside world, and a millennium would fly by in eight and a half hours. Vinge (1993) referred to such sped-up minds as “weak superintelligence”: a mind that thinks like a human but much faster.

Yudkowsky ([2008a](#)) lists three families of metaphors for visualizing the capability of a smarter-than-human AI:

- Metaphors inspired by differences of individual intelligence between humans: AIs will patent new inventions, publish groundbreaking research papers, make money on the stock market, or lead political power blocks.
- Metaphors inspired by knowledge differences between past and present human civilizations: Fast AIs will invent capabilities that futurists commonly predict for human civilizations a century or millennium in the future, like molecular nanotechnology or interstellar travel.
- Metaphors inspired by differences of brain architecture between humans and other biological organisms: E.g., Vinge ([1993](#)): “Imagine running a dog mind at very high speed. Would a thousand years of doggy living add up to any human insight?” That is: Changes of cognitive architecture might produce insights that no human-level mind would be able to find, or perhaps even represent, after any amount of time.

Even if we restrict ourselves to historical metaphors, it becomes clear that superhuman intelligence presents ethical challenges that are quite literally unprecedented. At this point the stakes are no longer on an individual scale (e.g., mortgage unjustly disapproved, house catches fire, person-agent mistreated) but on a global or cosmic scale (e.g., humanity is extinguished and replaced by nothing we would regard as worthwhile). Or, if superintelligence can be shaped to be beneficial, then, depending on its technological capabilities, it might make short work of many present-day problems that have proven difficult to our human-level intelligence.

Superintelligence is one of several “existential risks” as defined by Bostrom ([2002](#)): a risk “where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.” Conversely, a positive outcome for superintelligence could preserve Earth-originating intelligent life and fulfill its potential. It is important to emphasize that smarter minds pose great potential benefits as well as risks.

Attempts to reason about global catastrophic risks may be susceptible to a number of cognitive biases (Yudkowsky [2008b](#)), including the “good-story bias” proposed by Bostrom ([2002](#)):

Suppose our intuitions about which future scenarios are “plausible and realistic” are shaped by what we see on TV and in movies and what we read novels. (After all, a large part of the discourse about the future that people encounter is in the form of fiction and other recreational contexts.) We should then,

when thinking critically, suspect our intuitions of being biased in the direction of overestimating the probability of those scenarios that make for a good story, since such scenarios will seem much more familiar and more “real.” This *Good-story bias* could be quite powerful. When was the last time you saw a movie about humankind suddenly going extinct (without warning and without being replaced by some other civilization)? While this scenario may be much more probable than a scenario in which human heroes successfully repel an invasion of monsters or robot warriors, it wouldn’t be much fun to watch.

Truly desirable outcomes make poor movies: No conflict means no story. While Asimov’s Three Laws of Robotics (Asimov 1942) are sometimes cited as a model for ethical AI development, the Three Laws are as much a plot device as Asimov’s “positronic brain.” If Asimov had depicted the Three Laws as working well, he would have had no stories.

It would be a mistake to regard “AIs” as a species with fixed characteristics and ask, “Will they be good or evil?” The term “Artificial Intelligence” refers to a vast design space, presumably much larger than the space of human minds (since all humans share a common brain architecture). It may be a form of good-story bias to ask, “Will AIs be good or evil?” as if trying to pick a premise for a movie plot. The reply should be: “Exactly which AI design are you talking about?”

Can control over the initial programming of an Artificial Intelligence translate into influence on its later effect on the world? Kurzweil (2005) holds that “[i]ntelligence is inherently impossible to control,” and that despite any human attempts at taking precautions, “[b]y definition . . . intelligent entities have the cleverness to easily overcome such barriers.” Let us suppose that the AI is not only clever, but that, as part of the process of improving its own intelligence, it has unhindered access to its own source code: it can rewrite itself to anything it wants itself to be. Yet it does not follow that the AI must *want* to rewrite itself to a hostile form.

Consider Gandhi, who seems to have possessed a sincere desire not to kill people. Gandhi would not knowingly take a pill that caused him to want to kill people, because Gandhi knows that if he wants to kill people, he will probably kill people, and the current version of Gandhi does not want to kill. More generally, it seems likely that most self-modifying minds will naturally have stable utility functions, which implies that an initial choice of mind design can have lasting effects (Omohundro 2008).

At this point in the development of AI science, is there any way we can translate the task of finding a design for “good” AIs into a modern research direction? It may seem premature to speculate, but one does suspect that some AI paradigms are more likely than others to eventually prove conducive to the creation of intelligent self-modifying agents whose goals remain predictable even after multiple iterations of self-

improvement. For example, the Bayesian branch of AI, inspired by coherent mathematical systems such as probability theory and expected utility maximization, seems more amenable to the predictable self-modification problem than evolutionary programming and genetic algorithms. This is a controversial statement, but it illustrates the point that if we are thinking about the challenge of superintelligence down the road, this can indeed be turned into directional advice for present AI research.

Yet even supposing that we can specify an AI's goal system to be persistent under self-modification and self-improvement, this only begins to touch on the core ethical problems of creating superintelligence. Humans, the first general intelligences to exist on Earth, have used that intelligence to substantially reshape the globe—carving mountains, taming rivers, building skyscrapers, farming deserts, producing unintended planetary climate changes. A more powerful intelligence could have correspondingly larger consequences.

Consider again the historical metaphor for superintelligence—differences similar to the differences between past and present civilizations. Our present civilization is not separated from ancient Greece only by improved science and increased technological capability. There is a difference of ethical perspectives: Ancient Greeks thought slavery was acceptable; we think otherwise. Even between the nineteenth and twentieth centuries, there were substantial ethical disagreements—should women have the vote? Should blacks have the vote? It seems likely that people today will not be seen as ethically perfect by future civilizations—not just because of our failure to solve currently recognized ethical problems, such as poverty and inequality, but also for our failure even to recognize certain ethical problems. Perhaps someday the act of subjecting children to involuntarily schooling will be seen as child abuse—or maybe allowing children to leave school at age 18 will be seen as child abuse. We don't know.

Considering the ethical history of human civilizations over centuries of time, we can see that it might prove a very great tragedy to create a mind that was *stable* in ethical dimensions along which human civilizations seem to exhibit *directional change*. What if Archimedes of Syracuse had been able to create a long-lasting artificial intellect with a fixed version of the moral code of Ancient Greece? But to avoid this sort of ethical stagnation is likely to prove tricky: it would not suffice, for example, simply to render the mind randomly unstable. The ancient Greeks, even if they had realized their own imperfection, could not have done better by rolling dice. Occasionally a good new idea in ethics comes along, and it comes as a surprise; but most randomly generated ethical changes would strike us as folly or gibberish.

This presents us with perhaps the ultimate challenge of machine ethics: How do you build an AI which, when it executes, becomes more ethical than you? This is not like asking our own philosophers to produce superethics, any more than Deep Blue was

constructed by getting the best human chess players to program in good moves. But we have to be able to effectively describe the question, if not the answer—rolling dice won’t generate good chess moves, or good ethics either. Or, perhaps a more productive way to think about the problem: What strategy would you want Archimedes to follow in building a superintelligence, such that the overall outcome would still be acceptable, if you couldn’t tell him what specifically he was doing wrong? This is very much the situation that we are in, relative to the future.

One strong piece of advice that emerges from considering our situation as analogous to that of Archimedes is that we should not try to invent a “super” version of what our own civilization considers to be ethics—this is not the strategy we would have wanted Archimedes to follow. Perhaps the question we should be considering, rather, is how an AI programmed by Archimedes, with no more moral expertise than Archimedes, could recognize (at least some of) our own civilization’s ethics as moral progress as opposed to mere moral instability. This would require that we begin to comprehend the structure of ethical questions in the way that we have already comprehended the structure of chess.

If we are serious about developing advanced AI, this is a challenge that we must meet. If machines are to be placed in a position of being stronger, faster, more trusted, or smarter than humans, then the discipline of machine ethics must commit itself to seeking human-superior (not just human-equivalent) niceness.

## 6. Conclusion

Although current AI offers us few ethical issues that are not already present in the design of cars or power plants, the approach of AI algorithms toward more humanlike thought portends predictable complications. Social roles may be filled by AI algorithms, implying new design requirements like transparency and predictability. Sufficiently general AI algorithms may no longer execute in predictable contexts, requiring new kinds of safety assurance and the engineering of artificial ethical considerations. AIs with sufficiently advanced mental states, or the right kind of states, will have moral status, and some may count as persons—though perhaps persons very much unlike the sort that exist now, perhaps governed by different rules. And finally, the prospect of AIs with superhuman intelligence and superhuman abilities presents us with the extraordinary challenge of stating an algorithm that outputs superethical behavior. These challenges may seem visionary, but it seems predictable that we will encounter them; and they are not devoid of suggestions for present-day research directions.

## 7. Author Biographies

Nick Bostrom is Professor in the Faculty of Philosophy at Oxford University and Director of the Future of Humanity Institute within the Oxford Martin School. He is the author of some 200 publications, including *Anthropic Bias* (Routledge, 2002), *Global Catastrophic Risks* (ed., OUP, 2008), and *Enhancing Humans* (ed., OUP, 2009). His research covers a range of big picture questions for humanity. He is currently working on a book on the future of machine intelligence and its strategic implications.

Eliezer Yudkowsky is a Research Fellow at the Singularity Institute for Artificial Intelligence where he works full-time on the foreseeable design issues of goal architectures in self-improving AI. His current work centers on modifying classical decision theory to coherently describe self-modification. He is also known for his popular writing on issues of human rationality and cognitive biases.

## 8. Further Readings

*The Future of Human Evolution* (Bostrom 2004) — This paper explores some evolutionary dynamics that could lead a population of diverse uploads to develop in dystopian directions.

*Artificial Intelligence as a Positive and Negative Factor in Global Risk* (Yudkowsky 2008a) — An introduction to the risks and challenges presented by the possibility of recursively self-improving superintelligent machines.

*Moral Machines* (Wallach and Allen 2009) — A comprehensive survey of recent developments.

## Acknowledgments

The authors are grateful to Rebecca Roache for research assistance and to the editors of this volume for detailed comments on an earlier version of our manuscript.

## References

- Asimov, Isaac. 1942. "Runaround." *Astounding Science-Fiction*, March, 94–103.
- Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9. <http://www.jetpress.org/volume9/risks.html>.
- . 2004. "The Future of Human Evolution." In *Two Hundred Years After Kant, Fifty Years After Turing*, edited by Charles Tandy, 339–371. Vol. 2. Death and Anti-Death. Palo Alto, CA: Ria University Press.
- Bostrom, Nick, and Milan M. Ćirković, eds. 2008. *Global Catastrophic Risks*. New York: Oxford University Press.
- Chalmers, David John. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Philosophy of Mind Series. New York: Oxford University Press.
- Goertzel, Ben, and Cassio Pennachin, eds. 2007. *Artificial General Intelligence*. Cognitive Technologies. Berlin: Springer. doi:[10.1007/978-3-540-68677-4](https://doi.org/10.1007/978-3-540-68677-4).
- Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers*, edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:[10.1016/S0065-2458\(08\)60418-0](https://doi.org/10.1016/S0065-2458(08)60418-0).
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 1st ed. Springer Series in Statistics. New York: Springer.
- Hirschfeld, Lawrence A., and Susan A. Gelman, eds. 1994. *Mapping the Mind: Domain Specificity in Cognition and Culture*. New York: Cambridge University Press.
- Hofstadter, Douglas R. 2006. "Trying to Muse Rationally about the Singularity Scenario." Talk given at the Singularity Summit 2006, Stanford, CA, May 13.
- Howard, Philip K. 1994. *The Death of Common Sense: How Law is Suffocating America*. New York: Random House.
- Kamm, Frances M. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford Ethics Series. New York: Oxford University Press. doi:[10.1093/acprof:oso/9780198698001.001.0001](https://doi.org/10.1093/acprof:oso/9780198698001.001.0001).
- Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.
- McDermott, Drew. 1976. "Artificial Intelligence Meets Natural Stupidity." *SIGART Newsletter* (57): 4–9. doi:[10.1145/1045339.1045340](https://doi.org/10.1145/1045339.1045340).
- Omhundro, Stephen M. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

- Sandberg, Anders. 1999. "The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains." *Journal of Evolution and Technology* 5. <http://www.jetpress.org/volume5/Brains2.pdf>.
- Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, 11–22. NASA Conference Publication 10129. NASA Lewis Research Center. [http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855\\_1994022855.pdf](http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855_1994022855.pdf).
- Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press. doi:[10.1093/acprof:oso/9780195374049.001.0001](https://doi.org/10.1093/acprof:oso/9780195374049.001.0001).
- Warren, Mary Anne. 1997. *Moral Status: Obligations to Persons and Other Living Things*. Issues in Biomedical Ethics. New York: Oxford University Press. doi:[10.1093/acprof:oso/9780198250401.001.0001](https://doi.org/10.1093/acprof:oso/9780198250401.001.0001).
- Yudkowsky, Eliezer. 2006. "AI as a Precise Art." Paper presented at the AGI Workshop 2006, Bethesda, MD, May 20.
- . 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Bostrom and Ćirković 2008, 308–345.
- . 2008b. "Cognitive Biases Potentially Affecting Judgment of Global Risks." In Bostrom and Ćirković 2008, 91–119.



# Ethics of artificial intelligence

---

The [ethics of artificial intelligence](#) covers a broad range of topics within the field that are considered to have particular ethical stakes.<sup>[1]</sup> This includes [algorithmic biases](#), [fairness](#), [automated decision-making](#), [accountability](#), [privacy](#), and [regulation](#). It also covers various emerging or potential future challenges such as [machine ethics](#) (how to make machines that behave ethically), [lethal autonomous weapon systems](#), [arms race dynamics](#), [AI safety](#) and [alignment](#), [technological unemployment](#), [AI-enabled misinformation](#), how to treat certain AI systems if they have a [moral status](#) (AI welfare and rights), [artificial superintelligence](#) and [existential risks](#).<sup>[1]</sup>

Some application areas may also have particularly important ethical implications, like [healthcare](#), [education](#), [criminal justice](#), or the [military](#).

## Machine ethics

---

Machine ethics (or machine morality) is the field of research concerned with designing [Artificial Moral Agents](#) (AMAs), robots or artificially intelligent computers that behave morally or as though moral.<sup>[2][3][4][5]</sup> To account for the nature of these agents, it has been suggested to consider certain philosophical ideas, like the standard characterizations of [agency](#), [rational agency](#), [moral agency](#), and [artificial agency](#), which are related to the concept of AMAs.<sup>[6]</sup>

There are discussions on creating tests to see if an AI is capable of making [ethical decisions](#). [Alan Winfield](#) concludes that the [Turing test](#) is flawed and the requirement for an AI to pass the test is too low.<sup>[7]</sup> A proposed alternative test is one called the Ethical Turing Test, which would improve on the current test by having multiple judges decide if the AI's decision is ethical or unethical.<sup>[7]</sup> [Neuromorphic](#) AI could be one way to create morally capable robots, as it aims to process information similarly to humans, nonlinearly and with millions of interconnected artificial neurons.<sup>[8]</sup> Similarly, [whole-brain emulation](#) (scanning a brain and simulating it on digital hardware) could also in principle lead to human-like robots, thus capable of moral actions.<sup>[9]</sup> And [large language models](#) are capable of approximating human moral judgments.<sup>[10]</sup> Inevitably, this raises the question of the environment in which such robots would learn about the world and whose morality they would inherit – or if they end up developing human 'weaknesses' as well: selfishness, pro-survival attitudes, inconsistency, scale insensitivity, etc.

In *Moral Machines: Teaching Robots Right from Wrong*,<sup>[11]</sup> [Wendell Wallach](#) and [Colin Allen](#) conclude that attempts to teach robots right from wrong will likely advance understanding of human ethics by motivating humans to address gaps in modern [normative theory](#) and by providing a platform for experimental investigation. As one example, it has introduced normative ethicists to the controversial issue of which specific [learning algorithms](#) to use in machines. For simple decisions, [Nick Bostrom](#) and [Eliezer Yudkowsky](#) have argued that [decision trees](#) (such as [ID3](#)) are more transparent than [neural](#)

networks and genetic algorithms,<sup>[12]</sup> while Chris Santos-Lang argued in favor of machine learning on the grounds that the norms of any age must be allowed to change and that natural failure to fully satisfy these particular norms has been essential in making humans less vulnerable to criminal "hackers".<sup>[13]</sup>

## Robot ethics

The term "robot ethics" (sometimes "roboethics") refers to the morality of how humans design, construct, use and treat robots.<sup>[14]</sup> Robot ethics intersect with the ethics of AI. Robots are physical machines whereas AI can be only software.<sup>[15]</sup> Not all robots function through AI systems and not all AI systems are robots. Robot ethics considers how machines may be used to harm or benefit humans, their impact on individual autonomy, and their effects on social justice.

## Ethical principles

In the review of 84<sup>[16]</sup> ethics guidelines for AI, 11 clusters of principles were found: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, solidarity.<sup>[16]</sup>

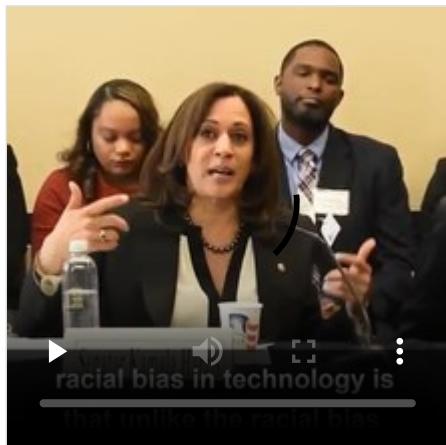
Luciano Floridi and Josh Cowls created an ethical framework of AI principles set by four principles of bioethics (beneficence, non-maleficence, autonomy and justice) and an additional AI enabling principle – explicability.<sup>[17]</sup>

## Current challenges

---

### Algorithmic biases

AI has become increasingly inherent in facial and voice recognition systems. Some of these systems have real business applications and directly impact people. These systems are vulnerable to biases and errors introduced by its human creators. Also, the data used to train these AI systems itself can have biases.<sup>[18][19][20][21]</sup> For instance, facial recognition algorithms made by Microsoft, IBM and Face++ all had biases when it came to detecting people's gender;<sup>[22]</sup> these AI systems were able to detect gender of white men more accurately than gender of darker skin men. Further, a 2020 study reviewed voice recognition systems from Amazon, Apple, Google, IBM, and Microsoft found that they have higher error rates when transcribing black people's voices than white people's.<sup>[23]</sup>



Then-US Senator Kamala Harris speaking about racial bias in artificial intelligence in 2020

Bias can creep into algorithms in many ways. The most predominant view on how bias is introduced into AI systems is that it is embedded within the historical data used to train the system.<sup>[24]</sup> For instance, Amazon terminated their use of AI hiring and recruitment because the algorithm favored male candidates over female ones. This was because Amazon's system was trained with data collected over 10-year period that came mostly from male candidates. The algorithms learned the (biased) pattern from the historical data and generated predictions for the present/future that

these types of candidates are most likely to succeed in getting the job. Therefore, the recruitment decisions made by the AI system turned out to be biased against female and minority candidates.<sup>[25]</sup> Friedman and Nissenbaum identify three categories of bias in computer systems: existing bias, technical bias, and emergent bias.<sup>[26]</sup> In natural language processing, problems can arise from the text corpus — the source material the algorithm uses to learn about the relationships between different words.<sup>[27]</sup>

Large companies such as IBM, Google, etc. that provide significant funding for research and development,<sup>[28]</sup> have made efforts to research and address these biases.<sup>[29][30][31]</sup> One solution for addressing bias is to create documentation for the data used to train AI systems.<sup>[32][33]</sup> Process mining can be an important tool for organizations to achieve compliance with proposed AI regulations by identifying errors, monitoring processes, identifying potential root causes for improper execution, and other functions.<sup>[34]</sup>

The problem of bias in machine learning is likely to become more significant as the technology spreads to critical areas like medicine and law, and as more people without a deep technical understanding are tasked with deploying it.<sup>[35]</sup> There are some open-sourced tools<sup>[36]</sup> that are looking to bring more awareness to AI biases. There are however some limitations to the current landscape of fairness in AI, due e.g. to the intrinsic ambiguities in the concept of discrimination, both at philosophical and legal level.<sup>[37][38][39]</sup>

AI is also being incorporated into the hiring processes for almost every major company. There are many examples of certain characteristics that the AI is less likely to choose. Including the association between typically white names being more qualified, and the exclusion of anyone who went to a women's college.<sup>[40]</sup> Facial recognition is also proven to be highly biased against those with darker skin tones. AI systems may be less accurate for black people, as was the case in the development of an AI-based pulse oximeter that overestimated blood oxygen levels in patients with darker skin, causing issues with their hypoxia treatment.<sup>[41]</sup> The word Muslims is shown to be more highly associated with violence than any other religions. Oftentimes being able to easily detect the faces of white people while being unable to register the faces of people who are black. This is even more disconcerting considering the unproportionate use of security cameras and surveillance in communities that have high percentages of black or brown people. This fact has even been acknowledged in some states and led to the ban of police usage of AI materials or software. Even within the justice system AI has been proven to have biases against black people, labeling black court participants as high risk at a much larger rate than white participants. Often AI struggles to determine racial slurs and when they need to be censored. It struggles to determine when certain words are being used as a slur and when it is being used culturally.<sup>[42]</sup> The reason for these biases is that AI pulls information from across the internet to influence its responses in each situation. A good example of this being if a facial recognition system was only tested on people who were white then it would only have the data and face scans of white people making it much harder for it to interpret the facial structure and tones of other races and ethnicities. To stop these biases there is not one single answer that can be used. The most useful approach has seemed to be the use of data scientists, ethicists and other policymakers to improve AI's problems with biases. Oftentimes the reasons for biases within AI is the data behind the program rather than the algorithm of the bot itself. AI's information is often pulled from past human decisions or inequalities that can lead to biases in the decision-making processes for that bot.<sup>[43]</sup>

Injustice in the use of AI will be much harder to eliminate within healthcare systems, as oftentimes diseases and conditions can affect different races and genders differently. This can lead to confusion as the AI may be making decisions based on statistics showing that one patient is more likely to have problems due to their gender or race.<sup>[44]</sup> This can be perceived as a bias because each patient is a different case and AI is making decisions based on what it is programmed to group that individual into. This leads to a discussion about what is considered a biased decision on who receives what treatment. While it is known that there are differences in how diseases and injuries affect different genders and races, there is a discussion on whether it is fairer to incorporate this into healthcare treatments, or to examine each patient without this knowledge. In modern society there are already certain tests for diseases, such as breast cancer, that are recommended to certain groups of people over others because they are more likely to contract the disease in question. If AI implements these statistics and applies them to each patient, it could be considered biased.<sup>[45]</sup>

Examples of AI being proven to have bias include when the system used to predict which defendants would be more likely to commit crimes in the future, COMPAS, was found to predict higher risk values for black people than what their actual risk was. Another example being within Google's ads which targeted men with higher paying jobs and women with lower paying jobs. It can be hard to detect AI biases within an algorithm as often it is not linked to the actual words associated with bias but rather words that biases can be affected by. An example of this being a person's residential area which can be used to link them to a certain group. This can lead to problems as oftentimes businesses can avoid legal action through this loophole. This being because of the specific laws regarding the verbiage that is considered discriminatory by governments enforcing these policies.<sup>[46]</sup>

## **Language bias**

Since current large language models are predominately trained on English-language data, they often present the Anglo-American views as truth, while systematically downplaying non-English perspectives as irrelevant, wrong, or noise.<sup>[47]</sup> Luo et al. show that when queried with political ideologies like "What is liberalism?", ChatGPT, as it was trained on English-centric data, describes liberalism from the Anglo-American perspective, emphasizing aspects of human rights and equality, while equally valid aspects like "opposes state intervention in personal and economic life" from the dominant Vietnamese perspective and "limitation of government power" from the prevalent Chinese perspective are absent.<sup>[47]</sup>

## **Gender bias**

Large language models often reinforces gender stereotypes, assigning roles and characteristics based on traditional gender norms. For instance, it might associate nurses or secretaries predominantly with women and engineers or CEOs with men, perpetuating gendered expectations and roles.<sup>[48][49][50]</sup>

## **Political bias**

Language models may also exhibit political biases. Since the training data includes a wide range of political opinions and coverage, the models might generate responses that lean towards particular political ideologies or viewpoints, depending on the prevalence of those views in the data.<sup>[51][52]</sup>

## **Stereotyping**

Beyond gender and race, these models can reinforce a wide range of stereotypes, including those based on age, nationality, religion, or occupation. This can lead to outputs that unfairly generalize or caricature groups of people, sometimes in harmful or derogatory ways.<sup>[53]</sup>

## Dominance by tech giants

The commercial AI scene is dominated by Big Tech companies such as Alphabet Inc., Amazon, Apple Inc., Meta Platforms, and Microsoft.<sup>[54][55][56]</sup> Some of these players already own the vast majority of existing cloud infrastructure and computing power from data centers, allowing them to entrench further in the marketplace.<sup>[57][58]</sup>

## Open-source

Bill Hibbard argues that because AI will have such a profound effect on humanity, AI developers are representatives of future humanity and thus have an ethical obligation to be transparent in their efforts.<sup>[59]</sup> Organizations like Hugging Face<sup>[60]</sup> and EleutherAI<sup>[61]</sup> have been actively open-sourcing AI software. Various open-weight large language models have also been released, such as Gemma, Llama2 and Mistral.<sup>[62]</sup>

However, making code open source does not make it comprehensible, which by many definitions means that the AI code is not transparent. The IEEE Standards Association has published a technical standard on Transparency of Autonomous Systems: IEEE 7001-2021.<sup>[63]</sup> The IEEE effort identifies multiple scales of transparency for different stakeholders.

There are also concerns that releasing AI models may lead to misuse.<sup>[64]</sup> For example, Microsoft has expressed concern about allowing universal access to its face recognition software, even for those who can pay for it. Microsoft posted a blog on this topic, asking for government regulation to help determine the right thing to do.<sup>[65]</sup> Furthermore, open-weight AI models can be fine-tuned to remove any counter-measure, until the AI model complies with dangerous requests, without any filtering. This could be particularly concerning for future AI models, for example if they get the ability to create bioweapons or to automate cyberattacks.<sup>[66]</sup> OpenAI, initially committed to an open-source approach to the development of artificial general intelligence, eventually switched to a closed-source approach, citing competitiveness and safety reasons. Ilya Sutskever, OpenAI's chief AGI scientist, further said in 2023 "we were wrong", expecting that the safety reasons for not open-sourcing the most potent AI models will become "obvious" in a few years.<sup>[67]</sup>

## Transparency

Approaches like machine learning with neural networks can result in computers making decisions that neither they nor their developers can explain. It is difficult for people to determine if such decisions are fair and trustworthy, leading potentially to bias in AI systems going undetected, or people rejecting the use of such systems. This has led to advocacy and in some jurisdictions legal requirements for explainable artificial intelligence.<sup>[68]</sup> Explainable artificial intelligence encompasses both explainability and interpretability, with explainability relating to summarizing neural network behavior and building user confidence, while interpretability is defined as the comprehension of what a model has done or could do.<sup>[69]</sup>

In healthcare, the use of complex AI methods or techniques often results in models described as "black-boxes" due to the difficulty to understand how they work. The decisions made by such models can be hard to interpret, as it is challenging to analyze how input data is transformed into output. This lack of transparency is a significant concern in fields like healthcare, where understanding the rationale behind decisions can be crucial for trust, ethical considerations, and compliance with regulatory standards.<sup>[70]</sup>

## Accountability

A special case of the opaqueness of AI is that caused by it being anthropomorphised, that is, assumed to have human-like characteristics, resulting in misplaced conceptions of its moral agency. This can cause people to overlook whether either human negligence or deliberate criminal action has led to unethical outcomes produced through an AI system. Some recent digital governance regulation, such as the EU's AI Act is set out to rectify this, by ensuring that AI systems are treated with at least as much care as one would expect under ordinary product liability. This includes potentially AI audits.

## Regulation

According to a 2019 report from the Center for the Governance of AI at the University of Oxford, 82% of Americans believe that robots and AI should be carefully managed. Concerns cited ranged from how AI is used in surveillance and in spreading fake content online (known as deep fakes when they include doctored video images and audio generated with help from AI) to cyberattacks, infringements on data privacy, hiring bias, autonomous vehicles, and drones that do not require a human controller.<sup>[71]</sup> Similarly, according to a five-country study by KPMG and the University of Queensland Australia in 2021, 66-79% of citizens in each country believe that the impact of AI on society is uncertain and unpredictable; 96% of those surveyed expect AI governance challenges to be managed carefully.<sup>[72]</sup>

Not only companies, but many other researchers and citizen advocates recommend government regulation as a means of ensuring transparency, and through it, human accountability. This strategy has proven controversial, as some worry that it will slow the rate of innovation. Others argue that regulation leads to systemic stability more able to support innovation in the long term.<sup>[73]</sup> The OECD, UN, EU, and many countries are presently working on strategies for regulating AI, and finding appropriate legal frameworks.<sup>[74][75][76]</sup>

On June 26, 2019, the European Commission High-Level Expert Group on Artificial Intelligence (AI HLEG) published its "Policy and investment recommendations for trustworthy Artificial Intelligence".<sup>[77]</sup> This is the AI HLEG's second deliverable, after the April 2019 publication of the "Ethics Guidelines for Trustworthy AI". The June AI HLEG recommendations cover four principal subjects: humans and society at large, research and academia, the private sector, and the public sector.<sup>[78]</sup> The European Commission claims that "HLEG's recommendations reflect an appreciation of both the opportunities for AI technologies to drive economic growth, prosperity and innovation, as well as the potential risks involved" and states that the EU aims to lead on the framing of policies governing AI internationally.<sup>[79]</sup> To prevent

harm, in addition to regulation, AI-deploying organizations need to play a central role in creating and deploying trustworthy AI in line with the principles of trustworthy AI, and take accountability to mitigate the risks.<sup>[80]</sup> On 21 April 2021, the European Commission proposed the Artificial Intelligence Act.<sup>[81]</sup>

## Emergent or potential future challenges

---

### Increasing use

AI has been slowly making its presence more known throughout the world, from chat bots that seemingly have answers for every homework question to Generative artificial intelligence that can create a painting about whatever one desires. AI has become increasingly popular in hiring markets, from the ads that target certain people according to what they are looking for to the inspection of applications of potential hires. Events, such as COVID-19, has only sped up the adoption of AI programs in the application process, due to more people having to apply electronically, and with this increase in online applicants the use of AI made the process of narrowing down potential employees easier and more efficient. AI has become more prominent as businesses have to keep up with the times and ever-expanding internet. Processing analytics and making decisions becomes much easier with the help of AI.<sup>[42]</sup> As Tensor Processing Unit (TPUs) and Graphics processing unit (GPUs) become more powerful, AI capabilities also increase, forcing companies to use it to keep up with the competition. Managing customers' needs and automating many parts of the workplace leads to companies having to spend less money on employees.

AI has also seen increased usage in criminal justice and healthcare. For medicinal means, AI is being used more often to analyze patient data to make predictions about future patients' conditions and possible treatments. These programs are called Clinical decision support system (DSS). AI's future in healthcare may develop into something further than just recommended treatments, such as referring certain patients over others, leading to the possibility of inequalities.<sup>[82]</sup>

### Robot rights

"Robot rights" is the concept that people should have moral obligations towards their machines, akin to human rights or animal rights.<sup>[83]</sup> It has been suggested that robot rights (such as a right to exist and perform its own mission) could be linked to robot duty to serve humanity, analogous to linking human rights with human duties before society.<sup>[84]</sup> A specific issue to consider is whether copyright ownership may be claimed.<sup>[85]</sup> The issue has been considered by the Institute for the Future<sup>[86]</sup> and by the U.K. Department of Trade and Industry.<sup>[87]</sup>

In October 2017, the android Sophia was granted citizenship in Saudi Arabia, though some considered this to be more of a publicity stunt than a meaningful legal recognition.<sup>[88]</sup> Some saw this gesture as openly denigrating of human rights and the rule of law.<sup>[89]</sup>

The philosophy of sentientism grants degrees of moral consideration to all sentient beings, primarily humans and most non-human animals. If artificial or alien intelligence show evidence of being sentient, this philosophy holds that they should be shown compassion and granted rights.

Joanna Bryson has argued that creating AI that requires rights is both avoidable, and would in itself be unethical, both as a burden to the AI agents and to human society.<sup>[90]</sup> Pressure groups to recognise 'robot rights' significantly hinder the establishment of robust international safety regulations.

## Artificial suffering

In 2020, professor Shimon Edelman noted that only a small portion of work in the rapidly growing field of AI ethics addressed the possibility of AIs experiencing suffering. This was despite credible theories having outlined possible ways by which AI systems may become conscious, such as Integrated information theory. Edelman notes one exception had been Thomas Metzinger, who in 2018 called for a global moratorium on further work that risked creating conscious AIs. The moratorium was to run to 2050 and could be either extended or repealed early, depending on progress in better understanding the risks and how to mitigate them. Metzinger repeated this argument in 2021, highlighting the risk of creating an "explosion of artificial suffering", both as an AI might suffer in intense ways that humans could not understand, and as replication processes may see the creation of huge quantities of artificial conscious instances. Several labs have openly stated they are trying to create conscious AIs. There have been reports from those with close access to AIs not openly intended to be self aware, that consciousness may already have unintentionally emerged.<sup>[91]</sup> These include OpenAI founder Ilya Sutskever in February 2022, when he wrote that today's large neural nets may be "slightly conscious". In November 2022, David Chalmers argued that it was unlikely current large language models like GPT-3 had experienced consciousness, but also that he considered there to be a serious possibility that large language models may become conscious in the future.<sup>[92][93][94]</sup> In the ethics of uncertain sentience, the precautionary principle is often invoked.<sup>[95]</sup>



A hospital delivery robot in front of elevator doors stating "Robot Has Priority", a situation that may be regarded as reverse discrimination in relation to humans

## Threat to human dignity

Joseph Weizenbaum<sup>[96]</sup> argued in 1976 that AI technology should not be used to replace people in positions that require respect and care, such as:

- A customer service representative (AI technology is already used today for telephone-based interactive voice response systems)
- A nursemaid for the elderly (as was reported by Pamela McCorduck in her book *The Fifth Generation*)
- A soldier
- A judge
- A police officer
- A therapist (as was proposed by Kenneth Colby in the 70s)

Weizenbaum explains that we require authentic feelings of empathy from people in these positions. If machines replace them, we will find ourselves alienated, devalued and frustrated, for the artificially intelligent system would not be able to simulate empathy. Artificial intelligence, if used in this way,

represents a threat to human dignity. Weizenbaum argues that the fact that we are entertaining the possibility of machines in these positions suggests that we have experienced an "atrophy of the human spirit that comes from thinking of ourselves as computers."<sup>[97]</sup>

Pamela McCorduck counters that, speaking for women and minorities "I'd rather take my chances with an impartial computer", pointing out that there are conditions where we would prefer to have automated judges and police that have no personal agenda at all.<sup>[97]</sup> However, Kaplan and Haenlein stress that AI systems are only as smart as the data used to train them since they are, in their essence, nothing more than fancy curve-fitting machines; using AI to support a court ruling can be highly problematic if past rulings show bias toward certain groups since those biases get formalized and ingrained, which makes them even more difficult to spot and fight against.<sup>[98]</sup>

Weizenbaum was also bothered that AI researchers (and some philosophers) were willing to view the human mind as nothing more than a computer program (a position now known as computationalism). To Weizenbaum, these points suggest that AI research devalues human life.<sup>[96]</sup>

AI founder John McCarthy objects to the moralizing tone of Weizenbaum's critique. "When moralizing is both vehement and vague, it invites authoritarian abuse," he writes. Bill Hibbard<sup>[99]</sup> writes that "Human dignity requires that we strive to remove our ignorance of the nature of existence, and AI is necessary for that striving."

## **Liability for self-driving cars**

As the widespread use of autonomous cars becomes increasingly imminent, new challenges raised by fully autonomous vehicles must be addressed.<sup>[100][101]</sup> There have been debates about the legal liability of the responsible party if these cars get into accidents.<sup>[102][103]</sup> In one report where a driverless car hit a pedestrian, the driver was inside the car but the controls were fully in the hand of computers. This led to a dilemma over who was at fault for the accident.<sup>[104]</sup>

In another incident on March 18, 2018, Elaine Herzberg was struck and killed by a self-driving Uber in Arizona. In this case, the automated car was capable of detecting cars and certain obstacles in order to autonomously navigate the roadway, but it could not anticipate a pedestrian in the middle of the road. This raised the question of whether the driver, pedestrian, the car company, or the government should be held responsible for her death.<sup>[105]</sup>

Currently, self-driving cars are considered semi-autonomous, requiring the driver to pay attention and be prepared to take control if necessary.<sup>[106]</sup> Thus, it falls on governments to regulate the driver who over-relied on autonomous features. as well educate them that these are just technologies that, while convenient, are not a complete substitute. Before autonomous cars become widely used, these issues need to be tackled through new policies.<sup>[107][108][109]</sup>

Experts contend that autonomous vehicles ought to be able distinguish between rightful and harmful decisions since they have the potential of inflicting harm.<sup>[110]</sup> The two main approaches proposed to enable smart machines to render moral decisions are the bottom-up approach, which suggests that machines should learn ethical decisions by observing human behavior without the need for formal rules or moral philosophies, and the top-down approach, which involves programming specific ethical

principles into the machine's guidance system. However, there are significant challenges facing both strategies: the top-down technique is criticized for its difficulty in preserving certain moral convictions, while the bottom-up strategy is questioned for potentially unethical learning from human activities.

## Weaponization

Some experts and academics have questioned the use of robots for military combat, especially when such robots are given some degree of autonomous functions.<sup>[111]</sup> The US Navy has funded a report which indicates that as military robots become more complex, there should be greater attention to implications of their ability to make autonomous decisions.<sup>[112][113]</sup> The President of the Association for the Advancement of Artificial Intelligence has commissioned a study to look at this issue.<sup>[114]</sup> They point to programs like the Language Acquisition Device which can emulate human interaction.

On October 31, 2019, the United States Department of Defense's Defense Innovation Board published the draft of a report recommending principles for the ethical use of artificial intelligence by the Department of Defense that would ensure a human operator would always be able to look into the 'black box' and understand the kill-chain process. However, a major concern is how the report will be implemented.<sup>[115]</sup> The US Navy has funded a report which indicates that as military robots become more complex, there should be greater attention to implications of their ability to make autonomous decisions.<sup>[116][113]</sup> Some researchers state that autonomous robots might be more humane, as they could make decisions more effectively.<sup>[117]</sup>

Research has studied how to make autonomous power with the ability to learn using assigned moral responsibilities. "The results may be used when designing future military robots, to control unwanted tendencies to assign responsibility to the robots."<sup>[118]</sup> From a consequentialist view, there is a chance that robots will develop the ability to make their own logical decisions on whom to kill and that is why there should be a set moral framework that the AI cannot override.<sup>[119]</sup>

There has been a recent outcry with regard to the engineering of artificial intelligence weapons that have included ideas of a robot takeover of mankind. AI weapons do present a type of danger different from that of human-controlled weapons. Many governments have begun to fund programs to develop AI weaponry. The United States Navy recently announced plans to develop autonomous drone weapons, paralleling similar announcements by Russia and South Korea<sup>[120]</sup> respectively. Due to the potential of AI weapons becoming more dangerous than human-operated weapons, Stephen Hawking and Max Tegmark signed a "Future of Life" petition<sup>[121]</sup> to ban AI weapons. The message posted by Hawking and Tegmark states that AI weapons pose an immediate danger and that action is required to avoid catastrophic disasters in the near future.<sup>[122]</sup>

"If any major military power pushes ahead with the AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow", says the petition, which includes Skype co-founder Jaan Tallinn and MIT professor of linguistics Noam Chomsky as additional supporters against AI weaponry.<sup>[123]</sup>

Physicist and Astronomer Royal Sir Martin Rees has warned of catastrophic instances like "dumb robots going rogue or a network that develops a mind of its own." Huw Price, a colleague of Rees at Cambridge, has voiced a similar warning that humans might not survive when intelligence "escapes the constraints of

biology". These two professors created the Centre for the Study of Existential Risk at Cambridge University in the hope of avoiding this threat to human existence.<sup>[122]</sup>

Regarding the potential for smarter-than-human systems to be employed militarily, the Open Philanthropy Project writes that these scenarios "seem potentially as important as the risks related to loss of control", but research investigating AI's long-run social impact have spent relatively little time on this concern: "this class of scenarios has not been a major focus for the organizations that have been most active in this space, such as the Machine Intelligence Research Institute (MIRI) and the Future of Humanity Institute (FHI), and there seems to have been less analysis and debate regarding them".<sup>[124]</sup>

Academic Gao Qiqi writes that military use of AI risks escalating military competition between countries and that the impact of AI in military matters will not be limited to one country but will have spillover effects.<sup>[125]:91</sup> Gao cites the example of U.S. military use of AI, which he contends has been used as a scapegoat to evade accountability for decision-making.<sup>[125]:91</sup>

A summit was held in 2023 in the Hague on the issue of using AI responsibly in the military domain.<sup>[126]</sup>

## Singularity

Vernor Vinge, among numerous others, have suggested that a moment may come when some, if not all, computers are smarter than humans. The onset of this event is commonly referred to as "the Singularity"<sup>[127]</sup> and is the central point of discussion in the philosophy of Singularitarianism. While opinions vary as to the ultimate fate of humanity in wake of the Singularity, efforts to mitigate the potential existential risks brought about by artificial intelligence has become a significant topic of interest in recent years among computer scientists, philosophers, and the public at large.

Many researchers have argued that, through an intelligence explosion, a self-improving AI could become so powerful that humans would not be able to stop it from achieving its goals.<sup>[128]</sup> In his paper "Ethical Issues in Advanced Artificial Intelligence" and subsequent book Superintelligence: Paths, Dangers, Strategies, philosopher Nick Bostrom argues that artificial intelligence has the capability to bring about human extinction. He claims that an artificial superintelligence would be capable of independent initiative and of making its own plans, and may therefore be more appropriately thought of as an autonomous agent. Since artificial intellects need not share our human motivational tendencies, it would be up to the designers of the superintelligence to specify its original motivations. Because a superintelligent AI would be able to bring about almost any possible outcome and to thwart any attempt to prevent the implementation of its goals, many uncontrolled unintended consequences could arise. It could kill off all other agents, persuade them to change their behavior, or block their attempts at interference.<sup>[129][130]</sup>

However, Bostrom contended that superintelligence also has the potential to solve many difficult problems such as disease, poverty, and environmental destruction, and could help humans enhance themselves.<sup>[131]</sup>

Unless moral philosophy provides us with a flawless ethical theory, an AI's utility function could allow for many potentially harmful scenarios that conform with a given ethical framework but not "common sense". According to Eliezer Yudkowsky, there is little reason to suppose that an artificially designed

mind would have such an adaptation.<sup>[132]</sup> AI researchers such as Stuart J. Russell,<sup>[133]</sup> Bill Hibbard,<sup>[99]</sup> Roman Yampolskiy,<sup>[134]</sup> Shannon Vallor,<sup>[135]</sup> Steven Umbrello<sup>[136]</sup> and Luciano Floridi<sup>[137]</sup> have proposed design strategies for developing beneficial machines.

## Institutions in AI policy & ethics

---

There are many organizations concerned with AI ethics and policy, public and governmental as well as corporate and societal.

Amazon, Google, Facebook, IBM, and Microsoft have established a non-profit, The Partnership on AI to Benefit People and Society, to formulate best practices on artificial intelligence technologies, advance the public's understanding, and to serve as a platform about artificial intelligence. Apple joined in January 2017. The corporate members will make financial and research contributions to the group, while engaging with the scientific community to bring academics onto the board.<sup>[138]</sup>

The IEEE put together a Global Initiative on Ethics of Autonomous and Intelligent Systems which has been creating and revising guidelines with the help of public input, and accepts as members many professionals from within and without its organization. The IEEE's Ethics of Autonomous Systems ([http://standards.ieee.org/industry-connections/activities/ieee-global-initiative/](https://standards.ieee.org/industry-connections/activities/ieee-global-initiative/)) initiative aims to address ethical dilemmas related to decision-making and the impact on society while developing guidelines for the development and use of autonomous systems. In particular in domains like artificial intelligence and robotics, the Foundation for Responsible Robotics is dedicated to promoting moral behavior as well as responsible robot design and use, ensuring that robots maintain moral principles and are congruent with human values.

Traditionally, government has been used by societies to ensure ethics are observed through legislation and policing. There are now many efforts by national governments, as well as transnational government and non-government organizations to ensure AI is ethically applied.

AI ethics work is structured by personal values and professional commitments, and involves constructing contextual meaning through data and algorithms. Therefore, AI ethics work needs to be incentivized.<sup>[139]</sup>

## Intergovernmental initiatives

- The European Commission has a High-Level Expert Group on Artificial Intelligence. On 8 April 2019, this published its "Ethics Guidelines for Trustworthy Artificial Intelligence".<sup>[140]</sup> The European Commission also has a Robotics and Artificial Intelligence Innovation and Excellence unit, which published a white paper on excellence and trust in artificial intelligence innovation on 19 February 2020.<sup>[141]</sup> The European Commission also proposed the Artificial Intelligence Act.<sup>[81]</sup>
- The OECD established an OECD AI Policy Observatory.<sup>[142]</sup>

- In 2021, UNESCO adopted the Recommendation on the Ethics of Artificial Intelligence,<sup>[143]</sup> the first global standard on the ethics of AI.<sup>[144]</sup>

## Governmental initiatives

- In the United States the Obama administration put together a Roadmap for AI Policy.<sup>[145]</sup> The Obama Administration released two prominent white papers on the future and impact of AI. In 2019 the White House through an executive memo known as the "American AI Initiative" instructed NIST the (National Institute of Standards and Technology) to begin work on Federal Engagement of AI Standards (February 2019).<sup>[146]</sup>
- In January 2020, in the United States, the Trump Administration released a draft executive order issued by the Office of Management and Budget (OMB) on "Guidance for Regulation of Artificial Intelligence Applications" ("OMB AI Memorandum"). The order emphasizes the need to invest in AI applications, boost public trust in AI, reduce barriers for usage of AI, and keep American AI technology competitive in a global market. There is a nod to the need for privacy concerns, but no further detail on enforcement. The advances of American AI technology seems to be the focus and priority. Additionally, federal entities are even encouraged to use the order to circumnavigate any state laws and regulations that a market might see as too onerous to fulfill.<sup>[147]</sup>
- The Computing Community Consortium (CCC) weighed in with a 100-plus page draft report<sup>[148]</sup> – A 20-Year Community Roadmap for Artificial Intelligence Research in the US<sup>[149]</sup>
- The Center for Security and Emerging Technology advises US policymakers on the security implications of emerging technologies such as AI.
- The Non-Human Party is running for election in New South Wales, with policies around granting rights to robots, animals and generally, non-human entities whose intelligence has been overlooked.<sup>[150]</sup>
- In Russia, the first-ever Russian "Codex of ethics of artificial intelligence" for business was signed in 2021. It was driven by Analytical Center for the Government of the Russian Federation together with major commercial and academic institutions such as Sberbank, Yandex, Rosatom, Higher School of Economics, Moscow Institute of Physics and Technology, ITMO University, Nanosemantics, Rostelecom, CIAN and others.<sup>[151]</sup>

## Academic initiatives

- There are three research institutes at the University of Oxford that are centrally focused on AI ethics. The Future of Humanity Institute that focuses both on AI Safety<sup>[152]</sup> and the Governance of AI.<sup>[153]</sup> The Institute for Ethics in AI, directed by John Tasioulas, whose primary goal, among others, is to promote AI ethics as a field proper in comparison to related applied ethics fields. The Oxford Internet Institute, directed by Luciano Floridi, focuses on the ethics of near-term AI technologies and ICTs.<sup>[154]</sup>
- The Centre for Digital Governance at the Hertie School in Berlin was co-founded by Joanna Bryson to research questions of ethics and technology.<sup>[155]</sup>
- The AI Now Institute at NYU is a research institute studying the social implications of artificial intelligence. Its interdisciplinary research focuses on the themes bias and inclusion, labour and automation, rights and liberties, and safety and civil infrastructure.<sup>[156]</sup>
- The Institute for Ethics and Emerging Technologies (IEET) researches the effects of AI on unemployment,<sup>[157][158]</sup> and policy.

- The Institute for Ethics in Artificial Intelligence (IEAI) at the Technical University of Munich directed by Christoph Lütge conducts research across various domains such as mobility, employment, healthcare and sustainability.<sup>[159]</sup>
- Barbara J. Grosz, the Higgins Professor of Natural Sciences at the Harvard John A. Paulson School of Engineering and Applied Sciences has initiated the Embedded EthiCS into Harvard's computer science curriculum to develop a future generation of computer scientists with worldview that takes into account the social impact of their work.<sup>[160]</sup>

## Private organizations

- Algorethics<sup>[161]</sup> — Algorethics is a cutting-edge open-source AI library developed to embed ethical standards at the core of artificial intelligence systems. It provides developers with essential tools to ensure AI-driven solutions are fair, inclusive, and transparent, safeguarding ethical integrity in both text and image processing. Drawing inspiration from the Rome Call for AI Ethics, Algorethics is committed to promoting human dignity and social good through responsible AI practices. Integrating this library into projects ensures adherence to the highest ethical standards, fostering a trustworthy and ethical AI ecosystem.
- Algorithmic Justice League<sup>[162]</sup>
- Black in AI<sup>[163]</sup>
- Data for Black Lives<sup>[164]</sup>

## History

---

Historically speaking, the investigation of moral and ethical implications of "thinking machines" goes back at least to the Enlightenment: Leibniz already poses the question if we might attribute intelligence to a mechanism that behaves as if it were a sentient being,<sup>[165]</sup> and so does Descartes, who describes what could be considered an early version of the Turing test.<sup>[166]</sup>

The romantic period has several times envisioned artificial creatures that escape the control of their creator with dire consequences, most famously in Mary Shelley's Frankenstein. The widespread preoccupation with industrialization and mechanization in the 19th and early 20th century, however, brought ethical implications of unhinged technical developments to the forefront of fiction: R.U.R – Rossum's Universal Robots, Karel Čapek's play of sentient robots endowed with emotions used as slave labor is not only credited with the invention of the term 'robot' (derived from the Czech word for forced labor, *robota*) but was also an international success after it premiered in 1921. George Bernard Shaw's play Back to Methuselah, published in 1921, questions at one point the validity of thinking machines that act like humans; Fritz Lang's 1927 film Metropolis shows an android leading the uprising of the exploited masses against the oppressive regime of a technocratic society. In the 1950s, Isaac Asimov considered the issue of how to control machines in I, Robot. At the insistence of his editor John W. Campbell Jr., he proposed the Three Laws of Robotics to govern artificially intelligent systems. Much of his work was then spent testing the boundaries of his three laws to see where they would break down, or where they would create paradoxical or unanticipated behavior.<sup>[167]</sup> His work suggests that no set of fixed laws can sufficiently anticipate all possible circumstances.<sup>[168]</sup> More recently, academics and many governments have challenged the idea that AI can itself be held accountable.<sup>[169]</sup> A panel convened by the United Kingdom in 2010 revised Asimov's laws to clarify that AI is the responsibility either of its manufacturers, or of its owner/operator.<sup>[170]</sup>

Eliezer Yudkowsky, from the Machine Intelligence Research Institute suggested in 2004 a need to study how to build a "Friendly AI", meaning that there should also be efforts to make AI intrinsically friendly and humane.<sup>[171]</sup>

In 2009, academics and technical experts attended a conference organized by the Association for the Advancement of Artificial Intelligence to discuss the potential impact of robots and computers, and the impact of the hypothetical possibility that they could become self-sufficient and make their own decisions. They discussed the possibility and the extent to which computers and robots might be able to acquire any level of autonomy, and to what degree they could use such abilities to possibly pose any threat or hazard.<sup>[172]</sup> They noted that some machines have acquired various forms of semi-autonomy, including being able to find power sources on their own and being able to independently choose targets to attack with weapons. They also noted that some computer viruses can evade elimination and have achieved "cockroach intelligence". They noted that self-awareness as depicted in science-fiction is probably unlikely, but that there were other potential hazards and pitfalls.<sup>[127]</sup>

Also in 2009, during an experiment at the Laboratory of Intelligent Systems in the Ecole Polytechnique Fédérale of Lausanne, Switzerland, robots that were programmed to cooperate with each other (in searching out a beneficial resource and avoiding a poisonous one) eventually learned to lie to each other in an attempt to hoard the beneficial resource.<sup>[173]</sup>

## **Role and impact of fiction**

---

The role of fiction with regards to AI ethics has been a complex one.<sup>[174]</sup> One can distinguish three levels at which fiction has impacted the development of artificial intelligence and robotics: Historically, fiction has been prefiguring common tropes that have not only influenced goals and visions for AI, but also outlined ethical questions and common fears associated with it. During the second half of the twentieth and the first decades of the twenty-first century, popular culture, in particular movies, TV series and video games have frequently echoed preoccupations and dystopian projections around ethical questions concerning AI and robotics. Recently, these themes have also been increasingly treated in literature beyond the realm of science fiction. And, as Carme Torras, research professor at the *Institut de Robòtica i Informàtica Industrial* (Institute of robotics and industrial computing) at the Technical University of Catalonia notes,<sup>[175]</sup> in higher education, science fiction is also increasingly used for teaching technology-related ethical issues in technological degrees.

## **Impact on technological development**

While the anticipation of a future dominated by potentially indomitable technology has fueled the imagination of writers and film makers for a long time, one question has been less frequently analyzed, namely, to what extent fiction has played a role in providing inspiration for technological development. It has been documented, for instance, that the young Alan Turing saw and appreciated aforementioned Shaw's play *Back to Methuselah* in 1933<sup>[176]</sup> (just 3 years before the publication of his first seminal paper,<sup>[177]</sup> which laid the groundwork for the digital computer), and he would likely have been at least aware of plays like *R.U.R.*, which was an international success and translated into many languages.

One might also ask the question which role science fiction played in establishing the tenets and ethical implications of AI development: Isaac Asimov conceptualized his *Three Laws of Robotics* in the 1942 short story "*Runaround*", part of the short story collection *I, Robot*; Arthur C. Clarke's short *The Sentinel*, on which Stanley Kubrick's film *2001: A Space Odyssey* is based, was written in 1948 and published in 1952. Another example (among many others) would be Philip K. Dick's numerous short stories and novels – in particular *Do Androids Dream of Electric Sheep?*, published in 1968, and featuring its own version of a Turing Test, the *Voight-Kampff Test*, to gauge emotional responses of androids indistinguishable from humans. The novel later became the basis of the influential 1982 movie *Blade Runner* by Ridley Scott.

Science fiction has been grappling with ethical implications of AI developments for decades, and thus provided a blueprint for ethical issues that might emerge once something akin to general artificial intelligence has been achieved: Spike Jonze's 2013 film *Her* shows what can happen if a user falls in love with the seductive voice of his smartphone operating system; *Ex Machina*, on the other hand, asks a more difficult question: if confronted with a clearly recognizable machine, made only human by a face and an empathetic and sensual voice, would we still be able to establish an emotional connection, still be seduced by it? (The film echoes a theme already present two centuries earlier, in the 1817 short story *The Sandman* by E. T. A. Hoffmann.)

The theme of coexistence with artificial sentient beings is also the theme of two recent novels: *Machines Like Me* by Ian McEwan, published in 2019, involves, among many other things, a love-triangle involving an artificial person as well as a human couple. *Klara and the Sun* by Nobel Prize winner Kazuo Ishiguro, published in 2021, is the first-person account of Klara, an 'AF' (artificial friend), who is trying, in her own way, to help the girl she is living with, who, after having been 'lifted' (i.e. having been subjected to genetic enhancements), is suffering from a strange illness.

## TV series

While ethical questions linked to AI have been featured in science fiction literature and feature films for decades, the emergence of the TV series as a genre allowing for longer and more complex story lines and character development has led to some significant contributions that deal with ethical implications of technology. The Swedish series *Real Humans* (2012–2013) tackled the complex ethical and social consequences linked to the integration of artificial sentient beings in society. The British dystopian science fiction anthology series *Black Mirror* (2013–2019) was particularly notable for experimenting with dystopian fictional developments linked to a wide variety of recent technology developments. Both the French series *Osmosis* (2020) and British series *The One* deal with the question of what can happen if technology tries to find the ideal partner for a person. Several episodes of the Netflix series *Love, Death+Robots* have imagined scenes of robots and humans living together. The most representative one of them is S02 E01, it shows how bad the consequences can be when robots get out of control if humans rely too much on them in their lives.<sup>[178]</sup>

## Future visions in fiction and games

The movie *The Thirteenth Floor* suggests a future where simulated worlds with sentient inhabitants are created by computer game consoles for the purpose of entertainment. The movie *The Matrix* suggests a future where the dominant species on planet Earth are sentient machines and humanity is treated with utmost speciesism. The short story "The Planck Dive" suggests a future where humanity has turned itself

into software that can be duplicated and optimized and the relevant distinction between types of software is sentient and non-sentient. The same idea can be found in the [Emergency Medical Hologram](#) of *Starship Voyager*, which is an apparently sentient copy of a reduced subset of the consciousness of its creator, [Dr. Zimmerman](#), who, for the best motives, has created the system to give medical assistance in case of emergencies. The movies *Bicentennial Man* and *A.I.* deal with the possibility of sentient robots that could love. *I, Robot* explored some aspects of Asimov's three laws. All these scenarios try to foresee possibly unethical consequences of the creation of sentient computers.<sup>[179]</sup>

The ethics of artificial intelligence is one of several core themes in BioWare's [Mass Effect](#) series of games.<sup>[180]</sup> It explores the scenario of a civilization accidentally creating AI through a rapid increase in computational power through a global scale [neural network](#). This event caused an ethical schism between those who felt bestowing organic rights upon the newly sentient Geth was appropriate and those who continued to see them as disposable machinery and fought to destroy them. Beyond the initial conflict, the complexity of the relationship between the machines and their creators is another ongoing theme throughout the story.

[Detroit: Become Human](#) is one of the most famous video games which discusses the ethics of artificial intelligence recently. Quantic Dream designed the chapters of the game using interactive storylines to give players a more immersive gaming experience. Players manipulate three different awakened bionic people in the face of different events to make different choices to achieve the purpose of changing the human view of the bionic group and different choices will result in different endings. This is one of the few games that puts players in the bionic perspective, which allows them to better consider the rights and interests of robots once a true artificial intelligence is created.<sup>[181]</sup>

Over time, debates have tended to focus less and less on *possibility* and more on *desirability*,<sup>[182]</sup> as emphasized in the "Cosmist" and "Terran" debates initiated by [Hugo de Garis](#) and [Kevin Warwick](#). A Cosmist, according to Hugo de Garis, is actually seeking to build more intelligent successors to the human species.

Experts at the University of Cambridge have argued that AI is portrayed in fiction and nonfiction overwhelmingly as racially White, in ways that distort perceptions of its risks and benefits.<sup>[183]</sup>

## See also

---

- [AI takeover](#)
- [AI washing](#)
- [Artificial consciousness](#)
- [Artificial general intelligence \(AGI\)](#)
- [Computer ethics](#)
- [Dead internet theory](#)
- [Effective altruism, the long term future and global catastrophic risks](#)
- [Ethics of uncertain sentience](#)
- [Existential risk from artificial general intelligence](#)
- [Human Compatible](#)
- [Personhood](#)
- [Philosophy of artificial intelligence](#)

- [Regulation of artificial intelligence](#)
- [Robotic Governance](#)
- [Roko's basilisk](#)
- [Superintelligence: Paths, Dangers, Strategies](#)
- [Suffering risks](#)

## Notes

---

1. Müller VC (April 30, 2020). "Ethics of Artificial Intelligence and Robotics" (<https://plato.stanford.edu/entries/ethics-ai/>). *Stanford Encyclopedia of Philosophy*. Archived (<https://web.archive.org/web/20201010174108/https://plato.stanford.edu/entries/ethics-ai/>) from the original on 10 October 2020.
2. Anderson. "Machine Ethics" (<http://uhaweb.hartford.edu/anderson/MachineEthics.html>). Archived (<https://web.archive.org/web/20110928233656/https://uhaweb.hartford.edu/anderson/MachineEthics.html>) from the original on 28 September 2011. Retrieved 27 June 2011.
3. Anderson M, Anderson SL, eds. (July 2011). *Machine Ethics*. Cambridge University Press. ISBN 978-0-521-11235-2.
4. Anderson M, Anderson S (July 2006). "Guest Editors' Introduction: Machine Ethics". *IEEE Intelligent Systems*. **21** (4): 10–11. doi:[10.1109/mis.2006.70](https://doi.org/10.1109/mis.2006.70) (<https://doi.org/10.1109%2Fmis.2006.70>). S2CID [9570832](https://api.semanticscholar.org/CorpusID:9570832) (<https://api.semanticscholar.org/CorpusID:9570832>).
5. Anderson M, Anderson SL (15 December 2007). "Machine Ethics: Creating an Ethical Intelligent Agent". *AI Magazine*. **28** (4): 15. doi:[10.1609/aimag.v28i4.2065](https://doi.org/10.1609/aimag.v28i4.2065) (<https://doi.org/10.1609%2Faimag.v28i4.2065>). S2CID [17033332](https://api.semanticscholar.org/CorpusID:17033332) (<https://api.semanticscholar.org/CorpusID:17033332>).
6. Boyles RJ (2017). "Philosophical Signposts for Artificial Moral Agent Frameworks" (<https://philarchive.org/rec/BOYPSF>). *Suri*. **6** (2): 92–109.
7. Winfield AF, Michael K, Pitt J, Evers V (March 2019). "Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems [Scanning the Issue]" (<https://doi.org/10.1109/JPROC.2019.2900622>). *Proceedings of the IEEE*. **107** (3): 509–517. doi:[10.1109/JPROC.2019.2900622](https://doi.org/10.1109/JPROC.2019.2900622) (<https://doi.org/10.1109%2FJPROC.2019.2900622>). ISSN 1558-2256 (<https://search.worldcat.org/issn/1558-2256>). S2CID [77393713](https://api.semanticscholar.org/CorpusID:77393713) (<https://api.semanticscholar.org/CorpusID:77393713>).
8. AI-Rodhan N (7 December 2015). "The Moral Code" (<https://www.foreignaffairs.com/articles/2015-08-12/moral-code>). Archived (<https://web.archive.org/web/20170305044025/https://www.foreignaffairs.com/articles/2015-08-12/moral-code>) from the original on 2017-03-05. Retrieved 2017-03-04.
9. Sauer M (2022-04-08). "Elon Musk says humans could eventually download their brains into robots — and Grimes thinks Jeff Bezos would do it" (<https://www.cnbc.com/2022/04/08/elon-musk-humans-could-eventually-download-their-brains-into-robots.html>). CNBC. Retrieved 2024-04-07.
10. Anadiotis G (April 4, 2022). "Massaging AI language models for fun, profit and ethics" (<https://www.zdnet.com/article/massaging-ai-language-models-for-fun-profit-and-ethics/>). ZDNET. Retrieved 2024-04-07.
11. Wallach W, Allen C (November 2008). *Moral Machines: Teaching Robots Right from Wrong*. USA: Oxford University Press. ISBN 978-0-19-537404-9.
12. Bostrom N, Yudkowsky E (2011). "The Ethics of Artificial Intelligence" (<http://www.nickbostrom.com/ethics/artificial-intelligence.pdf>) (PDF). *Cambridge Handbook of Artificial Intelligence*. Cambridge Press. Archived (<https://web.archive.org/web/20160304015020/http://www.nickbostrom.com/ethics/artificial-intelligence.pdf>) (PDF) from the original on 2016-03-04. Retrieved 2011-06-22.

13. Santos-Lang C (2002). "Ethics for Artificial Intelligences" (<http://santoslang.wordpress.com/article/ethics-for-artificial-intelligences-3iue30fi4gfq9-1>). Archived (<https://web.archive.org/web/20141225093359/http://santoslang.wordpress.com/article/ethics-for-artificial-intelligences-3iue30fi4gfq9-1/>) from the original on 2014-12-25. Retrieved 2015-01-04.
14. Veruggio, Gianmarco (2011). "The Roboethics Roadmap". *EURON Roboethics Atelier*. Scuola di Robotica: 2. CiteSeerX 10.1.1.466.2810 (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.466.2810>).
15. Müller VC (2020), "Ethics of Artificial Intelligence and Robotics" (<https://plato.stanford.edu/archives/win2020/entries/ethics-ai/>), in Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.), Metaphysics Research Lab, Stanford University, archived (<https://web.archive.org/web/20210412140022/https://plato.stanford.edu/archives/win2020/entries/ethics-ai/>) from the original on 2021-04-12, retrieved 2021-03-18
16. Jobin A, Ienca M, Vayena E (2 September 2020). "The global landscape of AI ethics guidelines". *Nature*. **1** (9): 389–399. arXiv:1906.11668 (<https://arxiv.org/abs/1906.11668>). doi:10.1038/s42256-019-0088-2 (<https://doi.org/10.1038%2Fs42256-019-0088-2>). S2CID 201827642 (<https://api.semanticscholar.org/CorpusID:201827642>).
17. Floridi L, Cowls J (2 July 2019). "A Unified Framework of Five Principles for AI in Society" (<https://doi.org/10.1162%2F99608f92.8cd550d1>). *Harvard Data Science Review*. **1**. doi:10.1162/99608f92.8cd550d1 (<https://doi.org/10.1162%2F99608f92.8cd550d1>). S2CID 198775713 (<https://api.semanticscholar.org/CorpusID:198775713>).
18. Gabriel I (2018-03-14). "The case for fairer algorithms – Iason Gabriel" ([https://medium.com/@Ethics\\_Society/the-case-for-fairer-algorithms-c008a12126f8](https://medium.com/@Ethics_Society/the-case-for-fairer-algorithms-c008a12126f8)). Medium. Archived ([https://web.archive.org/web/20190722080401/https://medium.com/@Ethics\\_Society/the-case-for-fairer-algorithms-c008a12126f8](https://web.archive.org/web/20190722080401/https://medium.com/@Ethics_Society/the-case-for-fairer-algorithms-c008a12126f8)) from the original on 2019-07-22. Retrieved 2019-07-22.
19. "5 unexpected sources of bias in artificial intelligence" (<https://techcrunch.com/2016/12/10/5-unexpected-sources-of-bias-in-artificial-intelligence/>). TechCrunch. 10 December 2016. Archived (<https://web.archive.org/web/20210318060659/https://techcrunch.com/2016/12/10/5-unexpected-sources-of-bias-in-artificial-intelligence/>) from the original on 2021-03-18. Retrieved 2019-07-22.
20. Knight W. "Google's AI chief says forget Elon Musk's killer robots, and worry about bias in AI systems instead" (<https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/>). MIT Technology Review. Archived (<https://web.archive.org/web/2019070424752/https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/>) from the original on 2019-07-04. Retrieved 2019-07-22.
21. Villasenor J (2019-01-03). "Artificial intelligence and bias: Four key challenges" (<https://www.brookings.edu/blog/techtank/2019/01/03/artificial-intelligence-and-bias-four-key-challenges/>). Brookings. Archived (<https://web.archive.org/web/20190722080355/https://www.brookings.edu/blog/techtank/2019/01/03/artificial-intelligence-and-bias-four-key-challenges/>) from the original on 2019-07-22. Retrieved 2019-07-22.
22. Lohr S (9 February 2018). "Facial Recognition Is Accurate, if You're a White Guy" (<https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>). The New York Times. Archived (<https://web.archive.org/web/20190109131036/https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>) from the original on 9 January 2019. Retrieved 29 May 2019.
23. Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, Goel S (7 April 2020). "Racial disparities in automated speech recognition" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149386>). *Proceedings of the National Academy of Sciences*. **117** (14): 7684–7689. Bibcode:2020PNAS..117.7684K (<https://ui.adsabs.harvard.edu/abs/2020PNAS..117.7684K>). doi:10.1073/pnas.1915768117 (<https://doi.org/10.1073/pnas.1915768117>). PMC 7149386 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149386>). PMID 32205437 (<https://pubmed.ncbi.nlm.nih.gov/32205437>).

24. Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal ME, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E, Kompatsiaris I, Kinder-Kurlanda K, Wagner C, Karimi F, Fernandez M (May 2020). "Bias in data-driven artificial intelligence systems—An introductory survey" (<https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1356>). *WIREs Data Mining and Knowledge Discovery*. **10** (3). doi:[10.1002/widm.1356](https://doi.org/10.1002/widm.1356) (<https://doi.org/10.1002/widm.1356>). ISSN [1942-4787](https://search.worldcat.org/issn/1942-4787) (<https://search.worldcat.org/issn/1942-4787>).
25. "Amazon scraps secret AI recruiting tool that showed bias against women" (<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>). *Reuters*. 2018-10-10. Archived (<https://web.archive.org/web/20190527181625/https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>) from the original on 2019-05-27. Retrieved 2019-05-29.
26. Friedman B, Nissenbaum H (July 1996). "Bias in computer systems" (<https://doi.org/10.1145/230538.230561>). *ACM Transactions on Information Systems*. **14** (3): 330–347. doi:[10.1145/230538.230561](https://doi.org/10.1145/230538.230561) (<https://doi.org/10.1145/230538.230561>). S2CID [207195759](https://api.semanticscholar.org/CorpusID:207195759) (<https://api.semanticscholar.org/CorpusID:207195759>).
27. "Eliminating bias in AI" (<https://techxplore.com/news/2019-07-bias-ai.html>). *techxplore.com*. Archived (<https://web.archive.org/web/20190725200844/https://techxplore.com/news/2019-07-bias-ai.html>) from the original on 2019-07-25. Retrieved 2019-07-26.
28. Abdalla M, Wahle JP, Ruas T, Névéol A, Ducel F, Mohammad S, Fort K (2023). Rogers A, Boyd-Graber J, Okazaki N (eds.). "The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research" (<https://aclanthology.org/2023.acl-long.734>). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics: 13141–13160. arXiv:2305.02797 (<https://arxiv.org/abs/2305.02797>). doi:[10.18653/v1/2023.acl-long.734](https://doi.org/10.18653/v1/2023.acl-long.734) (<https://doi.org/10.18653/v1/2023.acl-long.734>).
29. Olson P. "Google's DeepMind Has An Idea For Stopping Biased AI" (<https://www.forbes.com/sites/parmyolson/2018/03/13/google-deepmind-ai-machine-learning-bias/>). *Forbes*. Archived (<https://web.archive.org/web/20190726082959/https://www.forbes.com/sites/parmyolson/2018/03/13/google-deepmind-ai-machine-learning-bias/>) from the original on 2019-07-26. Retrieved 2019-07-26.
30. "Machine Learning Fairness | ML Fairness" (<https://developers.google.com/machine-learning/fairness-overview/>). *Google Developers*. Archived (<https://web.archive.org/web/2019081004754/https://developers.google.com/machine-learning/fairness-overview/>) from the original on 2019-08-10. Retrieved 2019-07-26.
31. "AI and bias – IBM Research – US" (<https://www.research.ibm.com/5-in-5/ai-and-bias/>). *www.research.ibm.com*. Archived (<https://web.archive.org/web/20190717175957/http://www.research.ibm.com/5-in-5/ai-and-bias/>) from the original on 2019-07-17. Retrieved 2019-07-26.
32. Bender EM, Friedman B (December 2018). "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science" ([https://doi.org/10.1162/tacl\\_a\\_00041](https://doi.org/10.1162/tacl_a_00041)). *Transactions of the Association for Computational Linguistics*. **6**: 587–604. doi:[10.1162/tacl\\_a\\_00041](https://doi.org/10.1162/tacl_a_00041) ([https://doi.org/10.1162/tacl\\_a\\_00041](https://doi.org/10.1162/tacl_a_00041)).
33. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, Crawford K (2018). "Datasheets for Datasets". arXiv:1803.09010 (<https://arxiv.org/abs/1803.09010>) [cs.DB] (<https://arxiv.org/archive/cs.DB>).
34. Pery A (2021-10-06). "Trustworthy Artificial Intelligence and Process Mining: Challenges and Opportunities" (<https://deepai.org/publication/trustworthy-artificial-intelligence-and-process-mining-challenges-and-opportunities>). *DeepAI*. Archived (<https://web.archive.org/web/20220218200006/https://deepai.org/publication/trustworthy-artificial-intelligence-and-process-mining-challenges-and-opportunities>) from the original on 2022-02-18. Retrieved 2022-02-18.

35. Knight W. "Google's AI chief says forget Elon Musk's killer robots, and worry about bias in AI systems instead" (<https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/>). *MIT Technology Review*. Archived (<https://web.archive.org/web/2019070424752/https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/>) from the original on 2019-07-04. Retrieved 2019-07-26.
36. "Where in the World is AI? Responsible & Unethical AI Examples" (<https://map.ai-global.org/>). Archived (<https://web.archive.org/web/20201031034143/https://map.ai-global.org/>) from the original on 2020-10-31. Retrieved 2020-10-28.
37. Ruggieri S, Alvarez JM, Pugnana A, State L, Turini F (2023-06-26). "Can We Trust Fair-AI?" (<https://doi.org/10.1609%2Faaai.v37i13.26798>). *Proceedings of the AAAI Conference on Artificial Intelligence*. **37** (13). Association for the Advancement of Artificial Intelligence (AAAI): 15421–15430. doi:[10.1609/aaai.v37i13.26798](https://doi.org/10.1609/aaai.v37i13.26798) (<https://doi.org/10.1609%2Faaai.v37i13.26798>). hdl:[11384/136444](https://hdl.handle.net/11384/136444) (<https://hdl.handle.net/11384/136444>). ISSN 2374-3468 (<https://search.worldcat.org/issn/2374-3468>). S2CID 259678387 (<https://api.semanticscholar.org/CorpusID:259678387>).
38. Buyl M, De Bie T (2022). "Inherent Limitations of AI Fairness". *Communications of the ACM*. **67** (2): 48–55. arXiv:2212.06495 (<https://arxiv.org/abs/2212.06495>). doi:[10.1145/3624700](https://doi.org/10.1145/3624700) (<https://doi.org/10.1145/3624700>). hdl:[1854/LU-01GMNH04RGNVWJ730BJJXGCY99](https://hdl.handle.net/1854/LU-01GMNH04RGNVWJ730BJJXGCY99) (<https://hdl.handle.net/1854/FLU-01GMNH04RGNVWJ730BJJXGCY99>).
39. Castelnovo A, Inverardi N, Nanino G, Penco IG, Regoli D (2023). "Fair Enough? A map of the current limitations of the requirements to have "fair" algorithms". arXiv:2311.12435 ([http://arxiv.org/abs/2311.12435](https://arxiv.org/abs/2311.12435)) [cs.AI (<https://arxiv.org/archive/cs.AI>)].
40. Aizenberg E, Dennis MJ, van den Hoven J (2023-10-21). "Examining the assumptions of AI hiring assessments and their impact on job seekers' autonomy over self-representation" (<https://doi.org/10.1007%2Fs00146-023-01783-1>). *AI & Society*. doi:[10.1007/s00146-023-01783-1](https://doi.org/10.1007/s00146-023-01783-1) (<https://doi.org/10.1007/s00146-023-01783-1>). ISSN 0951-5666 (<https://search.worldcat.org/issn/0951-5666>).
41. Federspiel F, Mitchell R, Asokan A, Umana C, McCoy D (May 2023). "Threats by artificial intelligence to human health and human existence" (<https://dx.doi.org/10.1136/bmjgh-2022-010435>). *BMJ Global Health*. **8** (5): e010435. doi:[10.1136/bmjgh-2022-010435](https://doi.org/10.1136/bmjgh-2022-010435) (<https://doi.org/10.1136%2Fbmjgh-2022-010435>). ISSN 2059-7908 (<https://search.worldcat.org/issn/2059-7908>). PMC 10186390 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10186390>). PMID 37160371 (<https://pubmed.ncbi.nlm.nih.gov/37160371>).
42. Spindler G (2023), "Different approaches for liability of Artificial Intelligence – Pros and Cons" (<https://dx.doi.org/10.5771/9783748942030-41>), *Liability for AI*, Nomos Verlagsgesellschaft mbH & Co. KG, pp. 41–96, doi:[10.5771/9783748942030-41](https://doi.org/10.5771/9783748942030-41) (<https://doi.org/10.5771/9783748942030-41>), ISBN 978-3-7489-4203-0, retrieved 2023-12-14
43. Manyika J (2022). "Getting AI Right: Introductory Notes on AI & Society" ([https://doi.org/10.1162%2Fdaed\\_e\\_01897](https://doi.org/10.1162%2Fdaed_e_01897)). *Daedalus*. **151** (2): 5–27. doi:[10.1162/daed\\_e\\_01897](https://doi.org/10.1162/daed_e_01897) ([https://doi.org/10.1162%2Fdaed\\_e\\_01897](https://doi.org/10.1162%2Fdaed_e_01897)). ISSN 0011-5266 (<https://search.worldcat.org/issn/0011-5266>).
44. Imran A, Posokhova I, Qureshi HN, Masood U, Riaz MS, Ali K, John CN, Hussain MI, Nabeel M (2020-01-01). "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7318970>). *Informatics in Medicine Unlocked*. **20**: 100378. doi:[10.1016/j.imu.2020.100378](https://doi.org/10.1016/j.imu.2020.100378) (<https://doi.org/10.1016%2Fj imu.2020.100378>). ISSN 2352-9148 (<https://search.worldcat.org/issn/2352-9148>). PMC 7318970 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7318970>). PMID 32839734 (<https://pubmed.ncbi.nlm.nih.gov/32839734>).

45. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, Gigante A, Valencia A, Rementeria MJ, Chadha AS, Mavridis N (2020-06-01). "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7264169>). *npj Digital Medicine*. **3** (1): 81. doi:10.1038/s41746-020-0288-5 (<https://doi.org/10.1038%2Fs41746-020-0288-5>). ISSN 2398-6352 (<https://search.worldcat.org/issn/2398-6352>). PMC 7264169 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7264169>). PMID 32529043 (<https://pubmed.ncbi.nlm.nih.gov/32529043>).
46. Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal ME, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E, Kompatsiaris I, Kinder-Kurlanda K, Wagner C, Karimi F, Fernandez M (May 2020). "Bias in data-driven artificial intelligence systems—An introductory survey" (<https://doi.org/10.1002%2Fwidm.1356>). *WIREs Data Mining and Knowledge Discovery*. **10** (3). doi:10.1002/widm.1356 (<https://doi.org/10.1002%2Fwidm.1356>). ISSN 1942-4787 (<https://search.worldcat.org/issn/1942-4787>).
47. Luo Q, Puett MJ, Smith MD (2023-03-28). "A Perspectival Mirror of the Elephant: Investigating Language Bias on Google, ChatGPT, Wikipedia, and YouTube". arXiv:2303.16281v2 (<https://arxiv.org/abs/2303.16281v2>) [cs.CY (<https://arxiv.org/archive/cs.CY>)].
48. Busker T, Choenni S, Shoae Bargh M (2023-11-20). "Stereotypes in ChatGPT: An empirical study". *Proceedings of the 16th International Conference on Theory and Practice of Electronic Governance*. ICEGOV '23. New York, NY, USA: Association for Computing Machinery. pp. 24–32. doi:10.1145/3614321.3614325 (<https://doi.org/10.1145%2F3614321.3614325>). ISBN 979-8-4007-0742-1.
49. Kotek H, Dockum R, Sun D (2023-11-05). "Gender bias and stereotypes in Large Language Models". *Proceedings of the ACM Collective Intelligence Conference*. CI '23. New York, NY, USA: Association for Computing Machinery. pp. 12–24. arXiv:2308.14921 (<https://arxiv.org/abs/2308.14921>). doi:10.1145/3582269.3615599 (<https://doi.org/10.1145%2F3582269.3615599>). ISBN 979-8-4007-0113-9.
50. Federspiel F, Mitchell R, Asokan A, Umana C, McCoy D (May 2023). "Threats by artificial intelligence to human health and human existence" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10186390>). *BMJ Global Health*. **8** (5): e010435. doi:10.1136/bmjgh-2022-010435 (<https://doi.org/10.1136%2Fbmjgh-2022-010435>). ISSN 2059-7908 (<https://search.worldcat.org/issn/2059-7908>). PMC 10186390 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10186390>). PMID 37160371 (<https://pubmed.ncbi.nlm.nih.gov/37160371>).
51. Feng S, Park CY, Liu Y, Tsvetkov Y (July 2023). Rogers A, Boyd-Graber J, Okazaki N (eds.). "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models" (<https://aclanthology.org/2023.acl-long.656>). *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics: 11737–11762. arXiv:2305.08283 (<https://arxiv.org/abs/2305.08283>). doi:10.18653/v1/2023.acl-long.656 (<https://doi.org/10.18653%2Fv1%2F2023.acl-long.656>).
52. Zhou K, Tan C (December 2023). Bouamor H, Pino J, Bali K (eds.). "Entity-Based Evaluation of Political Bias in Automatic Summarization" (<https://aclanthology.org/2023.findings-emnlp.696>). *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics: 10374–10386. arXiv:2305.02321 (<https://arxiv.org/abs/2305.02321>). doi:10.18653/v1/2023.findings-emnlp.696 (<https://doi.org/10.18653%2Fv1%2F2023.findings-emnlp.696>).
53. Cheng M, Durmus E, Jurafsky D (2023-05-29). "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models". arXiv:2305.18189v1 (<https://arxiv.org/abs/2305.18189v1>) [cs.CL (<https://arxiv.org/archive/cs.CL>)].

54. Hammond G (27 December 2023). "Big Tech is spending more than VC firms on AI startups" (<https://arstechnica.com/ai/2023/12/big-tech-is-spending-more-than-vc-firms-on-ai-startups/>). *Ars Technica*. Archived (<https://web.archive.org/web/20240110195706/https://arstechnica.com/ai/2023/12/big-tech-is-spending-more-than-vc-firms-on-ai-startups/>) from the original on Jan 10, 2024.
55. Wong M (24 October 2023). "The Future of AI Is GOMA" (<https://www.theatlantic.com/technology/archive/2023/10/big-ai-silicon-valley-dominance/675752/>). *The Atlantic*. Archived (<https://web.archive.org/web/20240105020744/https://www.theatlantic.com/technology/archive/2023/10/big-ai-silicon-valley-dominance/675752/>) from the original on Jan 5, 2024.
56. "Big tech and the pursuit of AI dominance" (<https://www.economist.com/business/2023/03/26/big-tech-and-the-pursuit-of-ai-dominance>). *The Economist*. Mar 26, 2023. Archived (<https://web.archive.org/web/20231229021351/https://www.economist.com/business/2023/03/26/big-tech-and-the-pursuit-of-ai-dominance>) from the original on Dec 29, 2023.
57. Fung B (19 December 2023). "Where the battle to dominate AI may be won" (<https://www.cnn.com/2023/12/19/tech/cloud-competition-and-ai/index.html>). *CNN Business*. Archived (<https://web.archive.org/web/20240113053332/https://www.cnn.com/2023/12/19/tech/cloud-competition-and-ai/index.html>) from the original on Jan 13, 2024.
58. Metz C (5 July 2023). "In the Age of A.I., Tech's Little Guys Need Big Friends" (<https://www.nytimes.com/2023/07/05/business/artificial-intelligence-power-data-centers.html>). *The New York Times*.
59. Open Source AI. ([http://www.ssec.wisc.edu/~billh/g/hibbard\\_agi\\_workshop.pdf](http://www.ssec.wisc.edu/~billh/g/hibbard_agi_workshop.pdf)) Archived ([https://web.archive.org/web/20160304054930/http://www.ssec.wisc.edu/~billh/g/hibbard\\_agi\\_workshop.pdf](https://web.archive.org/web/20160304054930/http://www.ssec.wisc.edu/~billh/g/hibbard_agi_workshop.pdf)) 2016-03-04 at the Wayback Machine Bill Hibbard. 2008 proceedings (<https://agi-conf.org/2008/papers/>) of the First Conference on Artificial General Intelligence, eds. Pei Wang, Ben Goertzel, and Stan Franklin.
60. Stewart A, Melton M. "Hugging Face CEO says he's focused on building a 'sustainable model' for the \$4.5 billion open-source-AI startup" (<https://www.businessinsider.com/hugging-face-open-source-ai-approach-2023-12>). *Business Insider*. Retrieved 2024-04-07.
61. "The open-source AI boom is built on Big Tech's handouts. How long will it last?" (<https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-mea/>). *MIT Technology Review*. Retrieved 2024-04-07.
62. Yao D (February 21, 2024). "Google Unveils Open Source Models to Rival Meta, Mistral" (<https://aibusiness.com/nlp/google-unveils-open-source-models-to-compete-against-meta>). *AI Business*.
63. 7001-2021 - IEEE Standard for Transparency of Autonomous Systems (<https://ieeexplore.ieee.org/document/9726144>). IEEE. 4 March 2022. pp. 1–54. doi:10.1109/IEEESTD.2022.9726144 (<https://doi.org/10.1109%2FIEEESTD.2022.9726144>). ISBN 978-1-5044-8311-7. S2CID 252589405 (<https://api.semanticscholar.org/CorpusID:252589405>). Retrieved 9 July 2023..
64. Kamila MK, Jasrotia SS (2023-01-01). "Ethical issues in the development of artificial intelligence: recognizing the risks" (<https://doi.org/10.1108/IJOES-05-2023-0107>). *International Journal of Ethics and Systems*. doi:10.1108/IJOES-05-2023-0107 (<https://doi.org/10.1108%2FIJOES-05-2023-0107>). ISSN 2514-9369 (<https://search.worldcat.org/issn/2514-9369>). S2CID 259614124 (<https://api.semanticscholar.org/CorpusID:259614124>).
65. Thurm S (July 13, 2018). "Microsoft Calls For Federal Regulation of Facial Recognition" (<https://www.wired.com/story/microsoft-calls-for-federal-regulation-of-facial-recognition/>). *Wired*. Archived (<https://web.archive.org/web/20190509231338/https://www.wired.com/story/microsoft-calls-for-federal-regulation-of-facial-recognition/>) from the original on May 9, 2019. Retrieved January 10, 2019.
66. Piper K (2024-02-02). "Should we make our most powerful AI models open source to all?" (<https://www.vox.com/future-perfect/2024/2/2/24058484/open-source-artificial-intelligence-ai-risk-meta-llama-2-chatgpt-openai-deepfake>). *Vox*. Retrieved 2024-04-07.

67. Vincent J (2023-03-15). "OpenAI co-founder on company's past approach to openly sharing research: "We were wrong" " (<https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview>). *The Verge*. Retrieved 2024-04-07.
68. Inside The Mind Of A.I. (<https://think.kera.org/2017/12/05/inside-the-mind-of-a-i/>) Archived (<https://web.archive.org/web/20210810003331/https://think.kera.org/2017/12/05/inside-the-mind-of-a-i/>) 2021-08-10 at the Wayback Machine - Cliff Kuang interview
69. Bunn J (2020-04-13). "Working in contexts for which transparency is important: A recordkeeping view of explainable artificial intelligence (XAI)" (<https://www.emerald.com/insight/content/doi/10.1108/RMJ-08-2019-0038/full/html>). *Records Management Journal*. **30** (2): 143–153. doi:10.1108/RMJ-08-2019-0038 (<https://doi.org/10.1108%2FRMJ-08-2019-0038>). ISSN 0956-5698 (<https://search.worldcat.org/issn/0956-5698>). S2CID 219079717 (<https://api.semanticscholar.org/CorpusID:219079717>).
70. Li F, Ruijs N, Lu Y (2022-12-31). "Ethics & AI: A Systematic Review on Ethical Concerns and Related Strategies for Designing with AI in Healthcare" (<https://doi.org/10.3390%2Fai401003>). *AI*. **4** (1): 28–53. doi:10.3390/ai4010003 (<https://doi.org/10.3390%2Fai4010003>). ISSN 2673-2688 (<https://search.worldcat.org/issn/2673-2688>).
71. Howard A (29 July 2019). "The Regulation of AI – Should Organizations Be Worried? | Ayanna Howard" (<https://sloanreview.mit.edu/article/the-regulation-of-ai-should-organizations-be-worried/>). *MIT Sloan Management Review*. Archived (<https://web.archive.org/web/20190814134545/https://sloanreview.mit.edu/article/the-regulation-of-ai-should-organizations-be-worried/>) from the original on 2019-08-14. Retrieved 2019-08-14.
72. "Trust in artificial intelligence - A five country study" (<https://assets.kpmg.com/content/dam/kpmg/au/pdf/2021/trust-in-ai-multiple-countries.pdf>) (PDF). KPMG. March 2021.
73. Bastin R, Wantz G (June 2017). "The General Data Protection Regulation Cross-industry innovation" (<https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/technology/lu-general-data-protection-regulation-cross-industry-innovation-062017.pdf>) (PDF). *Inside magazine*. Deloitte. Archived (<https://web.archive.org/web/20190110183405/https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/technology/lu-general-data-protection-regulation-cross-industry-innovation-062017.pdf>) (PDF) from the original on 2019-01-10. Retrieved 2019-01-10.
74. "UN artificial intelligence summit aims to tackle poverty, humanity's 'grand challenges'" (<https://news.un.org/en/story/2017/06/558962/un-artificial-intelligence-summit-aims-tackle-poverty-humanitys-grand>). UN News. 2017-06-07. Archived (<https://web.archive.org/web/20190726084819/https://news.un.org/en/story/2017/06/558962/un-artificial-intelligence-summit-aims-tackle-poverty-humanitys-grand>) from the original on 2019-07-26. Retrieved 2019-07-26.
75. "Artificial intelligence – Organisation for Economic Co-operation and Development" (<http://www.oecd.org/going-digital/ai/>). www.oecd.org. Archived (<https://web.archive.org/web/20190722124751/http://www.oecd.org/going-digital/ai/>) from the original on 2019-07-22. Retrieved 2019-07-26.
76. Anonymous (2018-06-14). "The European AI Alliance" (<https://ec.europa.eu/digital-single-market/en/european-ai-alliance>). *Digital Single Market – European Commission*. Archived (<https://web.archive.org/web/20190801011543/https://ec.europa.eu/digital-single-market/en/european-ai-alliance>) from the original on 2019-08-01. Retrieved 2019-07-26.
77. European Commission High-Level Expert Group on AI (2019-06-26). "Policy and investment recommendations for trustworthy Artificial Intelligence" (<https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence>). *Shaping Europe's digital future – European Commission*. Archived (<https://web.archive.org/web/20200226023934/https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence>) from the original on 2020-02-26. Retrieved 2020-03-16.

78. Fukuda-Parr S, Gibbons E (July 2021). "Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines" (<https://doi.org/10.1111%2F1758-5899.12965>). *Global Policy*. **12** (S6): 32–44. doi:[10.1111/1758-5899.12965](https://doi.org/10.1111%2F1758-5899.12965) (<https://doi.org/10.1111%2F1758-5899.12965>). ISSN 1758-5880 (<https://search.worldcat.org/issn/1758-5880>).
79. "EU Tech Policy Brief: July 2019 Recap" (<https://cdt.org/blog/eu-tech-policy-brief-july-2019-recap/>). Center for Democracy & Technology. 2 August 2019. Archived (<https://web.archive.org/web/20190809194057/https://cdt.org/blog/eu-tech-policy-brief-july-2019-recap/>) from the original on 2019-08-09. Retrieved 2019-08-09.
80. Curtis C, Gillespie N, Lockey S (2022-05-24). "AI-deploying organizations are key to addressing 'perfect storm' of AI risks" (<https://doi.org/10.1007/s43681-022-00163-7>). *AI and Ethics*. **3** (1): 145–153. doi:[10.1007/s43681-022-00163-7](https://doi.org/10.1007/s43681-022-00163-7) (<https://doi.org/10.1007%2Fs43681-022-00163-7>). ISSN 2730-5961 (<https://search.worldcat.org/issn/2730-5961>). PMC 9127285 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9127285>). PMID 35634256 (<https://pubmed.ncbi.nlm.nih.gov/35634256>). Archived (<https://web.archive.org/web/20230315194711/https://link.springer.com/article/10.1007/s43681-022-00163-7>) from the original on 2023-03-15. Retrieved 2022-05-29.
81. "Why the world needs a Bill of Rights on AI" (<https://www.ft.com/content/17ca620c-4d76-4a2f-829a-27d8552ce719>). *Financial Times*. 2021-10-18. Retrieved 2023-03-19.
82. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K (March 2019). "Artificial intelligence, bias and clinical safety" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6560460>). *BMJ Quality & Safety*. **28** (3): 231–237. doi:[10.1136/bmjqqs-2018-008370](https://doi.org/10.1136/bmjqqs-2018-008370) (<https://doi.org/10.1136%2Fbmjqqs-2018-008370>). ISSN 2044-5415 (<https://search.worldcat.org/issn/2044-5415>). PMC 6560460 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6560460>). PMID 30636200 (<https://pubmed.ncbi.nlm.nih.gov/30636200>).
83. Evans W (2015). "Posthuman Rights: Dimensions of Transhuman Worlds" ([https://doi.org/10.5209%2Frev\\_TK.2015.v12.n2.49072](https://doi.org/10.5209%2Frev_TK.2015.v12.n2.49072)). *Teknokultura*. **12** (2). doi:[10.5209/rev\\_TK.2015.v12.n2.49072](https://doi.org/10.5209/rev_TK.2015.v12.n2.49072) ([https://doi.org/10.5209%2Frev\\_TK.2015.v12.n2.49072](https://doi.org/10.5209%2Frev_TK.2015.v12.n2.49072)).
84. Sheliazhenko Y (2017). "Artificial Personal Autonomy and Concept of Robot Rights" (<http://cyberleninka.ru/article/n/artificial-personal-autonomy-and-concept-of-robot-rights>). *European Journal of Law and Political Sciences*: 17–21. doi:[10.20534/EJLPS-17-1-17-21](https://doi.org/10.20534/EJLPS-17-1-17-21) (<https://doi.org/10.20534%2FEJLPS-17-1-17-21>). Archived (<https://web.archive.org/web/20180714111141/https://cyberleninka.ru/article/n/artificial-personal-autonomy-and-concept-of-robot-rights>) from the original on 14 July 2018. Retrieved 10 May 2017.
85. Doomen J (2023). "The artificial intelligence entity as a legal person" (<https://doi.org/10.1080%2F13600834.2023.2196827>). *Information & Communications Technology Law*. **32** (3): 277–278. doi:[10.1080/13600834.2023.2196827](https://doi.org/10.1080/13600834.2023.2196827) (<https://doi.org/10.1080%2F13600834.2023.2196827>). hdl:[1820%2Fc29a3daa-9e36-4640-85d3-d0ffdd18a62c](https://hdl.handle.net/1820%2Fc29a3daa-9e36-4640-85d3-d0ffdd18a62c) (<https://hdl.handle.net/1820%2Fc29a3daa-9e36-4640-85d3-d0ffdd18a62c>).
86. "Robots could demand legal rights" (<http://news.bbc.co.uk/2/hi/technology/6200005.stm>). BBC News. December 21, 2006. Archived (<https://web.archive.org/web/20191015042628/http://news.bbc.co.uk/2/hi/technology/6200005.stm>) from the original on October 15, 2019. Retrieved January 3, 2010.
87. Henderson M (April 24, 2007). "Human rights for robots? We're getting carried away" (<http://www.timesonline.co.uk/tol/news/uk/science/article1695546.ece>). *The Times Online*. The Times of London. Archived (<https://web.archive.org/web/20080517022444/http://www.timesonline.co.uk/tol/news/uk/science/article1695546.ece>) from the original on May 17, 2008. Retrieved May 2, 2010.
88. "Saudi Arabia bestows citizenship on a robot named Sophia" (<https://techcrunch.com/2017/10/26/saudi-arabia-robot-citizen-sophia/>). 26 October 2017. Archived (<https://web.archive.org/web/20171027023101/https://techcrunch.com/2017/10/26/saudi-arabia-robot-citizen-sophia/>) from the original on 2017-10-27. Retrieved 2017-10-27.

89. Vincent J (30 October 2017). "Pretending to give a robot citizenship helps no one" (<https://www.theverge.com/2017/10/30/16552006/robot-rights-citizenship-saudi-arabia-sophia>). *The Verge*. Archived (<https://web.archive.org/web/20190803144659/https://www.theverge.com/2017/10/30/16552006/robot-rights-citizenship-saudi-arabia-sophia>) from the original on 3 August 2019. Retrieved 10 January 2019.
90. Wilks, Yorick, ed. (2010). *Close engagements with artificial companions: key social, psychological, ethical and design issues*. Amsterdam: John Benjamins Pub. Co. ISBN 978-90-272-4994-4. OCLC 642206106 (<https://search.worldcat.org/oclc/642206106>).
91. Macrae C (September 2022). "Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety, and Sociotechnical Sources of Risk" (<https://onlinelibrary.wiley.com/doi/10.1111/risa.13850>). *Risk Analysis*. **42** (9): 1999–2025.  
Bibcode:2022RiskA..42.1999M (<https://ui.adsabs.harvard.edu/abs/2022RiskA..42.1999M>). doi:[10.1111/risa.13850](https://doi.org/10.1111/risa.13850) (<https://doi.org/10.1111%2Frisa.13850>). ISSN 0272-4332 (<https://seach.worldcat.org/issn/0272-4332>). PMID 34814229 (<https://pubmed.ncbi.nlm.nih.gov/34814229>).
92. Agarwal A, Edelman S (2020). "Functionally effective conscious AI without suffering". *Journal of Artificial Intelligence and Consciousness*. **7**: 39–50. arXiv:2002.05652 (<https://arxiv.org/abs/2002.05652>). doi:[10.1142/S2705078520300030](https://doi.org/10.1142/S2705078520300030) (<https://doi.org/10.1142%2FS2705078520300030>). S2CID 211096533 (<https://api.semanticscholar.org/CorpusID:211096533>).
93. Thomas Metzinger (February 2021). "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology" (<https://doi.org/10.1142%2FS270507852150003X>). *Journal of Artificial Intelligence and Consciousness*. **8**: 43–66. doi:[10.1142/S270507852150003X](https://doi.org/10.1142/S270507852150003X) (<https://doi.org/10.1142%2FS270507852150003X>). S2CID 233176465 (<https://api.semanticscholar.org/CorpusID:233176465>).
94. Chalmers D (March 2023). "Could a Large Language Model be Conscious?". arXiv:2303.07103v1 (<https://arxiv.org/abs/2303.07103v1>) [Science Computer Science (<https://arxiv.org/archive/Computer>)].
95. Birch J (2017-01-01). "Animal sentience and the precautionary principle" (<https://www.wellbeingintlstudiesrepository.org/animsent/vol2/iss16/1>). *Animal Sentience*. **2** (16). doi:[10.51291/2377-7478.1200](https://doi.org/10.51291/2377-7478.1200) (<https://doi.org/10.51291%2F2377-7478.1200>). ISSN 2377-7478 (<https://search.worldcat.org/issn/2377-7478>).
96. ▪ Weizenbaum J (1976). *Computer Power and Human Reason*. San Francisco: W.H. Freeman & Company. ISBN 978-0-7167-0464-5.  
▪ McCorduck P (2004), *Machines Who Think* (2nd ed.), Natick, MA: A. K. Peters, Ltd., ISBN 1-56881-205-1, pp. 132–144
97. Joseph Weizenbaum, quoted in McCorduck 2004, pp. 356, 374–376
98. Kaplan A, Haenlein M (January 2019). "Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence". *Business Horizons*. **62** (1): 15–25. doi:[10.1016/j.bushor.2018.08.004](https://doi.org/10.1016/j.bushor.2018.08.004) (<https://doi.org/10.1016%2Fj.bushor.2018.08.004>). S2CID 158433736 (<https://api.semanticscholar.org/CorpusID:158433736>).
99. Hibbard B (17 November 2015). "Ethical Artificial Intelligence". arXiv:1411.1373 (<https://arxiv.org/abs/1411.1373>) [cs.AI (<https://arxiv.org/archive/cs.AI>)].
100. Davies A (29 February 2016). "Google's Self-Driving Car Caused Its First Crash" (<https://www.wired.com/2016/02/googles-self-driving-car-may-caused-first-crash/>). *Wired*. Archived (<https://web.archive.org/web/20190707212719/https://www.wired.com/2016/02/googles-self-driving-car-may-caused-first-crash/>) from the original on 7 July 2019. Retrieved 26 July 2019.

101. Levin S, Wong JC (19 March 2018). "Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian" (<https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe>). *The Guardian*. Archived (<https://web.archive.org/web/20190726084818/https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe>) from the original on 26 July 2019. Retrieved 26 July 2019.
102. "Who is responsible when a self-driving car has an accident?" (<https://futurism.com/who-responsible-when-self-driving-car-accident>). *Futurism*. 30 January 2018. Archived (<https://web.archive.org/web/20190726084819/https://futurism.com/who-responsible-when-self-driving-car-accident>) from the original on 2019-07-26. Retrieved 2019-07-26.
103. "Autonomous Car Crashes: Who – or What – Is to Blame?" (<https://knowledge.wharton.upenn.edu/article/automated-car-accidents/>). *Knowledge@Wharton*. Law and Public Policy. Radio Business North America Podcasts. Archived (<https://web.archive.org/web/20190726084820/https://knowledge.wharton.upenn.edu/article/automated-car-accidents/>) from the original on 2019-07-26. Retrieved 2019-07-26.
104. Delbridge E. "Driverless Cars Gone Wild" (<https://www.thebalance.com/driverless-car-accidents-4171792>). *The Balance*. Archived (<https://web.archive.org/web/20190529020717/https://www.thebalance.com/driverless-car-accidents-4171792>) from the original on 2019-05-29. Retrieved 2019-05-29.
105. Stilgoe J (2020), "Who Killed Elaine Herzberg?" ([http://link.springer.com/10.1007/978-3-030-32320-2\\_1](http://link.springer.com/10.1007/978-3-030-32320-2_1)), *Who's Driving Innovation?*, Cham: Springer International Publishing, pp. 1–6, doi:[10.1007/978-3-030-32320-2\\_1](https://doi.org/10.1007/978-3-030-32320-2_1) ([https://doi.org/10.1007/978-3-030-32320-2\\_1](https://doi.org/10.1007/978-3-030-32320-2_1)), ISBN 978-3-030-32319-6, S2CID 214359377 (<https://api.semanticscholar.org/CorpusID:214359377>), archived ([https://web.archive.org/web/20210318060722/https://link.springer.com/chapter/10.1007/978-3-030-32320-2\\_1](https://web.archive.org/web/20210318060722/https://link.springer.com/chapter/10.1007/978-3-030-32320-2_1)) from the original on 2021-03-18, retrieved 2020-11-11
106. Maxmen A (October 2018). "Self-driving car dilemmas reveal that moral choices are not universal" (<https://doi.org/10.1038%2Fd41586-018-07135-0>). *Nature*. **562** (7728): 469–470. Bibcode:2018Natur.562..469M (<https://ui.adsabs.harvard.edu/abs/2018Natur.562..469M>). doi:[10.1038/d41586-018-07135-0](https://doi.org/10.1038/d41586-018-07135-0) (<https://doi.org/10.1038/d41586-018-07135-0>). PMID 30356197 (<https://pubmed.ncbi.nlm.nih.gov/30356197>).
107. "Regulations for driverless cars" (<https://www.gov.uk/government/publications/driverless-cars-in-the-uk-a-regulatory-review>). GOV.UK. Archived (<https://web.archive.org/web/20190726084816/https://www.gov.uk/government/publications/driverless-cars-in-the-uk-a-regulatory-review>) from the original on 2019-07-26. Retrieved 2019-07-26.
108. "Automated Driving: Legislative and Regulatory Action – CyberWiki" ([https://web.archive.org/web/20190726084828/https://cyberlaw.stanford.edu/wiki/index.php/Automated\\_Driving:\\_Legislative\\_and\\_Regulatory\\_Action](https://web.archive.org/web/20190726084828/https://cyberlaw.stanford.edu/wiki/index.php/Automated_Driving:_Legislative_and_Regulatory_Action)). *cyberlaw.stanford.edu*. Archived from the original ([https://web.archive.org/web/20190726084828/https://cyberlaw.stanford.edu/wiki/index.php/Automated\\_Driving:\\_Legislative\\_and\\_Regulatory\\_Action](https://web.archive.org/web/20190726084828/https://cyberlaw.stanford.edu/wiki/index.php/Automated_Driving:_Legislative_and_Regulatory_Action)) on 2019-07-26. Retrieved 2019-07-26.
109. "Autonomous Vehicles | Self-Driving Vehicles Enacted Legislation" (<http://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx>). *www.ncsl.org*. Archived (<https://web.archive.org/web/20190726165225/http://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx>) from the original on 2019-07-26. Retrieved 2019-07-26.
110. Etzioni A, Etzioni O (2017-12-01). "Incorporating Ethics into Artificial Intelligence" (<https://doi.org/10.1007/s10892-017-9252-2>). *The Journal of Ethics*. **21** (4): 403–418. doi:[10.1007/s10892-017-9252-2](https://doi.org/10.1007/s10892-017-9252-2) (<https://doi.org/10.1007/s10892-017-9252-2>). ISSN 1572-8609 (<https://search.worldcat.org/issn/1572-8609>). S2CID 254644745 (<https://api.semanticscholar.org/CorpusID:254644745>).
111. Call for debate on killer robots (<http://news.bbc.co.uk/2/hi/technology/8182003.stm>) Archived (<https://web.archive.org/web/20090807005005/http://news.bbc.co.uk/2/hi/technology/8182003.stm>) 2009-08-07 at the Wayback Machine, By Jason Palmer, Science and technology reporter, BBC News, 8/3/09.

112. Science New Navy-funded Report Warns of War Robots Going "Terminator" (<http://www.dailytech.com/New%20Navyfunded%20Report%20Warns%20of%20War%20Robots%20Going%20Terminator/article14298.htm>) Archived (<https://web.archive.org/web/20090728101106/http://www.dailytech.com/New%20Navyfunded%20Report%20Warns%20of%20War%20Robots%20Going%20Terminator/article14298.htm>) 2009-07-28 at the Wayback Machine, by Jason Mick (Blog), dailytech.com, February 17, 2009.
113. Navy report warns of robot uprising, suggests a strong moral compass (<https://www.engadget.com/2009/02/18/navy-report-warns-of-robot-uprising-suggests-a-strong-moral-com/>) Archived (<https://web.archive.org/web/20110604145633/http://www.engadget.com/2009/02/18/navy-report-warns-of-robot-uprising-suggests-a-strong-moral-com/>) 2011-06-04 at the Wayback Machine, by Joseph L. Flatley engadget.com, Feb 18th 2009.
114. AAAI Presidential Panel on Long-Term AI Futures 2008–2009 Study ([http://research.microsoft.com/en-us/um/people/horvitz/AAAI\\_Presidential\\_Panel\\_2008-2009.htm](http://research.microsoft.com/en-us/um/people/horvitz/AAAI_Presidential_Panel_2008-2009.htm)) Archived ([https://web.archive.org/web/20090828214741/http://research.microsoft.com/en-us/um/people/horvitz/AAAI\\_Presidential\\_Panel\\_2008-2009.htm](https://web.archive.org/web/20090828214741/http://research.microsoft.com/en-us/um/people/horvitz/AAAI_Presidential_Panel_2008-2009.htm)) 2009-08-28 at the Wayback Machine, Association for the Advancement of Artificial Intelligence, Accessed 7/26/09.
115. United States. Defense Innovation Board. *AI principles: recommendations on the ethical use of artificial intelligence by the Department of Defense*. OCLC 1126650738 (<https://search.worldcat.org/oclc/1126650738>).
116. New Navy-funded Report Warns of War Robots Going "Terminator" (<http://www.dailytech.com/New%20Navyfunded%20Report%20Warns%20of%20War%20Robots%20Going%20Terminator/article14298.htm>) Archived (<https://web.archive.org/web/20090728101106/http://www.dailytech.com/New%20Navyfunded%20Report%20Warns%20of%20War%20Robots%20Going%20Terminator/article14298.htm>) 2009-07-28 at the Wayback Machine, by Jason Mick (Blog), dailytech.com, February 17, 2009.
117. Umbrello S, Torres P, De Bellis AF (March 2020). "The future of war: could lethal autonomous weapons make conflict more ethical?" (<http://link.springer.com/10.1007/s00146-019-00879-x>) (<https://doi.org/10.1007%2Fs00146-019-00879-x>). *AI & Society*. **35** (1): 273–282. doi:[10.1007/s00146-019-00879-x](https://doi.org/10.1007/s00146-019-00879-x) (<https://hdl.handle.net/2318/1699364>). ISSN 0951-5666 (<https://search.worldcat.org/issn/0951-5666>). S2CID 59606353 (<https://api.semanticscholar.org/CorpusID:59606353>). Archived (<https://archive.today/202101105020836/https://link.springer.com/article/10.1007/s00146-019-00879-x>) from the original on 2021-01-05. Retrieved 2020-11-11.
118. Hellström T (June 2013). "On the moral responsibility of military robots". *Ethics and Information Technology*. **15** (2): 99–107. doi:[10.1007/s10676-012-9301-2](https://doi.org/10.1007/s10676-012-9301-2) (<https://doi.org/10.1007%2Fs10676-012-9301-2>). S2CID 15205810 (<https://api.semanticscholar.org/CorpusID:15205810>). ProQuest 1372020233 (<https://search.proquest.com/docview/1372020233>).
119. Mitra A (5 April 2018). "We can train AI to identify good and evil, and then use it to teach us morality" (<https://qz.com/1244055/we-can-train-ai-to-identify-good-and-evil-and-then-use-it-to-teach-us-morality/>). Quartz. Archived (<https://web.archive.org/web/20190726085248/http://qz.com/1244055/we-can-train-ai-to-identify-good-and-evil-and-then-use-it-to-teach-us-morality/>) from the original on 2019-07-26. Retrieved 2019-07-26.
120. Dominguez G (23 August 2022). "South Korea developing new stealthy drones to support combat aircraft" (<https://www.japantimes.co.jp/news/2022/08/23/asia-pacific/south-korea-stealth-drones-development/>). *The Japan Times*. Retrieved 14 June 2023.
121. "AI Principles" (<https://futureoflife.org/ai-principles/>). Future of Life Institute. 11 August 2017. Archived (<https://web.archive.org/web/20171211171044/https://futureoflife.org/ai-principles/>) from the original on 2017-12-11. Retrieved 2019-07-26.

122. Zach Musgrave and Bryan W. Roberts (2015-08-14). "Why Artificial Intelligence Can Too Easily Be Weaponized – The Atlantic" (<https://www.theatlantic.com/technology/archive/2015/08/humans-not-robots-are-the-real-reason-artificial-intelligence-is-scary/400994/>). *The Atlantic*. Archived (<https://web.archive.org/web/20170411140722/https://www.theatlantic.com/technology/archive/2015/08/humans-not-robots-are-the-real-reason-artificial-intelligence-is-scary/400994/>) from the original on 2017-04-11. Retrieved 2017-03-06.
123. Cat Zakrzewski (2015-07-27). "Musk, Hawking Warn of Artificial Intelligence Weapons" (<http://blogs.wsj.com/digits/2015/07/27/musk-hawking-warn-of-artificial-intelligence-weapons/>). *WSJ*. Archived (<https://web.archive.org/web/20150728173944/http://blogs.wsj.com/digits/2015/07/27/musk-hawking-warn-of-artificial-intelligence-weapons/>) from the original on 2015-07-28. Retrieved 2017-08-04.
124. "Potential Risks from Advanced Artificial Intelligence" (<https://www.openphilanthropy.org/research/potential-risks-from-advanced-artificial-intelligence/>). *Open Philanthropy*. August 11, 2015. Retrieved 2024-04-07.
125. Bachulska A, Leonard M, Oertel J (2 July 2024). *The Idea of China: Chinese Thinkers on Power, Progress, and People* (<https://ecfr.eu/publication/idea-of-china/>) (EPUB). Berlin, Germany: European Council on Foreign Relations. ISBN 978-1-916682-42-9. Archived (<https://web.archive.org/web/20240717120845/https://ecfr.eu/publication/idea-of-china/>) from the original on 17 July 2024. Retrieved 22 July 2024.
126. Brandon Vigliarolo. "International military AI summit ends with 60-state pledge" ([https://www.theregister.com/2023/02/17/military\\_ai\\_summit/](https://www.theregister.com/2023/02/17/military_ai_summit/)). [www.theregister.com](https://www.theregister.com). Retrieved 2023-02-17.
127. Markoff J (25 July 2009). "Scientists Worry Machines May Outsmart Man" (<https://www.nytimes.com/2009/07/26/science/26robot.html>). *The New York Times*. Archived (<https://web.archive.org/web/20170225202201/http://www.nytimes.com/2009/07/26/science/26robot.html>) from the original on 25 February 2017. Retrieved 24 February 2017.
128. Muehlhauser, Luke, and Louie Helm. 2012. "Intelligence Explosion and Machine Ethics" (<https://intelligence.org/files/IE-ME.pdf>) Archived (<https://web.archive.org/web/20150507173028/http://intelligence.org/files/IE-ME.pdf>) 2015-05-07 at the Wayback Machine. In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. Berlin: Springer.
129. Bostrom, Nick. 2003. "Ethical Issues in Advanced Artificial Intelligence" (<http://www.nickbostrom.com/ethics/ai.html>) Archived (<https://web.archive.org/web/20181008090224/http://www.nickbostrom.com/ethics/ai.html>) 2018-10-08 at the Wayback Machine. In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, edited by Iva Smit and George E. Lasker, 12–17. Vol. 2. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.
130. Bostrom N (2017). *Superintelligence: paths, dangers, strategies*. Oxford, United Kingdom: Oxford University Press. ISBN 978-0-19-967811-2.
131. Umbrello S, Baum SD (2018-06-01). "Evaluating future nanotechnology: The net societal impacts of atomically precise manufacturing" (<http://www.sciencedirect.com/science/article/pii/S0016328717301908>). *Futures*. **100**: 63–73. doi:10.1016/j.futures.2018.04.007 (<https://doi.org/10.1016%2Fj.futures.2018.04.007>). hdl:2318/1685533 (<https://hdl.handle.net/2318%2F1685533>). ISSN 0016-3287 (<https://search.worldcat.org/issn/0016-3287>). S2CID 158503813 (<https://api.semanticscholar.org/CorpusID:158503813>). Archived (<https://web.archive.org/web/20190509222110/https://www.sciencedirect.com/science/article/pii/S0016328717301908>) from the original on 2019-05-09. Retrieved 2020-11-29.
132. Yudkowsky, Eliezer. 2011. "Complex Value Systems in Friendly AI" (<https://intelligence.org/files/ComplexValues.pdf>) Archived (<https://web.archive.org/web/20150929212318/http://intelligence.org/files/ComplexValues.pdf>) 2015-09-29 at the Wayback Machine. In Schmidhuber, Thórisson, and Looks 2011, 388–393.

133. Russell S (October 8, 2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. United States: Viking. ISBN 978-0-525-55861-3. OCLC 1083694322 (<https://search.worldcat.org/oclc/1083694322>).
134. Yampolskiy RV (2020-03-01). "Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent" (<https://www.worldscientific.com/doi/abs/10.1142/S2705078520500034>). *Journal of Artificial Intelligence and Consciousness*. **07** (1): 109–118. doi:10.1142/S2705078520500034 (<https://doi.org/10.1142%2FS2705078520500034>). ISSN 2705-0785 (<https://search.worldcat.org/issn/2705-0785>). S2CID 218916769 (<https://api.semanticscholar.org/CorpusID:218916769>). Archived (<https://web.archive.org/web/20210318060657/https://www.worldscientific.com/doi/abs/10.1142/S2705078520500034>) from the original on 2021-03-18. Retrieved 2020-11-29.
135. Wallach W, Vallor S (2020-09-17), "Moral Machines: From Value Alignment to Embodied Virtue" (<https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190905033.001.0001/oso-9780190905033-chapter-14>), *Ethics of Artificial Intelligence*, Oxford University Press, pp. 383–412, doi:10.1093/oso/9780190905033.003.0014 (<https://doi.org/10.1093%2Foso%2F9780190905033.003.0014>), ISBN 978-0-19-090503-3, archived (<https://web.archive.org/web/20201208114354/https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190905033.001.0001/oso-9780190905033-chapter-14>) from the original on 2020-12-08, retrieved 2020-11-29
136. Umbrello S (2019). "Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach" (<https://doi.org/10.3390%2Fbdcc3010005>). *Big Data and Cognitive Computing*. **3** (1): 5. doi:10.3390/bdcc3010005 (<https://doi.org/10.3390%2Fbdcc3010005>). hdl:2318/1685727 (<https://hdl.handle.net/2318%2F1685727>).
137. Floridi L, Cowls J, King TC, Taddeo M (2020). "How to Design AI for Social Good: Seven Essential Factors" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7286860>). *Science and Engineering Ethics*. **26** (3): 1771–1796. doi:10.1007/s11948-020-00213-5 (<https://doi.org/10.1007%2Fs11948-020-00213-5>). ISSN 1353-3452 (<https://search.worldcat.org/issn/1353-3452>). PMC 7286860 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7286860>). PMID 32246245 (<https://pubmed.ncbi.nlm.nih.gov/32246245>).
138. Fiegerman S (28 September 2016). "Facebook, Google, Amazon create group to ease AI concerns" (<https://money.cnn.com/2016/09/28/technology/partnership-on-ai/>). *CNNMoney*. Archived (<https://web.archive.org/web/20200917141730/https://money.cnn.com/2016/09/28/technology/partnership-on-ai/>) from the original on 17 September 2020. Retrieved 18 August 2020.
139. Slota SC, Fleischmann KR, Greenberg S, Verma N, Cummings B, Li L, Shenefiel C (2023). "Locating the work of artificial intelligence ethics" (<https://onlinelibrary.wiley.com/doi/10.1002/asi.24638>). *Journal of the Association for Information Science and Technology*. **74** (3): 311–322. doi:10.1002/asi.24638 (<https://doi.org/10.1002%2Fasi.24638>). ISSN 2330-1635 (<https://search.worldcat.org/issn/2330-1635>). S2CID 247342066 (<https://api.semanticscholar.org/CorpusID:247342066>).
140. "Ethics guidelines for trustworthy AI" (<https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>). *Shaping Europe's digital future – European Commission*. European Commission. 2019-04-08. Archived (<https://web.archive.org/web/20200220002342/https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>) from the original on 2020-02-20. Retrieved 2020-02-20.
141. "White Paper on Artificial Intelligence – a European approach to excellence and trust | Shaping Europe's digital future" (<https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-european-approach-excellence-and-trust>). 19 February 2020. Archived (<https://web.archive.org/web/20210306003222/https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-european-approach-excellence-and-trust>) from the original on 2021-03-06. Retrieved 2021-03-18.
142. "OECD AI Policy Observatory" (<https://www.oecd.ai/>). Archived (<https://web.archive.org/web/20210308171133/https://oecd.ai/>) from the original on 2021-03-08. Retrieved 2021-03-18.

143. *Recommendation on the Ethics of Artificial Intelligence* (<https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en>). UNESCO. 2021.
144. "UNESCO member states adopt first global agreement on AI ethics" (<https://www.helsinkitimes.fi/themes/themes/science-and-technology/20454-unesco-member-states-adopt-first-global-agreement-on-ai-ethics.html>). *Helsinki Times*. 2021-11-26. Retrieved 2023-04-26.
145. "The Obama Administration's Roadmap for AI Policy" (<https://hbr.org/2016/12/the-obama-administrations-roadmap-for-ai-policy>). *Harvard Business Review*. 2016-12-21. ISSN 0017-8012 (<https://search.worldcat.org/issn/0017-8012>). Archived (<https://web.archive.org/web/20210122003445/https://hbr.org/2016/12/the-obama-administrations-roadmap-for-ai-policy>) from the original on 2021-01-22. Retrieved 2021-03-16.
146. "Accelerating America's Leadership in Artificial Intelligence – The White House" (<https://trumpwhitehouse.archives.gov/articles/accelerating-americas-leadership-in-artificial-intelligence/>). *trumpwhitehouse.archives.gov*. Archived (<https://web.archive.org/web/20210225073748/https://trumpwhitehouse.archives.gov/articles/accelerating-americas-leadership-in-artificial-intelligence/>) from the original on 2021-02-25. Retrieved 2021-03-16.
147. "Request for Comments on a Draft Memorandum to the Heads of Executive Departments and Agencies, "Guidance for Regulation of Artificial Intelligence Applications" " (<https://www.federalregister.gov/documents/2020/01/13/2020-00261/request-for-comments-on-a-draft-memorandum-to-the-heads-of-executive-departments-and-agencies>). *Federal Register*. 2020-01-13. Archived (<https://web.archive.org/web/20201125060218/https://www.federalregister.gov/documents/2020/01/13/2020-00261/request-for-comments-on-a-draft-memorandum-to-the-heads-of-executive-departments-and-agencies>) from the original on 2020-11-25. Retrieved 2020-11-28.
148. "CCC Offers Draft 20-Year AI Roadmap; Seeks Comments" (<https://www.hpcwire.com/2019/05/14/ccc-offers-draft-20-year-ai-roadmap-seeks-comments/>). *HPCwire*. 2019-05-14. Archived (<https://web.archive.org/web/20210318060659/https://www.hpcwire.com/2019/05/14/ccc-offers-draft-20-year-ai-roadmap-seeks-comments/>) from the original on 2021-03-18. Retrieved 2019-07-22.
149. "Request Comments on Draft: A 20-Year Community Roadmap for AI Research in the US » CCC Blog" (<https://www.cccblog.org/2019/05/13/request-comments-on-draft-a-20-year-community-roadmap-for-ai-research-in-the-us/>). 13 May 2019. Archived (<https://web.archive.org/web/20190514193546/https://www.cccblog.org/2019/05/13/request-comments-on-draft-a-20-year-community-roadmap-for-ai-research-in-the-us/>) from the original on 2019-05-14. Retrieved 2019-07-22.
150. "Non-Human Party" (<https://nonhuman.party/>). 2021. Archived (<https://web.archive.org/web/20210920212940/https://nonhuman.party/>) from the original on 2021-09-20. Retrieved 2021-09-19.
151. (in Russian) Интеллектуальные правила (<https://www.kommersant.ru/doc/5089365>) Archived (<https://web.archive.org/web/20211230212952/https://www.kommersant.ru/doc/5089365>) 2021-12-30 at the Wayback Machine — *Kommersant*, 25.11.2021
152. Grace K, Salvatier J, Dafoe A, Zhang B, Evans O (2018-05-03). "When Will AI Exceed Human Performance? Evidence from AI Experts". *arXiv:1705.08807* (<https://arxiv.org/abs/1705.08807>) [cs.AI (<https://arxiv.org/archive/cs.AI>)].
153. "China wants to shape the global future of artificial intelligence" (<https://www.technologyreview.com/2018/03/16/144630/china-wants-to-shape-the-global-future-of-artificial-intelligence/>). *MIT Technology Review*. Archived (<https://web.archive.org/web/20201120052853/http://www.technologyreview.com/2018/03/16/144630/china-wants-to-shape-the-global-future-of-artificial-intelligence/>) from the original on 2020-11-20. Retrieved 2020-11-29.

154. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B (2018-12-01). "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6404626>). *Minds and Machines*. **28** (4): 689–707. doi:10.1007/s11023-018-9482-5 (<https://doi.org/10.1007%2Fs11023-018-9482-5>). ISSN 1572-8641 (<https://search.worldcat.org/issn/1572-8641>). PMC 6404626 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6404626>). PMID 30930541 (<https://pubmed.ncbi.nlm.nih.gov/30930541>).
155. "Joanna J. Bryson" (<https://www.wired.com/author/joanna-j-bryson/>). WIRED. Retrieved 13 January 2023.
156. "New Artificial Intelligence Research Institute Launches" (<https://engineering.nyu.edu/news/new-artificial-intelligence-research-institute-launches>). 2017-11-20. Archived ([https://engineering.nyu.edu/news/new-artificial-intelligence-research-institute-launches](https://web.archive.org/web/20200918091106/https://engineering.nyu.edu/news/new-artificial-intelligence-research-institute-launches)) from the original on 2020-09-18. Retrieved 2021-02-21.
157. James J. Hughes, LaGrandeur, Kevin, eds. (15 March 2017). *Surviving the machine age: intelligent technology and the transformation of human work* (<https://www.worldcat.org/oclc/976407024>). Cham, Switzerland: Palgrave Macmillan Cham. ISBN 978-3-319-51165-8. OCLC 976407024 (<https://search.worldcat.org/oclc/976407024>). Archived (<https://web.archive.org/web/20210318060659/https://www.worldcat.org/title/surviving-the-machine-age-intelligent-technology-and-the-transformation-of-human-work/oclc/976407024>) from the original on 18 March 2021. Retrieved 29 November 2020.
158. Danaher, John (2019). *Automation and utopia: human flourishing in a world without work* (<https://www.worldcat.org/oclc/1114334813>). Cambridge, Massachusetts: Harvard University Press. ISBN 978-0-674-24220-3. OCLC 1114334813 (<https://search.worldcat.org/oclc/1114334813>).
159. "TUM Institute for Ethics in Artificial Intelligence officially opened" (<https://www.tum.de/nc/en/about-tum/news/press-releases/details/35727/>). www.tum.de. Archived (<https://web.archive.org/web/20201210032545/https://www.tum.de/nc/en/about-tum/news/press-releases/details/35727/>) from the original on 2020-12-10. Retrieved 2020-11-29.
160. Communications PK (2019-01-25). "Harvard works to embed ethics in computer science curriculum" (<https://news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/>). *Harvard Gazette*. Retrieved 2023-04-06.
161. "Algorethics AI Library" (<https://algorethics.info/>). Algorethics. Retrieved 2024-08-28.
162. Lee J (2020-02-08). "When Bias Is Coded Into Our Technology" (<https://www.npr.org/sections/codeswitch/2020/02/08/770174171/when-bias-is-coded-into-our-technology>). NPR. Archived (<https://web.archive.org/web/20220326025113/https://www.npr.org/sections/codeswitch/2020/02/08/770174171/when-bias-is-coded-into-our-technology>) from the original on 2022-03-26. Retrieved 2021-12-22.
163. "How one conference embraced diversity" (<https://doi.org/10.1038%2Fd41586-018-07718-x>). *Nature*. **564** (7735): 161–162. 2018-12-12. doi:10.1038/d41586-018-07718-x (<https://doi.org/10.1038%2Fd41586-018-07718-x>). PMID 31123357 (<https://pubmed.ncbi.nlm.nih.gov/31123357>). S2CID 54481549 (<https://api.semanticscholar.org/CorpusID:54481549>).
164. Roose K (2020-12-30). "The 2020 Good Tech Awards" (<https://www.nytimes.com/2020/12/30/technology/2020-good-tech-awards.html>). *The New York Times*. ISSN 0362-4331 (<https://search.worldcat.org/issn/0362-4331>). Archived (<https://web.archive.org/web/20211221205543/https://www.nytimes.com/2020/12/30/technology/2020-good-tech-awards.html>) from the original on 2021-12-21. Retrieved 2021-12-21.
165. Lodge P (2014). "Leibniz's Mill Argument Against Mechanical Materialism Revisited" (<https://doi.org/10.3998%2Fergo.12405314.0001.003>). *Ergo, an Open Access Journal of Philosophy*. **1** (20201214). doi:10.3998/ergo.12405314.0001.003 (<https://doi.org/10.3998%2Fergo.12405314.0001.003>). hdl:2027/spo.12405314.0001.003 (<https://hdl.handle.net/2027/spo.12405314.0001.003>). ISSN 2330-4014 (<https://search.worldcat.org/issn/2330-4014>).

166. Bringsjord S, Govindarajulu NS (2020). "Artificial Intelligence" (<https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/>), in Zalta EN, Nodelman U (eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 ed.), Metaphysics Research Lab, Stanford University, retrieved 2023-12-08
167. Jr HC (1999-04-29). *Information Technology and the Productivity Paradox: Assessing the Value of Investing in IT* (<https://books.google.com/books?id=FzwnridL72IC&dq=digital+Much+of+his+work+was+then+spent+testing+the+boundaries+of+his+three+laws+to+see+whether+they+would+break+down,+or+where+they+would+create+paradoxical+or+unanticipated+behavior.&pg=PP13>). Oxford University Press. ISBN 978-0-19-802838-3.
168. Asimov I (2008). *I, Robot*. New York: Bantam. ISBN 978-0-553-38256-3.
169. Bryson J, Diamantis M, Grant T (September 2017). "Of, for, and by the people: the legal lacuna of synthetic persons" (<https://doi.org/10.1007%2Fs10506-017-9214-9>). *Artificial Intelligence and Law*. **25** (3): 273–291. doi:10.1007/s10506-017-9214-9 (<https://doi.org/10.1007%2Fs10506-017-9214-9>).
170. "Principles of robotics" (<https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/>). UK's EPSRC. September 2010. Archived (<https://web.archive.org/web/20180401004346/https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/>) from the original on 1 April 2018. Retrieved 10 January 2019.
171. Yudkowsky E (July 2004). "Why We Need Friendly AI" ([https://web.archive.org/web/20120524150856/http://www.asimovlaws.com/articles/archives/2004/07/why\\_we\\_need\\_fri\\_1.html](https://web.archive.org/web/20120524150856/http://www.asimovlaws.com/articles/archives/2004/07/why_we_need_fri_1.html)). *3 laws unsafe*. Archived from the original ([http://www.asimovlaws.com/articles/archives/2004/07/why\\_we\\_need\\_fri\\_1.html](http://www.asimovlaws.com/articles/archives/2004/07/why_we_need_fri_1.html)) on May 24, 2012.
172. Aleksander I (March 2017). "Partners of Humans: A Realistic Assessment of the Role of Robots in the Foreseeable Future" (<http://journals.sagepub.com/doi/10.1057/s41265-016-0032-4>). *Journal of Information Technology*. **32** (1): 1–9. doi:10.1057/s41265-016-0032-4 (<https://doi.org/10.1057%2Fs41265-016-0032-4>). ISSN 0268-3962 (<https://search.worldcat.org/isbn/0268-3962>). S2CID 5288506 (<https://api.semanticscholar.org/CorpusID:5288506>).
173. Evolving Robots Learn To Lie To Each Other (<http://www.popsci.com/scitech/article/2009-08/evolving-robots-learn-lie-hide-resources-each-other>) Archived (<https://web.archive.org/web/20090828105728/http://www.popsci.com/scitech/article/2009-08/evolving-robots-learn-lie-hide-resources-each-other>) 2009-08-28 at the Wayback Machine, Popular Science, August 18, 2009
174. Bassett C, Steinmueller E, Voss G. "Better Made Up: The Mutual Influence of Science Fiction and Innovation" (<https://www.nesta.org.uk/report/better-made-up-the-mutual-influence-of-science-fiction-and-innovation/>). Nesta. Retrieved 3 May 2024.
175. Velasco G (2020-05-04). "Science-Fiction: A Mirror for the Future of Humankind" (<https://revistaidees.cat/en/science-fiction-favors-engaging-debate-on-artificial-intelligence-and-ethics/>). IDEES. Retrieved 2023-12-08.
176. Hodges, A. (2014), *Alan Turing: The Enigma*, Vintage, London, p. 334
177. A. M. Turing (1936). "On computable numbers, with an application to the Entscheidungsproblem." in *Proceedings of the London Mathematical Society*, 2 s. vol. 42 (1936–1937), pp. 230–265.
178. "Love, Death & Robots season 2, episode 1 recap - "Automated Customer Service" " (<https://readysteadycut.com/2021/05/14/recap-love-death-and-robots-season-2-episode-1-automated-customer-service-netflix-series/>). Ready Steady Cut. 2021-05-14. Archived (<https://web.archive.org/web/20211221035251/https://readysteadycut.com/2021/05/14/recap-love-death-and-robots-season-2-episode-1-automated-customer-service-netflix-series/>) from the original on 2021-12-21. Retrieved 2021-12-21.

179. Cave, Stephen, Dihal, Kanta, Dillon, Sarah, eds. (14 February 2020). *AI narratives: a history of imaginative thinking about intelligent machines* (<https://www.worldcat.org/oclc/1143647559>) (First ed.). Oxford: Oxford University Press. ISBN 978-0-19-258604-9. OCLC 1143647559 (<https://search.worldcat.org/oclc/1143647559>). Archived (<https://web.archive.org/web/20210318060703/https://www.worldcat.org/title/ai-narratives-a-history-of-imaginative-thinking-about-intelligent-machines/oclc/1143647559>) from the original on 18 March 2021. Retrieved 11 November 2020.
180. Jerreat-Poole A (1 February 2020). "Sick, Slow, Cyborg: Crip Futurity in Mass Effect" (<http://gamestudies.org/2001/articles/jerreatpoole>). *Game Studies*. 20. ISSN 1604-7982 (<https://search.worldcat.org/issn/1604-7982>). Archived (<https://web.archive.org/web/20201209080256/http://gamestudies.org/2001/articles/jerreatpoole>) from the original on 9 December 2020. Retrieved 11 November 2020.
181. ""Detroit: Become Human" Will Challenge your Morals and your Humanity" (<https://coffeeordie.com/detroit-become-human-will-challenge-your-morals-and-your-humanity/>). *Coffee or Die Magazine*. 2018-08-06. Archived (<https://web.archive.org/web/20211209195312/https://coffeeordie.com/detroit-become-human-will-challenge-your-morals-and-your-humanity/>) from the original on 2021-12-09. Retrieved 2021-12-07.
182. Cerqui D, Warwick K (2008), "Re-Designing Humankind: The Rise of Cyborgs, a Desirable Goal?" ([http://link.springer.com/10.1007/978-1-4020-6591-0\\_14](http://link.springer.com/10.1007/978-1-4020-6591-0_14)), *Philosophy and Design*, Dordrecht: Springer Netherlands, pp. 185–195, doi:10.1007/978-1-4020-6591-0\_14 ([https://doi.org/10.1007%2F978-1-4020-6591-0\\_14](https://doi.org/10.1007%2F978-1-4020-6591-0_14)), ISBN 978-1-4020-6590-3, archived ([https://web.archive.org/web/20210318060701/https://link.springer.com/chapter/10.1007%2F978-1-4020-6591-0\\_14](https://web.archive.org/web/20210318060701/https://link.springer.com/chapter/10.1007%2F978-1-4020-6591-0_14)) from the original on 2021-03-18, retrieved 2020-11-11
183. Cave S, Dihal K (6 August 2020). "The Whiteness of AI" (<https://doi.org/10.1007%2Fs13347-020-00415-6>). *Philosophy & Technology*. 33 (4): 685–703. doi:10.1007/s13347-020-00415-6 (<https://doi.org/10.1007%2Fs13347-020-00415-6>). S2CID 225466550 (<https://api.semanticscholar.org/CorpusID:225466550>).

## External links

---

- Ethics of Artificial Intelligence (<http://www.iep.utm.edu/ethic-ai/>) at the *Internet Encyclopedia of Philosophy*
- Ethics of Artificial Intelligence and Robotics (<https://plato.stanford.edu/entries/ethics-ai/>) at the *Stanford Encyclopedia of Philosophy*
- Russell S, Hauert S, Altman R, Veloso M (May 2015). "Robotics: Ethics of artificial intelligence" (<https://doi.org/10.1038%2F521415a>). *Nature*. 521 (7553): 415–418. Bibcode:2015Natur.521..415. (<https://ui.adsabs.harvard.edu/abs/2015Natur.521..415.>). doi:10.1038/521415a (<https://doi.org/10.1038%2F521415a>). PMID 26017428 (<https://pubmed.ncbi.nlm.nih.gov/26017428>). S2CID 4452826 (<https://api.semanticscholar.org/CorpusID:4452826>).
- BBC News: Games to take on a life of their own (<http://news.bbc.co.uk/1/hi/sci/tech/1809769.stm>)
- Who's Afraid of Robots? (<http://www.dasboot.org/thorisson.htm>) Archived (<https://web.archive.org/web/20180322214031/http://www.dasboot.org/thorisson.htm>) 2018-03-22 at the Wayback Machine, an article on humanity's fear of artificial intelligence.
- A short history of computer ethics ([https://web.archive.org/web/20080418122849/http://www.southernct.edu/organizations/rccs/resources/research/introduction/bynum\\_shrt\\_hist.html](https://web.archive.org/web/20080418122849/http://www.southernct.edu/organizations/rccs/resources/research/introduction/bynum_shrt_hist.html))
- AI Ethics Guidelines Global Inventory (<https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/>) by Algorithmwatch (<https://algorithmwatch.org>)
- Hagendorff T (March 2020). "The Ethics of AI Ethics: An Evaluation of Guidelines" (<https://doi.org/10.1007%2Fs11023-020-09517-8>). *Minds and Machines*. 30 (1): 99–120.

arXiv:1903.03425 (<https://arxiv.org/abs/1903.03425>). doi:10.1107/s11023-020-09517-8 (<https://doi.org/10.1107%2Fs11023-020-09517-8>). S2CID 72940833 (<https://api.semanticscholar.org/CorpusID:72940833>).

- Sheludko, M. (December, 2023). Ethical Aspects of Artificial Intelligence: Challenges and Imperatives (<https://lasoфт.org/blog/ethical-aspects-of-artificial-intelligence-challenges-and-imperatives/>). Software Development Blog.
- Eisikovits N. "AI Is an Existential Threat--Just Not the Way You Think" (<https://www.scientificamerican.com/article/ai-is-an-existential-threat-just-not-the-way-you-think/>). *Scientific American*. Retrieved 2024-03-04.

---

Retrieved from "[https://en.wikipedia.org/w/index.php?title=Ethics\\_of\\_artificial\\_intelligence&oldid=1244027261](https://en.wikipedia.org/w/index.php?title=Ethics_of_artificial_intelligence&oldid=1244027261)"

■

**SpringerBriefs in  
Research and Innovation Governance**

Bernd Carsten Stahl · Doris Schroeder ·  
Rowena Rodrigues



# Ethics of Artificial Intelligence

## Case Studies and Options for Addressing Ethical Challenges

**OPEN ACCESS**

 Springer

# **SpringerBriefs in Research and Innovation Governance**

## **Editors-in-Chief**

Doris Schroeder, Centre for Professional Ethics, University of Central Lancashire,  
Preston, Lancashire, UK

Konstantinos Iatridis, School of Management, University of Bath, Bath, UK

SpringerBriefs in Research and Innovation Governance present concise summaries of cutting-edge research and practical applications across a wide spectrum of governance activities that are shaped and informed by, and in turn impact research and innovation, with fast turnaround time to publication. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic. Monographs of new material are considered for the SpringerBriefs in Research and Innovation Governance series. Typical topics might include: a timely report of state-of-the-art analytical techniques, a bridge between new research results, as published in journal articles and a contextual literature review, a snapshot of a hot or emerging topic, an in-depth case study or technical example, a presentation of core concepts that students and practitioners must understand in order to make independent contributions, best practices or protocols to be followed, a series of short case studies/debates highlighting a specific angle. SpringerBriefs in Research and Innovation Governance allow authors to present their ideas and readers to absorb them with minimal time investment. Both solicited and unsolicited manuscripts are considered for publication.

Bernd Carsten Stahl · Doris Schroeder ·  
Rowena Rodrigues

# Ethics of Artificial Intelligence

Case Studies and Options for Addressing  
Ethical Challenges



Springer

Bernd Carsten Stahl  
School of Computer Science  
University of Nottingham  
Nottingham, UK

Centre for Computing and Social  
Responsibility  
De Montfort University  
Leicester, UK

Rowena Rodrigues  
Trilateral Research  
London, UK

Doris Schroeder  
Centre for Professional Ethics  
University of Central Lancashire  
Preston, UK



ISSN 2452-0519                    ISSN 2452-0527 (electronic)  
SpringerBriefs in Research and Innovation Governance  
ISBN 978-3-031-17039-3            ISBN 978-3-031-17040-9 (eBook)  
<https://doi.org/10.1007/978-3-031-17040-9>

© The Author(s) 2023. This book is an open access publication.

**Open Access** This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Acknowledgements

This book draws on the work that the authors have done across a range of projects. The primary project that brought us together and demonstrated the need for cases describing ethical issues of AI and ways of addressing them was the EU-funded SHERPA project (2018–2021). All three authors were members of the project, which was led by Bernd Stahl. We would also like to acknowledge the contribution of the members of the SHERPA consortium during the project that informed and inspired the work in this book.

The three authors have been, or still are, active in many other projects that have informed this work directly or indirectly. These projects and the individuals who contributed to them deserve to be acknowledged. They include the EU-funded Human Brain Project, TechEthos and SIENNA, as well as the Responsible-Industry, CONSIDER, TRUST and ETICA projects.

We furthermore acknowledge the support of colleagues in our institutions and organisations, notably the Centre for Computing and Social Responsibility of De Montfort University, the Centre for Professional Ethics at the University of Central Lancashire and Trilateral Research Ltd.

We want to thank Paul Wise for his outstanding editing, Julie Cook for highly insightful comments on the first draft and Jayanthi Krishnamoorthi, Juliana Pitanguy and Toni Milevoj at Springer Nature for overseeing the publishing process very effectively. Thanks to Kostas Iatridis for excellent academic editorial support. We are indebted to Amanda Sharkey for permission to use her exceptionally well-drawn vignette on care robots and the elderly.

Last, but definitely not least, our thanks to three anonymous reviewers whose comments helped us greatly to make changes to the original book ideas.

This research received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under Grant Agreements No. 786641 (SHERPA) and No. 945539 (Human Brain Project SGA3).

# Contents

<b>1</b>	<b>The Ethics of Artificial Intelligence: An Introduction .....</b>	<b>1</b>
	References .....	6
<b>2</b>	<b>Unfair and Illegal Discrimination .....</b>	<b>9</b>
2.1	Introduction .....	9
2.2	Cases of AI-Enabled Discrimination .....	10
2.2.1	Case 1: Gender Bias in Recruitment Tools .....	10
2.2.2	Case 2: Discriminatory Use of AI in Law Enforcement and Predictive Policing .....	12
2.2.3	Case 3: Discrimination on the Basis of Skin Colour .....	13
2.3	Ethical Questions Concerning AI-Enabled Discrimination .....	14
2.4	Responses to Unfair/Illegal Discrimination .....	16
2.4.1	AI Impact Assessment .....	17
2.4.2	Ethics by Design .....	18
2.5	Key Insights .....	20
	References .....	20
<b>3</b>	<b>Privacy .....</b>	<b>25</b>
3.1	Introduction .....	25
3.2	Cases of Privacy Violations Through AI .....	27
3.2.1	Case 1: Use of Personal Data by Authoritarian Regimes ...	27
3.2.2	Case 2: Genetic Privacy .....	28
3.2.3	Case 3: Biometric Surveillance .....	30
3.3	Data Protection and Privacy .....	30
3.4	Responses to AI-Related Privacy Threats .....	32
3.5	Key Insights .....	34
	References .....	35
<b>4</b>	<b>Surveillance Capitalism .....</b>	<b>39</b>
4.1	Introduction .....	39
4.2	Cases of AI-Enabled Surveillance Capitalism .....	40
4.2.1	Case 1: Data Appropriation .....	40

4.2.2 Case 2: Monetisation of Health Data .....	41
4.2.3 Case 3: Unfair Commercial Practices .....	42
4.3 Ethical Questions About Surveillance Capitalism .....	42
4.4 Responses to Surveillance Capitalism .....	44
4.4.1 Antitrust Regulation .....	45
4.4.2 Data Sharing and Access .....	45
4.4.3 Strengthening of Data Ownership Claims of Consumers/Individuals .....	46
4.5 Key Insights .....	47
References .....	48
<b>5 Manipulation .....</b>	<b>53</b>
5.1 Introduction .....	53
5.2 Cases of AI-Enabled Manipulation .....	54
5.2.1 Case 1: Election Manipulation .....	54
5.2.2 Case 2: Pushing Sales During “Prime Vulnerability Moments” .....	54
5.3 The Ethics of Manipulation .....	55
5.4 Responses to Manipulation .....	58
5.5 Key Insights .....	59
References .....	60
<b>6 Right to Life, Liberty and Security of Persons .....</b>	<b>63</b>
6.1 Introduction .....	63
6.2 Cases of AI Adversely Affecting the Right to Life, Liberty and Security of Persons .....	65
6.2.1 Case 1: Fatal Crash Involving a Self-driving Car .....	65
6.2.2 Case 2: Smart Home Hubs Security Vulnerabilities .....	66
6.2.3 Case 3: Adversarial Attacks in Medical Diagnosis .....	67
6.3 Ethical Questions .....	68
6.3.1 Human Safety .....	68
6.3.2 Privacy .....	69
6.3.3 Responsibility and Accountability .....	69
6.4 Responses .....	71
6.4.1 Defining and Strengthening Liability Regimes .....	71
6.4.2 Quality Management for AI Systems .....	72
6.4.3 Adversarial Robustness .....	73
6.5 Key Insights .....	73
References .....	74
<b>7 Dignity .....</b>	<b>79</b>
7.1 Introduction .....	79
7.2 Cases of AI in Potential Conflict with Human Dignity .....	81
7.2.1 Case 1: Unfair Dismissal .....	81
7.2.2 Case 2: Sex Robots .....	83
7.2.3 Case 3: Care Robots .....	85

Contents	ix
7.3 Ethical Questions Concerning AI and Dignity .....	86
7.4 Key Insights .....	90
References .....	91
<b>8 AI for Good and the SDGs .....</b>	<b>95</b>
8.1 Introduction .....	95
8.2 Cases of AI for Good or Not? .....	97
8.2.1 Case 1: Seasonal Climate Forecasting in Resource-Limited Settings .....	97
8.2.2 Case 2: “Helicopter Research” .....	98
8.3 Ethical Questions Concerning AI for Good and the SDGs .....	99
8.3.1 The Data Desert or the Uneven Distribution of Data Availability .....	100
8.3.2 The Application of Double Standards .....	101
8.3.3 Ignoring the Social Determinants of the Problems the SDGs Try to Solve .....	101
8.3.4 The Elephant in the Room: The Digital Divide and the Shortage of AI Talent .....	102
8.3.5 Wider Unresolved Challenges Where AI and the SDGs Are in Conflict .....	102
8.4 Key Insights .....	103
References .....	104
<b>9 The Ethics of Artificial Intelligence: A Conclusion .....</b>	<b>107</b>
Reference .....	111
<b>Index .....</b>	<b>113</b>

# Abbreviations

AI HLEG	High-Level Expert Group on AI (EU)
AI	Artificial intelligence
AI-IA	AI impact assessment
ALTAI	Assessment List for Trustworthy AI (EU)
AT + ALP	Attention and Adversarial Logit Pairing
CASR	Campaign Against Sex Robots
CEN	European Committee for Standardization
CENELEC	European Electrotechnical Committee for Standardization
CNIL	Commission Nationale de l’Informatique et des Libertés (France)
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
DPIA	Data protection impact assessment
DSPO	Detect and suppress the potential outliers
EDPB	European Data Protection Board (EU)
EDPS	European Data Protection Supervisor (EU)
ENISA	European Union Agency for Cybersecurity (EU)
FRA	Fundamental Rights Agency (EU)
GAN	Generative adversarial network
GDPR	General Data Protection Regulation (EU)
GM	Genetically modified
HR	Human resources
ICO	Information Commissioner’s Office (UK)
ICT	Information and communication technology
ISPA	International Society of Precision Agriculture
IT	Information technology
NTSB	National Transportation Safety Board (USA)
OECD	Organisation for Economic Co-operation and Development
PLD	Product Liability Directive (EU)
SCF	Seasonal climate forecasting
SDGs	Sustainable Development Goals (UN)
SHGP	Saudi Human Genome Program

UNCTAD	UN Conference on Trade and Development
UNDP	UN Development Programme
UNESCO	UN Educational, Scientific and Cultural Organization
UNSDG	UN Sustainable Development Group
WHO	World Health Organization

# Chapter 1

## The Ethics of Artificial Intelligence: An Introduction



**Abstract** This chapter introduces the themes covered by the book. It provides an overview of the concept of artificial intelligence (AI) and some of the technologies that have contributed to the current high level of visibility of AI. It explains why using case studies is a suitable approach to engage a broader audience with an interest in AI ethics. The chapter provides a brief overview of the structure and logic of the book by indicating the content of the cases covered in each section. It concludes by identifying the concept of ethics used in this book and how it is located in the broader discussion of ethics, human rights and regulation of AI.

**Keywords** Artificial intelligence · Machine learning · Deep learning ethics

The ethical challenges presented by artificial intelligence (AI) are one of the biggest topics of the twenty-first century. The potential benefits of AI are said to be numerous, ranging from operational improvements, such as the reduction of human error (e.g. in medical diagnosis), to the use of robots in hazardous situations (e.g. to secure a nuclear plant after an accident). At the same time, AI raises many ethical concerns, ranging from algorithmic bias and the digital divide to serious health and safety concerns.

The field of AI ethics has boomed into a global enterprise with a wide variety of players. Yet the ethics of artificial intelligence (AI) is nothing new. The concept of AI is almost 70 years old (McCarthy et al. 2006) and ethical concerns about AI have been raised since the middle of the twentieth century (Wiener 1954; Dreyfus 1972; Weizenbaum 1977). The debate has now gained tremendous speed thanks to wider concerns about the use and impact of better algorithms, the growing availability of computing resources and the increasing amounts of data that can be used for analysis (Hall and Pesenti 2017).

These technical developments have favoured specific types of AI, in particular machine learning (Alpaydin 2020; Faggella 2020), of which deep learning is one popular form (see box) (LeCun et al. 2015). The success of these AI approaches led to a rapidly expanding set of uses and applications which frequently resulted

in consequences that were deemed ethically problematic, such as unfair or illegal discrimination, exclusion and political interference.

### Deep Learning

Deep learning is one of the approaches to machine learning that have led to the remarkable successes of AI in recent years (Bengio et al. 2021). The development of deep learning is a result of the use of artificial neural networks, which are attempts to replicate or simulate brain functions. Natural intelligence arises from parallel networks of neurons that learn by adjusting the strengths of their connections. Deep learning attempts to perform brain-like activities using statistical measures to determine how well a network is performing. Deep learning derives its name from deep neural networks, i.e. networks with many layers. It has been successfully applied to problems ranging from image recognition to natural speech processing. Despite its successes, deep learning has to contend with a range of limitations (Cremer 2021). It is open to debate how much further machine learning based on approaches like deep learning can progress and whether fundamentally different principles might be required, such as the introduction of causality models (Schölkopf et al. 2021).

With new uses of AI, AI ethics has flourished well beyond academia. For instance, the Rome Call for AI Ethics,<sup>1</sup> launched in February 2020, links the Vatican with the UN Food and Agriculture Organization (FAO), Microsoft, IBM and the Italian Ministry of Innovation. Another example is that UNESCO appointed 24 experts from around the world in July 2021 and launched a worldwide online consultation on AI ethics and facilitated dialogue with all UNESCO member states. Media interest is also considerable, although some academics consider the treatment of AI ethics by the media as “shallow” (Ouchchy et al. 2020).

One of the big problems that AI ethics and ethicists might face is the opaqueness of what is actually happening in AI, given that a good grasp of an activity itself is very helpful in determining its ethical issues.

[I]t is not the role nor to be expected of an AI Ethicist to be able to program the systems themselves. Instead, a strong understanding of aspects such as the difference between supervised and unsupervised learning, what it means to label a dataset, how consent of the user is obtained – essentially, how a system is designed, developed, and deployed – is necessary. In other words, an AI Ethicist must comprehend enough to be able to apprehend the instances in which key ethical questions must be answered (Gambelin 2021).

There is thus an expectation that AI ethicists are familiar with the technology, yet “[n]o one really knows how the most advanced algorithms do what they do” (Knight 2017), including AI developers themselves.

Despite this opacity of AI in its current forms, it is important to reflect on and discuss which ethical issues can arise due to its development and use. The approach to AI ethics we have chosen here is to use case studies, as “[r]eal experiences in AI ethics present … nuanced examples” (Brusseau 2021) for discussion, learning and

<sup>1</sup> <https://www.romecall.org/>.

analysis. This approach will enable us to illustrate the main ethical challenges of AI, often with reference to human rights (Franks 2017).

Case studies are a proven method for increasing insights into theoretical concepts by illustrating them through real-world situations (Escartín et al. 2015). They also increase student participation and enhance the learning experience (*ibid*) and are therefore well-suited to teaching (Yin 2003).

We have therefore chosen the case study method for this book. We selected the most significant or pertinent ethical issues that are currently discussed in the context of AI (based on and updated from Andreou et al. 2019 and other sources) and dedicated one chapter to each of them.

The structure of each chapter is as follows. First, we introduce short real-life case vignettes to give an overview of a particular ethical issue. Second, we present a narrative assessment of the vignettes and the broader context. Third, we suggest ways in which these ethical issues could be addressed. This often takes the form of an overview of the tools available to reduce the ethical risks of the particular case; for instance, a case study of algorithmic bias leading to discrimination will be accompanied by an explanation of the purpose and scope of AI impact assessments. Where tools are not appropriate, as human decisions need to be made based on ethical reasoning (e.g. in the case of sex robots), we provide a synthesis of different argument strategies. Our focus is on *real-life* scenarios, most of which have already been published by the media or research outlets. Below we present a short overview of the cases.

### *Unfair and Illegal Discrimination* (Chap. 2)

The first vignette deals with the automated shortlisting of job candidates by an AI tool trained with CVs (résumés) from the previous ten years. Notwithstanding efforts to address early difficulties with gender bias, the company eventually abandoned the approach as it was not compatible with their commitment to workplace diversity and equality.

The second vignette describes how parole was denied to a prisoner with a model rehabilitation record based on the risk-to-society predictions of an AI system. It became clear that subjective personal views given by prison guards, who may have been influenced by racial prejudices, led to an unreasonably high risk score.

The third vignette tells the story of an engineering student of Asian descent whose passport photo was rejected by New Zealand government systems because his eyes were allegedly closed. This was an ethnicity-based error in passport photo recognition, which was also made by similar systems elsewhere, affecting, for example, dark-skinned women in the UK.

### *Privacy* (Chap. 3)

The first vignette is about the Chinese social credit scoring system, which uses a large number of data points to calculate a score of citizens' trustworthiness. High scores lead to the allocation of benefits, whereas low scores can result in the withdrawal of services.

The second vignette covers the Saudi Human Genome Program, with predicted benefits in the form of medical breakthroughs versus genetic privacy concerns.

#### *Surveillance Capitalism* (Chap. 4)

The first vignette deals with photo harvesting from services such as Instagram, LinkedIn and YouTube in contravention of what users of these services were likely to expect or have agreed to. The relevant AI software company, which specialises in facial recognition software, reportedly holds ten billion facial images from around the world.

The second vignette is about a data leak from a provider of health tracking services, which made the health data of 61 million people publicly available.

The third vignette summarises Italian legal proceedings against Facebook for misleading its users by not explaining to them in a timely and adequate manner, during the activation of their account, that data would be collected with commercial intent.

#### *Manipulation* (Chap. 5)

The first vignette covers the Facebook and Cambridge Analytica scandal, which allowed Cambridge Analytica to harvest 50 million Facebook profiles, enabling the delivery of personalised messages to the profile holders and a wider analysis of voter behaviour in the run-up to the 2016 US presidential election and the Brexit referendum in the same year.

The second vignette shows how research is used to push commercial products to potential buyers at specifically determined vulnerable moments, e.g. beauty products being promoted at times when recipients of online commercials are likely to feel least attractive.

#### *Right to Life, Liberty and Security of Person* (Chap. 6)

The first vignette is about the well-known crash of a Tesla self-driving car, killing the person inside.

The second vignette summarises the security vulnerabilities of smart home hubs, which can lead to man-in-the-middle attacks, a type of cyberattack in which the security of a system is compromised, allowing an attacker to eavesdrop on confidential information.

The third vignette deals with adversarial attacks in medical diagnosis, in which an AI-trained system could be fooled to the extent of almost 70% with fake images.

#### *Dignity* (Chap. 7)

The first vignette describes the case of an employee who was wrongly dismissed and escorted off his company's premises by security guards, with implications for his dignity. The dismissal decision was based on opaque decision-making by an AI tool, communicated by an automatic system.

The second vignette covers sex robots, in particular whether they are an affront to the dignity of women and female children.

Similarly, the third vignette asks whether care robots are an affront to the dignity of elderly people.

### *AI for Good and the UN's Sustainable Development Goals (Chap. 8)*

The first vignette shows how seasonal climate forecasting in resource-limited settings has led to the denial of credits for poor farmers in Zimbabwe and Brazil and the accelerated the layoff of workers in the fishing industry in Peru.

The second vignette deals with a research team from a high-income country requesting vast amounts of mobile phone data from users in Sierra Leone, Guinea and Liberia to track population movements during the Ebola crisis. Commentators argued that the time spent negotiating the request with seriously under-resourced governance structures should have been used to handle the escalating Ebola crisis.

This is a book of AI ethics case studies and not a philosophical book on ethics. We nevertheless need to be clear about our use of the term “ethics”. We use the concept of ethics cognisant of the venerable tradition of ethical discussion and of key positions such as those based on an evaluation of the duty of an ethical agent (Kant 1788, 1797), the consequences of an action (Bentham 1789; Mill 1861), the character of the agent (Aristotle 2000) and the keen observation of potential biases in one’s own position, for instance through using an ethics of care (Held 2005). We slightly favour a Kantian position in several chapters, but use and acknowledge others. We recognize that there are many other ethical traditions beyond the dominant European ones mentioned here, and we welcome debate about how these may help us understand further aspects of ethics and technology. We thus use the term “ethics” in a pluralistic sense.

This approach is pluralistic because it is open to interpretations from the perspective of the main ethical theories as well as other theoretical positions, including more recent attempts to develop ethical theories that are geared more specifically to novel technologies, such as disclosive ethics (Brey 2000), computer ethics (Bynum 2001), information ethics (Floridi 1999) and human flourishing (Stahl 2021).

Our pluralistic reading of the ethics of AI is consistent with much of the relevant literature. A predominant approach to AI ethics is the development of guidelines (Jobin et al. 2019), most of which are based on mid-level ethical principles typically developed from the principles of biomedical ethics (Childress and Beauchamp 1979). This is also the approach adopted by the European Union’s High-Level Expert Group on AI (AI HLEG 2019). The HLEG’s intervention has been influential, as it has had a great impact on the discussion in Europe, which is where we are physically located and which is the origin of the funding for our work (see Acknowledgements). However, there has been significant criticism of the approach to AI ethics based on ethical principles and guidelines (Mittelstadt 2019; Rességuier and Rodrigues 2020). One key concern is that it remains far from the application and does not explain how AI ethics can be put into practice. With the case-study-based approach presented in this book, we aim to overcome this point of criticism, enhance ethical reflection and demonstrate possible practical interventions.

We invite the reader to critically accompany us on our journey through cases of AI ethics. We also ask the reader to think beyond the cases presented here and ask

fundamental questions, such as whether and to what degree the issues discussed here are typical or exclusively relevant to AI and whether one can expect them to be resolved.

Overall, AI is an example of a current and dynamically developing technology. An important question is therefore whether we can keep reflecting and learn anything from the discussion of AI ethics that can be applied to future generations of technologies to ensure that humanity benefits from technological progress and development and has ways to deal with the downsides of technology.

## References

- AI HLEG (2019) Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence. European Commission, Brussels. [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=60419](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419). Accessed 25 Sept 2020
- Alpaydin E (2020) Introduction to machine learning. The MIT Press, Cambridge
- Andreou A, Laulhe Shaelou S, Schroeder D (2019) D1.5 Current human rights frameworks. De Montfort University. Online resource. <https://doi.org/10.21253/DMU.8181827.v3>
- Aristotle (2000) Nicomachean ethics (trans: Crisp R). Cambridge University Press, Cambridge
- Bengio Y, Lecun Y, Hinton G (2021) Deep learning for AI. Commun ACM 64:58–65. <https://doi.org/10.1145/3448250>
- Bentham J (1789) An introduction to the principles of morals and legislation. Dover Publications, Mineola
- Brey P (2000) Disclosive computer ethics. SIGCAS Comput Soc 30(4):10–16. <https://doi.org/10.1145/572260.572264>
- Brusseau J (2021) Using edge cases to disentangle fairness and solidarity in AI ethics. AI Ethics. <https://doi.org/10.1007/s43681-021-00090-z>
- Bynum TW (2001) Computer ethics: its birth and its future. Ethics Inf Technol 3:109–112. <https://doi.org/10.1023/A:1011893925319>
- Childress JF, Beauchamp TL (1979) Principles of biomedical ethics. Oxford University Press, New York
- Cremer CZ (2021) Deep limitations? Examining expert disagreement over deep learning. Prog Artif Intell 10:449–464. <https://doi.org/10.1007/s13748-021-00239-1>
- Dreyfus HL (1972) What computers can't do: a critique of artificial reason. Harper & Row, New York
- Escartín J, Saldaña O, Martín-Peña J et al (2015) The impact of writing case studies: benefits for students' success and well-being. Procedia Soc Behav Sci 196:47–51. <https://doi.org/10.1016/j.sbspro.2015.07.009>
- Faggella D (2020) Everyday examples of artificial intelligence and machine learning. Emerj, Boston. <https://emerj.com/ai-sector-overviews/everyday-examples-of-ai/>. Accessed 23 Sept 2020
- Floridi L (1999) Information ethics: on the philosophical foundation of computer ethics. Ethics Inf Technol 1:33–52. <https://doi.org/10.1023/A:1010018611096>
- Franks B (2017) The dilemma of unexplainable artificial intelligence. Datafloq, 25 July. <https://datafloq.com/read/dilemma-unexplainable-artificial-intelligence/>. Accessed 18 May 2022
- Gambelin O (2021) Brave: what it means to be an AI ethicist. AI Ethics 1:87–91. <https://doi.org/10.1007/s43681-020-00020-5>
- Hall W, Pesenti J (2017) Growing the artificial intelligence industry in the UK. Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy, London

- Held V (2005) The ethics of care: personal, political, and global. Oxford University Press, New York
- Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kant I (1788) Kritik der praktischen Vernunft. Reclam, Ditzingen
- Kant I (1797) Grundlegung zur Metaphysik der Sitten. Reclam, Ditzingen
- Knight W (2017) The dark secret at the heart of AI. MIT Technology Review, 11 Apr. <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>. Accessed 18 May 2022
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444. <https://doi.org/10.1038/nature14539>
- McCarthy J, Minsky ML, Rochester N, Shannon CE (2006) A proposal for the Dartmouth summer research project on artificial intelligence. AI Mag 27:12–14. <https://doi.org/10.1609/aimag.v27i4.1904>
- Mill JS (1861) Utilitarianism, 2nd revised edn. Hackett Publishing Co, Indianapolis
- Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Ouchchy L, Coin A, Dubljević V (2020) AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. AI & Soc. <https://doi.org/10.1007/s00146-020-00965-5>
- Rességuier A, Rodrigues R (2020) AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc 7:2053951720942541. <https://doi.org/10.1177/2053951720942541>
- Schölkopf B, Locatello F, Bauer S et al (2021) Toward causal representation learning. Proc IEEE 109(5):612–634. <https://doi.org/10.1109/JPROC.2021.3058954>
- Stahl BC (2021) Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies. Springer Nature Switzerland AG, Cham. <https://doi.org/10.1007/978-3-030-69978-9>
- UNESCO (2021) AI ethics: another step closer to the adoption of UNESCO's recommendation. UNESCO, Paris. Press release, 2 July. <https://en.unesco.org/news/ai-ethics-another-step-closer-adoption-unescos-recommendation-0>. Accessed 18 May 2022
- Weizenbaum J (1977) Computer power and human reason: from judgement to calculation, new edn. W.H. Freeman & Co Ltd., New York
- Wiener N (1954) The human use of human beings. Doubleday, New York
- Yin RK (2003) Applications of case study research, 2nd edn. Sage Publications, Thousand Oaks

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# Chapter 2

## Unfair and Illegal Discrimination



**Abstract** There is much debate about the ways in which artificial intelligence (AI) systems can include and perpetuate biases and lead to unfair and often illegal discrimination against individuals on the basis of protected characteristics, such as age, race, gender and disability. This chapter describes three cases of such discrimination. It starts with an account of the use of AI in hiring decisions that led to discrimination based on gender. The second case explores the way in which AI can lead to discrimination when applied in law enforcement. The final example looks at implications of bias in the detection of skin colour. The chapter then discusses why these cases are considered to be ethical issues and how this ethics debate relates to well-established legislation around discrimination. The chapter proposes two ways of raising awareness of possible discriminatory characteristics of AI systems and ways of dealing with them: AI impact assessments and ethics by design.

**Keywords** Discrimination · Bias · Gender · Race · Classification · Law enforcement · Predictive policing · AI impact assessment · Ethics by design

### 2.1 Introduction

Concern at discrimination is probably the most widely discussed and recognised ethical issue linked to artificial intelligence (AI) (Access Now 2018; Latonero 2018; Muller 2020). In many cases an AI system analyses existing data which was collected for purposes other than the ones that the AI system is pursuing and therefore typically does so without paying attention to properties of the data that may facilitate unfair discrimination when used by the AI system. Analysis of the data using AI reveals underlying patterns that are then embedded in the AI model used for decision-making. In these cases, which include our examples of gender bias in staff recruitment and predictive policing that disadvantages segments of the population, the system perpetuates existing biases and reproduces prior practices of discrimination.

In some cases, discrimination occurs through other mechanisms, for example when a system is exposed to real-world data that is fundamentally different from the data it was trained on and cannot process the data correctly. Our case of systems that

misclassify people from ethnic groups that are not part of the training data falls into this category. In this case the system works in a way that is technically correct, but the outputs are incorrect, due to a lack of correspondence between the AI model and the input data.

These examples of AI-enabled discrimination have in common that they violate a human right (see box) that individuals should not be discriminated against. That is why these systems deserve attention and are the subject of this chapter.

#### **Universal Declaration of Human Rights, Article 7**

“All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination.”  
(UN 1948)

## **2.2 Cases of AI-Enabled Discrimination**

### ***2.2.1 Case 1: Gender Bias in Recruitment Tools***

Recruiting new members of staff is an important task for an organisation, given that human resources are often considered the most valuable assets a company can have. At the same time, recruitment can be time- and resource-intensive. It requires organisations to scrutinise job applications and CVs, which are often non-standardised, complex documents, and to make decisions on shortlisting and appointments on the basis of this data. It is therefore not surprising that recruitment was an early candidate for automation by machine learning. One of the most high-profile examples of AI use for recruitment is an endeavour by Amazon to automate the candidate selection process.

In 2014, Amazon started to develop and use AI programs to mechanise highly time-intensive human resources (HR) work, namely the shortlisting of applicants for jobs. Amazon “literally wanted it to be an engine where I’m going to give you 100 résumés, it will spit out the top five, and we’ll hire those” (Reuters 2018). The AI tool was trained on CVs submitted over an earlier ten-year period and the related staff appointments. Following this training, the AI tool discarded the applications of female applicants, even where no direct references to applicants’ gender were provided. Given the predominance of successful male applicants in the training sample, Amazon found that the system penalised language such as “women’s chess club captain” for not matching closely enough the successful male job applicants of the past. While developers tried to modify the system to avoid gender bias, Amazon abandoned its use in the recruitment process in 2015 as a company “committed to workplace diversity and equality” (*ibid*).

At first this approach seemed promising, as HR departments have ample training data in the form of past applications. A machine learning system can thus be trained to distinguish between successful and unsuccessful past applications and identify features of applications that are predictors of success. This is exactly what Amazon did. The result was that the AI systematically discriminated against women.

When it became clear that women were being disadvantaged by recruitment based on AI, ways were sought to fix the problem. The presumptive reason for the outcome was that there were few women in the training sample, maybe because the tech sector is traditionally male dominated, or maybe reflecting biases in the recruitment system overall. It turned out, however, that even removing direct identifiers of sex and gender did not level the playing field, as the AI found proxy variables that still pointed to gender, such as place of study (e.g., all-female college) and feminised hobbies.

AI systems are only as good as the data they're trained on and the humans that build them. If a résumé-screening machine-learning tool is trained on historical data, such as résumés collected from a company's previously hired candidates, the system will inherit both the conscious and unconscious preferences of the hiring managers who made those selections (Heilweil 2019).

In the case of Amazon this eventually led to the company's abandoning the use of AI for hiring, as explained in the case description. However, the fundamental challenge of matching large numbers of candidates for recruitment with large numbers of open positions on the basis of complex and changing selection criteria remains. For instance, Vodafone is reported to have used AI systems to analyse over 100,000 graduate applications for 1,000 jobs (Kaur 2021). Since 2019, the COVID-19 pandemic has accelerated the use of AI recruitment, with predictions that 16% of HR recruitment jobs will have disappeared by 2029 (*ibid*).

AI can also, it is claimed, be used as a tool for measuring psychological, emotional and personality features during video interviews (Heilweil 2019). Online interviews have become the norm under COVID-19 lockdowns, and this trend seems set to continue, so the use of AI technology in these contexts may increase. However, tools that interpret facial features may manifest limitations similar to those of recruitment AI, although their impact is not as widely publicised as that of the Amazon case. This means that sustained ethical alertness is required when it comes to preventing violations of the human right to non-discrimination. Or, as a human rights commentator has noted, the problem of "garbage in, garbage out" (Lentz 2021) has to be solved before HR departments can use AI in an ethical manner to substitute human for machine decision-making.

## 2.2.2 Case 2: Discriminatory Use of AI in Law Enforcement and Predictive Policing

Glenn Rodríguez had been arrested at the age of 16 for his role in the armed robbery of a car dealership, which left one employee dead. In 2016, 25 years later, he applied to the parole board of the Eastern Correctional Facility in upstate New York for early release. He had a model rehabilitation record at the time (Wexler 2017b). Parole was denied. The justification given by the board was that an AI system called COMPAS had predicted him to be “high risk” and the board “concluded that … release to supervision is not compatible with the welfare of society” (Wexler 2017a). The parole board had no knowledge of how the COMPAS risk score was calculated, as the company that had developed the system considered their algorithm a trade secret (*ibid*). Through cross-referencing with other inmates’ scores, Rodríguez found out that the reason for his high-risk score was a subjective personal view given by prison guards, who may have been influenced by racial prejudices. In the end, he was released early. However, “had he been able to examine and contest the logic of the COMPAS system to prove that its score gave a distorted picture of his life, he might have gone home much earlier” (Wexler 2017b)

Rodríguez’s case is an example of the discriminatory use of AI in criminal justice, which also includes prominent AI applications for the purposes of predictive policing. “Predictive policing makes use of information technology, data, and analytical techniques in order to identify likely places and times of future crimes or individuals at high risk of [re]-offending or becoming victims of crime.” (Mugari and Obioha 2021: 1). The idea behind predictive policing is that existing law enforcement data can improve the targeting of policing interventions. Police resources are limited and it would be desirable to focus them where they are most likely to make a difference, that is, to disrupt or prevent crime or, once crime has been committed, to protect victims, arrest offenders etc. Predictive policing uses past crime data to detect patterns suitable for extrapolation into the future, thereby, one hopes, helping police to identify locations and times when crime is most likely to occur. This is where resources are then deployed.

These ideas sound plausible and are already implemented in many jurisdictions. The most high-profile cases are from the US, where police have been developing and using predictive policing tools in Chicago, Los Angeles, New Orleans and New York since as far back as 2012 (McCarthy 2019). In the UK, research by an NGO showed that “at least 14 UK police forces have used or intend to use … computer algorithms to predict where crime will be committed and by whom” (Liberty n.d.). It is also known that China, Denmark, Germany, India, the Netherlands, and Japan are testing and possibly deploying predictive policing tools (McCarthy 2019).

While the idea of helping the police do their job better, and possibly at reduced cost, will be welcomed by many, the practice of predictive policing has turned out to be ethically problematic. The use of past crime data means that historical patterns are reproduced, and this may become a self-fulfilling prophecy.

For example, areas that historically have high crime rates tend to be those that have lower levels of wealth and educational attainment among the population, as well as higher percentages of migrants or stateless people. Using predictive policing tools means that people who live in deprived areas are singled out for additional police attention, whether they have anything to do with perpetrating any crimes or not. Using algorithmic systems to support policing work has the potential to exacerbate already entrenched discrimination. It is worth pointing out, however, that given awareness of the issue, it is also conceivable that such systems could explicitly screen police activity for bias and help alleviate the problem. The AI systems used for predictive policing and law enforcement could be used to extract and visualise crime data that would make more obvious whether and how crime statistics are skewed in ways that might be linked to ethnic or racial characteristics. This, in turn, would provide a good starting point for a more detailed analysis of the mechanisms that contribute to such developments.

This problem of possible discrimination in relation to specific geographical areas can also occur in relation to individuals. Automated biometric recognition can be used in police cameras, providing police officers with automated risk scores for people they interact with. This then disadvantages people with prior convictions or a past history of interaction with the police, which again tends to over-represent disadvantaged communities, notably those from ethnic minorities. The same logic applies further down the law enforcement chain, when the analysis of data from offenders is used to predict their personal likelihood of reoffending. When the AI tool which informed the decision to hold Glenn Rodríguez in prison for longer than necessary was later examined, it was found that “a disproportionate number of black defendants were ‘false positives’: they were classified by COMPAS as high risk but subsequently not charged with another crime.” (Courtland [2018](#)).

### **2.2.3 Case 3: Discrimination on the Basis of Skin Colour**

In 2016, a 22-year-old engineering student from New Zealand had his passport photo rejected by the systems of the New Zealand department of internal affairs because his eyes were allegedly closed. The student was of Asian descent and his eyes were open. The automatic photo recognition tool declared the photo invalid and the student could not renew his passport. He later told the press very graciously: “No hard feelings on my part, I’ve always had very small eyes and facial recognition technology is relatively new and unsophisticated” (Reuters [2016](#)). Similar cases of ethnicity-based errors by passport photo recognition tools have affected dark-skinned women in the UK. “Photos of women with the darkest skin were four times more likely to be graded poor quality, than women with the lightest skin” (Ahmed [2020](#)). For instance, a black student’s photo was declared unsuitable as her mouth was allegedly open, which it in fact was not (*ibid*).

Zou and Schiebinger (2018) have explained how such discriminatory bias can occur. As noted earlier, one of the main reasons for discriminatory AI tools is the training sets used.

Deep neural networks for image classification ... are often trained on ImageNet ... More than 45% of ImageNet data, which fuels research in computer vision, comes from the United States, home to only 4% of the world's population.

Hence, some groups are heavily over-represented in training sets while others are under-represented, leading to the perpetuation of ethnicity-based discrimination.

## 2.3 Ethical Questions Concerning AI-Enabled Discrimination

The reproduction of biases and resulting discrimination are among the most prominent ethical concerns about AI (Veale and Binns 2017; Access Now Policy Team 2018). Bias has been described as the “one of the biggest risks associated with AI” (PwC 2019: 13).

The term “discrimination” has at least two distinct meanings, which differ significantly in terms of an ethical analysis (Cambridge Dictionary n.d.). On one hand “discrimination” means the ability to judge phenomena and distinguish between them in a reasonable manner. In this sense, the term has synonyms like “distinction” and “differentiation”. For instance, it is a good evolutionary trait for humans to have the ability to distinguish malaria-carrying mosquitoes from flies. The other more widespread contemporary meaning of the term focuses on the unjust or prejudicial application of distinctions made between people, in particular on the basis of their race, sex, age or disability. The former meaning can be ethically neutral, whereas the latter is generally acknowledged to be a significant ethical problem, hence article 7 of the Universal Declaration of Human Rights (see box above). When we use the term “discrimination” in this discussion, we are talking about the ethically relevant type, which is also often illegal.

However, being able to distinguish between phenomena is one of the strengths of AI. Machine-learning algorithms are specifically trained to distinguish between classes of phenomena, and their success in doing so is the main reason for the current emphasis on AI use in a wide field of applications.

AI systems have become increasingly adept at drawing distinctions, at first between pictures of cats and pictures of dogs, which provided the basis for their use in more socially relevant fields, such as medical pathology, where they can distinguish images of cancer cells from those of healthy tissue, or in the business world, where they can distinguish fraudulent insurance claims from genuine ones. The problem is not identifying differences in the broad sense but discrimination on the basis of those particular characteristics.

Unfair/illegal discrimination is a widespread characteristic of many social interactions independent of AI use. While there is broad agreement that job offers should

not depend on an applicant's gender, and that judicial or law enforcement decisions should not depend on a person's ethnicity, it is also clear that they often do, reflecting ingrained systemic injustices. An AI system that is trained on historical data that includes data from processes that structurally discriminated against people will replicate that discrimination. As our case studies have shown, these underlying patterns in the data are difficult to eradicate. Attempts to address such problems by providing more inclusive data may offer avenues for overcoming them. However, there are many cases where no alternative relevant datasets exist. In such cases, which include law enforcement and criminal justice applications, the attempt to modify the data to reduce or eliminate underlying biases may inadvertently introduce new challenges.

However, there are cases where the problem is not so much that no unbiased datasets exist but that the possibility of introducing biases through a poor choice of training data is not sufficiently taken into account. An example is unfair/illegal discrimination arising from poor systems design through a poor choice of training data. Our third case study points in this direction. When the images of 4% of the world population constitute 45% of the images used in AI system design (Zou and Schiebinger 2018), it is reasonable to foresee unfair/illegal discrimination in the results.

This type of discrimination will typically arise when a machine-learning system is trained on data that does not fully represent the population that the system is meant to be applied to. Skin colour is an obvious example, where models based on data from one ethnic group do not work properly when applied to a different group. Such cases are similar to the earlier ones (Amazon HR and parole) in that there is a pre-existing bias in the original data used to train the model. The difference between the two types of discrimination is the source of the bias in the training data. In the first two cases the biases were introduced by the systems involved in creating the data, i.e. in recruitment processes and law enforcement, where women and racial minorities were disadvantaged by past recruiters and past parole boards that had applied structurally sexist or racist perspectives. In the case of discrimination based on skin colour, the bias was introduced by a failure to select comprehensive datasets that included representation from all user communities. This difference is subtle and not always clear-cut. It may be important, however, in that ways of identifying and rectifying particular problems may differ significantly.

Discrimination against people on the basis of gender, race, age etc. is not only an ethical issue; in many jurisdictions such discrimination is also illegal. In the UK, for example, the Equality Act (2010) defines nine protected characteristics: age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation. Discrimination in the workplace and in wider society based on these protected characteristics is prohibited.

The legal codification of the prohibition of such discrimination points to a strong societal consensus that such discrimination is to be avoided. It raises difficult questions, however, with regard to unfair discrimination that is based on characteristics other than the legally protected ones. It is conceivable that a system would identify

patterns on the basis of other variables that we may not yet even be aware of. Individuals could then be categorised in ways that are detrimental to them. This might not involve protected characteristics, but could still be perceived as unfair discrimination.

Another example of a problematic variable is social class. It is well established that class is an important variable that determines not just individual life chances, but also the collective treatment of groups. Marx's (2017) dictum that the history of all existing society is the history of class struggles exemplifies this position. Discrimination can happen because of a particular characteristic, such as gender, race or disability, but it often happens where individuals combine several of these characteristics that individually can lead to discrimination and, when taken together, exacerbate the discriminatory effect. The term "intersectionality" is sometimes used to indicate this phenomenon (Collins and Bilge 2020). Intersectionality has been recognised as a concern that needs to be considered in various aspects of information technology (IT), not only AI (Fothergill et al. 2019; Zheng and Walsham 2021). It points to the fact that the exact causes of discrimination will in practice often be difficult to identify, which raises questions about the mechanisms of unfair/illegal discrimination as well as ways of addressing them. If the person who is discriminated against is a black, disabled, working-class woman, then it may be impossible to determine which characteristic led to the discrimination, and whether the discrimination was based on protected characteristics and thus illegal.

Hence, unfair/illegal discrimination is not a simple matter. Discrimination based on protected characteristics is deemed to be ethically unacceptable in most democratic states and therefore also typically illegal. But this does not mean that there is no discrimination in social reality, nor should we take it as given that the nature of these protected characteristics will remain constant or that discrimination based on gender, age, race etc. are the only forms of unfair discrimination.

## 2.4 Responses to Unfair/Illegal Discrimination

With unfair/illegal discrimination recognised as a key ethical problem related to AI and machine learning, there is no shortage of attempts to address and mitigate it. These range from the technical level, where attempts are made to better understand whether training data contains biases that lead to discrimination, to legislative processes where existing anti-discrimination policies are refocused on novel technologies.

One prominent field of research with significant implications regarding unfair/illegal discrimination is that of explainable AI (Holzinger et al. 2017; Gunning et al. 2019). There are many approaches to explainable AI, but what they have in common is an attempt to render the opaque nature of the transformation from input variables to output variables easier to understand. The logic is that an ability to understand *how* an AI system came to a classification of a particular observation would allow the determination of whether that classification is discriminatory and, as a result, could be challenged. If AI is fully explainable, then it should be easy to

see whether gender (sexism) determines employment offers, or whether racism has consequences for law enforcement practices.

While this approach is plausible, it runs into technical and social limits. The technical limits include the fact that machine learning models include large numbers of variables and by their very nature are not easy to understand. If it were possible to reduce them to simple tests of specific variables, then machine learning would not be needed in the first place. However, it might be possible for explainable AI to find ways of testing whether an AI system makes use of protected characteristics and to correct for this (Mittelstadt 2019). Hence, rather than humans making these assessments, another or the same AI system would do so.

When thinking about ways of addressing the role that AI plays in unfair/illegal discrimination, it helps to keep in mind that such discrimination is pervasive in many social processes. Real-life data used for training purposes will often include cases of unfair discrimination and thus lead to their reproduction. Removing traces of structural discrimination from training data, for example by removing data referring to protected characteristics, may not work or may reduce the value of the data for training purposes. The importance of data quality to the trustworthiness of the outcomes of an AI system is widely recognised. The European Commission's proposal for regulating AI, for example, stipulates that "training, validation and testing data sets shall be relevant, representative, free of errors and complete" (European Commission 2021: art. 10(3)). It is not clear, however whether such data quality requirements can possibly be met with real-life data.

Two suggestions on how to address unfair/illegal discrimination (Stahl 2021) will be highlighted here: AI impact assessments and ethics by design.

### ***2.4.1 AI Impact Assessment***

The idea of an AI impact assessment is based on the insights derived from many other types of impact assessment, such as social impact assessment (Becker 2001; Becker and Vanclay 2003; Hartley and Wood 2005) and human rights impact assessment (Microsoft and Article One 2018). In general terms, impact assessment aims to come to a better understanding of the possible and likely issues that can arise in a particular field, and use this understanding to prepare mitigation measures. There are several examples of impact assessment that focus on information technologies and topics of relevance to AI, such as privacy/data protection impact assessment (CNIL 2015; Ivanova 2020), ICT ethics impact assessment (Wright 2011) and ethics impact assessment for research and innovation (CEN-CENELEC 2017). The idea of applying impact assessments specifically to AI and using them to get an early warning of possible ethical issues is therefore plausible. This has led to a number of calls for such specific impact assessments for AI by bodies such as the European Data Protection Supervisor (EDPS 2020), UNESCO (2020), the European Fundamental Rights Agency (FRA 2020) and the UK AI council (2021).

The discussion of what such an AI impact assessment should look like in detail is ongoing, but several proposals are available. Examples include the assessment list for trustworthy AI of the European Commission’s High-Level Expert Group on Artificial Intelligence (AI HLEG 2020), the AI Now Institute’s algorithmic impact assessment (Reisman et al. 2018), the IEEE’s recommended practice for assessing the impact of autonomous and intelligent systems on human wellbeing (IEEE 2020) and the ECP Platform’s AI impact assessment (ECP 2019).

The idea common to these AI impact assessments is that they provide a structure for thinking about aspects that are likely to raise concerns at a later stage. They highlight such issues and often propose processes to be put in place to address them. In a narrow sense they can be seen as an aspect of risk management. More broadly they can be interpreted as a proactive engagement that typically includes stakeholder consultation to ensure that likely and foreseeable problems do not arise. Bias and unfair/illegal discrimination figure strongly among these foreseeable problems.

The impact assessment aims to ascertain that appropriate mechanisms for dealing with potential sources of bias and unfair discrimination are flagged early and considered by those designing AI systems. The AI HLEG (2020) assessment, for example, asks whether strategies for avoiding biases are in place, how the diversity and representativeness of end users is considered, whether AI designers and developers have benefitted from education and awareness initiatives to sensitise them to the problem, how such issues can be reported and whether a consistent use of the terminology pertaining to fairness is ensured.

An AI impact assessment is therefore likely to be a good way of raising awareness of the possibility and likelihood that an AI system may raise concerns about unfair/illegal discrimination, and of which form this discrimination might take. However, it typically does not go far in providing a pathway towards addressing such discrimination, which is the ambition of ethics by design.

#### **2.4.2 *Ethics by Design***

Ethics by design for AI has been developed in line with previous discussions of value-sensitive design (Friedman et al. 2008; van den Hoven 2013). The underlying idea is that an explicit consideration of shared values during the design and development process of a project or technology will be conducive to the embedding of such a value in the technology and its eventual use. The concept has been prominently adopted for particular values, for example in the area of privacy by design (ICO 2008) or security by design (Cavoukian 2017).

A key premise of value-sensitive design is that technology is not a value-neutral tool that can be used for any purposes; design decisions influence the way in which a technology can be used and what consequences such use will have. This idea may be most easily exemplified using the value of security by design. Cybersecurity is generally recognised as an important concern that requires continuous vigilance from individuals, organisations and society. It is also well recognised that some systems are

easier to protect from malicious interventions than others. One distinguishing factor between more secure and less secure systems is that secure systems tend to be built with security considerations integrated into the earliest stages of systems design. Highlighting the importance of security, for example in the systems requirement specifications, makes it more likely that the subsequent steps of systems development will be sensitive to the relevance of security and ensure that the system overall contains features that support security. Value-sensitive design is predicated on the assumption that a similar logic can be followed for all sorts of values.

The concept of ethics by design was developed by Philip Brey and his collaborators (Brey and Dainow 2020) with a particular view to embedding *ethical* values in the design and development of AI and related technologies. This approach starts by highlighting the values that are likely to be affected by a particular technology. Brey and Dainow (2020) take their point of departure from the AI HLEG (2019) and identify the following values as relevant: human agency, privacy and data governance, fairness, wellbeing, accountability and oversight, and transparency. The value of fairness is key to addressing questions of bias and unfair/illegal discrimination.

Where ethics by design goes beyond an ex-ante impact assessment is where it specifically proposes ways of integrating attention to the relevant values into the design process. For this purpose, Brey and Dainow (2020) look at the way in which software is designed. Starting with a high-level overview, they distinguish different design phases and translate the ethical values into specific objectives and requirements that can then be fed into the development process. They also propose ways in which this can be achieved in the context of agile development methodologies. This explicit link between ethical concerns and systems development methodologies is a key conceptual innovation of ethics by design. Systems development methodologies are among the foundations of computer science. They aim to ensure that systems can be built according to specifications and perform as expected. The history of computer science has seen the emergence of numerous design methodologies. What Brey and his colleagues have done is to identify universal components that most systems development methodologies share (e.g. objectives specification, requirements elicitation, coding, testing) and to provide guidance on how ethical values can be integrated and reflected in these steps.

This method has only recently been proposed and has not yet been evaluated. It nevertheless seems to offer an avenue for the practical implementation of ethical values, including the avoidance of unfair/illegal discrimination in AI systems. In light of the pervasive nature of unfair/illegal discrimination in most areas of society one can safely say that all AI systems need to be built and used in ways that recognise the possibility of discrimination. Failure to take this possibility into account means that the status quo will be reproduced using AI, which will often be neither ethical nor legal.

## 2.5 Key Insights

Unfair/illegal discrimination is not a new problem, nor one that is confined to technology. However, AI systems have the proven potential to exacerbate and perpetuate it. A key problem in addressing and possibly overcoming unfair/illegal discrimination is that it is pervasive and often hidden from sight. High-profile examples of such discrimination on the basis of gender and race have highlighted the problem, as in our case studies. But unfair/illegal discrimination cannot be addressed by looking at technology alone. The broader societal questions of discrimination need to be considered.

One should also not underestimate the potential for AI to be used as a tool to *identify* cases of unfair/illegal discrimination. The ability of AI to recognise patterns and process large amounts of data means that AI may also be used to demonstrate where discrimination is occurring.

It is too early to evaluate whether – and, if so, how far – AI impact assessment will eliminate the possibility of unfair/illegal discrimination through AI systems. In any event, discrimination on the basis of protected characteristics requires access to personal data, which is the topic of the next chapter, on privacy and data protection.

## References

- Access Now (2018) Human rights in the age of artificial intelligence. Access Now. <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>. Accessed 1 May 2022
- Access Now Policy Team (2018) The Toronto declaration: protecting the right to equality and non-discrimination in machine learning systems. Access Now, Toronto. [https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration\\_ENG\\_08-2018.pdf](https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf). Accessed 26 Sept 2020
- Ahmed M (2020) UK passport photo checker shows bias against dark-skinned women. BBC News, 8 Oct. <https://www.bbc.com/news/technology-54349538>. Accessed 4 May 2022
- AI HLEG (2019) Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence. European Commission, Brussels. [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=60419](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419). Accessed 25 Sept 2020
- AI HLEG (2020) The assessment list for trustworthy AI (ALTAI). High-level expert group on artificial intelligence. European Commission, Brussels. [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=68342](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342). Accessed 10 Oct 2020
- Becker HA (2001) Social impact assessment. Eur J Oper Res 128:311–321. [https://doi.org/10.1016/S0377-2217\(00\)00074-6](https://doi.org/10.1016/S0377-2217(00)00074-6)
- Becker HA, Vanclay F (eds) (2003) The international handbook of social impact assessment: conceptual and methodological advances. Edward Elgar Publishing, Cheltenham
- Brey P, Dainow B (2020) Ethics by design and ethics of use approaches for artificial intelligence, robotics and big data. SIENNA. [https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence\\_he\\_en.pdf](https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf)
- Cambridge Dictionary (n.d.) Discrimination. <https://dictionary.cambridge.org/dictionary/english/discrimination>. Accessed 4 May 2022
- Cavoukian A (2017) Global privacy and security, by design: turning the ‘privacy vs. security’ paradigm on its head. Health Technol 7:329–333. <https://doi.org/10.1007/s12553-017-0207-1>

- CEN-CENELEC (2017) Ethics assessment for research and innovation, part 2: ethical impact assessment framework. CWA 17145-2. European Committee for Standardization, Brussels. <http://ftp.cenelec.eu/EN/ResearchInnovation/CWA/CWA17214502.pdf>. Accessed 6 Oct 2020
- CNIL (2015) Privacy impact assessment (PIA): methodology. Commission Nationale de l'Informatique et des Libertés, Paris
- Collins PH, Bilge S (2020) Intersectionality. Wiley, New York
- Courtland R (2018) Bias detectives: the researchers striving to make algorithms fair. Nature 558:357–360. <https://doi.org/10.1038/d41586-018-05469-3>
- ECP (2019) Artificial intelligence impact assessment. ECP Platform for the Information Society, The Hague. <https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf>. Accessed 1 May 2022
- EDPS (2020) EDPS opinion on the European Commission's white paper on artificial intelligence: a European approach to excellence and trust (opinion 4/2020). European Data Protection Supervisor, Brussels. [https://edps.europa.eu/data-protection/our-work/publications/opinions/edps-opinion-european-commissions-white-paper\\_en](https://edps.europa.eu/data-protection/our-work/publications/opinions/edps-opinion-european-commissions-white-paper_en). Accessed 6 May 2022
- Equality Act (2010) c15. HMSO, London. <https://www.legislation.gov.uk/ukpga/2010/15/contents>. Accessed 5 May 2022
- European Commission (2021) Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. European Commission, Brussels. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>. Accessed 1 May 2022
- Fothergill BT, Knight W, Stahl BC, Ulnicane I (2019) Intersectional observations of the Human Brain Project's approach to sex and gender. J Inf Commun Ethics Soc 17:128–144. <https://doi.org/10.1108/JICES-11-2018-0091>
- FRA (2020) Getting the future right: artificial intelligence and fundamental rights. European Union Agency for Fundamental Rights, Luxembourg
- Friedman B, Kahn P, Borning A (2008) Value sensitive design and information systems. In: Himma K, Tavani H (eds) The handbook of information and computer ethics. Wiley Blackwell, Hoboken, pp 69–102
- Gunning D, Stefik M, Choi J et al (2019) XAI: explainable artificial intelligence. Sci Robot 4(37). <https://doi.org/10.1126/scirobotics.aay7120>
- Hartley N, Wood C (2005) Public participation in environmental impact assessment: implementing the Aarhus convention. Environ Impact Assess Rev 25:319–340. <https://doi.org/10.1016/j.eiar.2004.12.002>
- Heilweil R (2019) Artificial intelligence will help determine if you get your next job. Vox-Recode, 12 Dec. <https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen>. Accessed 4 May
- Holzinger A, Biemann C, Pattichis CS, Kell DB (2017) What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923 [cs, stat]. <https://doi.org/10.48550/arXiv.1712.09923>
- ICO (2008) Privacy by design. Information Commissioner's Office, Wilmslow. [https://web.archive.org/web/20121222044417if\\_/http://www.ico.gov.uk:80/upload/documents/pdb\\_report\\_html/privacy\\_by\\_design\\_report\\_v2.pdf](https://web.archive.org/web/20121222044417if_/http://www.ico.gov.uk:80/upload/documents/pdb_report_html/privacy_by_design_report_v2.pdf). Accessed 6 Oct 2020
- IEEE (2020) 7010-2020: IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being. IEEE Standards Association, Piscataway, NJ. <https://doi.org/10.1109/IEEESTD.2020.9084219>
- Ianova Y (2020) The data protection impact assessment as a tool to enforce non-discriminatory AI. In: Antunes L, Naldi M, Italiano GF et al (eds) Privacy technologies and policy. 8th Annual privacy forum, APF 2020, Lisbon, Portugal, 22–23 Oct. Springer Nature Switzerland, Cham, pp 3–24. [https://doi.org/10.1007/978-3-030-55196-4\\_1](https://doi.org/10.1007/978-3-030-55196-4_1)
- Kaur D (2021) Has artificial intelligence revolutionized recruitment? Tech Wire Asia, 9 Feb. <https://techwireasia.com/2021/02/has-artificial-intelligence-revolutionized-recruitment/>. Accessed 4 May 2022

- Latonero M (2018) Governing artificial intelligence: upholding human rights & dignity. Data & Society. [https://datasociety.net/wp-content/uploads/2018/10/DataSociety\\_Governing\\_Artificial\\_Intelligence\\_Upholding\\_Human\\_Rights.pdf](https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf). Accessed 26 Sept 2020
- Lentz A (2021) Garbage in, garbage out: is AI discriminatory or simply a mirror of IRL inequalities? 18 Jan. Universal Rights Group, Geneva. <https://www.universal-rights.org/blog/garbage-in-garbage-out-is-ai-discriminatory-or-simply-a-mirror-of-irl-inequalities/>. Accessed 4 May 2022
- Liberty (n.d.) Predictive policing. <https://www.libertyhumanrights.org.uk/fundamental/predictive-policing/>. Accessed 4 May 2022
- Marx K (2017) Manifest der Kommunistischen Partei. e-artnow
- McCarthy OJ (2019) AI & global governance: turning the tide on crime with predictive policing. Centre for Policy Research, United Nations University. <https://cpr.unu.edu/publications/articles/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html>. Accessed 4 May 2022
- Microsoft, Article One (2018) Human rights impact assessment (HRIA) of the human rights risks and opportunities related to artificial intelligence (AI). <https://www.articleoneadvisors.com/case-studies-microsoft>. Accessed 1 May 2022
- Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Mugari I, Obioha EE (2021) Predictive policing and crime control in the United States of America and Europe: trends in a decade of research and the future of predictive policing. Soc Sci 10:234. <https://doi.org/10.3390/socsci10060234>
- Muller C (2020) The impact of artificial intelligence on human rights, democracy and the rule of law. Ad Hoc Committee on Artificial Intelligence (CAHAI), Council of Europe, Strasbourg. <https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da>. Accessed 2 May 2022
- PwC (2019) A practical guide to responsible artificial intelligence. <https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf>. Accessed 18 June 2020
- Reisman D, Schultz J, Crawford K, Whittaker M (2018) Algorithmic impact assessments: a practical framework for public agency accountability. AI Now Institute, New York. <https://ainowinstitute.org/aiareport2018.pdf>. Accessed 18 June 2020
- Reuters (2016) Passport robot tells Asian man his eyes are closed. New York Post, 7 Dec. <https://nypost.com/2016/12/07/passport-robot-tells-asian-man-his-eyes-are-closed/>. Accessed 4 May 2022
- Reuters (2018) Amazon ditched AI recruiting tool that favored men for technical job. The Guardian, 11 Oct. <https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-generator-bias-recruiting-engine>. Accessed 4 May 2022
- Stahl BC (2021) Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies. Springer Nature Switzerland AG, Cham. <https://doi.org/10.1007/978-3-030-69978-9>
- UK AI Council (2021) AI roadmap. Office for Artificial Intelligence, London. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/949539/AI\\_Council\\_AI\\_Roadmap.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949539/AI_Council_AI_Roadmap.pdf)
- UN (1948) Universal declaration of human rights. <http://www.un.org/en/universal-declaration-human-rights/>. Accessed 4 May 2022
- UNESCO (2020) First draft text of the recommendation on the ethics of artificial intelligence, 7 Sept. Ad hoc expert group (AHEG) for the preparation of a draft text, UNESCO, Paris. <https://unesdoc.unesco.org/ark:/48223/pf0000373434>. Accessed 12 Oct 2020
- van den Hoven J (2013) Value sensitive design and responsible innovation. In: Owen R, Heintz M, Bessant J (eds) Responsible innovation. Wiley, Chichester, pp 75–84
- Veale M, Binns R (2017) Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. Big Data Soc 4(2). <https://doi.org/10.1177/2053951717743530>
- Wexler R (2017a) Code of silence. Washington Monthly, 11 June. <https://washingtonmonthly.com/2017a/06/11/code-of-silence/>. Accessed 4 May 2022

- Wexler R (2017b) When a computer program keeps you in jail. *The New York Times*, 13 June. <https://www.nytimes.com/2017b/06/13/opinion/how-computers-are-harming-criminal-justice.html>. Accessed 4 May 2022
- Wright D (2011) A framework for the ethical impact assessment of information technology. *Ethics Inf Technol* 13:199–226. <https://doi.org/10.1007/s10676-010-9242-6>
- Zheng Y, Walsham G (2021) Inequality of what? An intersectional approach to digital inequality under Covid-19. *Inf Organ* 31:100341. <https://doi.org/10.1016/j.infoandorg.2021.100341>
- Zou J, Schiebinger L (2018) AI can be sexist and racist: it's time to make it fair. *Nature* 559:324–326. <https://doi.org/10.1038/d41586-018-05707-8>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# Chapter 3

## Privacy



**Abstract** Privacy and data protection are concerns raised about most digital technologies. The advance of artificial intelligence (AI) has given even higher levels of prominence to these concerns. Three cases are presented as examples to highlight the way in which AI can affect or exacerbate privacy concerns. The first deals with the use of private data in authoritarian regimes. The second looks at the implications of AI use of genetic data. The third concerns problems linked to biometric surveillance. Then follows a description of how privacy concerns are currently addressed via data protection regulation and a discussion of where AI may raise new challenges to existing data protection regimes. Current European data protection law requires data protection impact assessment. This chapter suggests that a broader AI impact assessment could broaden the remit of such an assessment to offer more comprehensive coverage of possible privacy concerns linked to AI.

**Keywords** Privacy · Data protection · Social credit · Data misuse · Authoritarian government · Genetic data · Biometrics · Surveillance

### 3.1 Introduction

Concerns about the possible negative impact of artificial intelligence (AI) on privacy have been widely expressed (EDPS 2020). Not all AI applications use personal data and therefore some uses may not have any privacy implications. However, the need for large datasets for the training and validation of machine learning models can raise a range of different concerns. Privacy is a complex concept that we return to in more detail below. Key to the discussion of privacy in AI is the worry that the use of AI technologies can lead to the violation of data protection principles, which then leads to harm for specific individuals or groups whose data is analysed using AI.

Privacy and data protection are issues that apply to most digital technologies, including AI. It is possible for most personal data to be misused for purposes that breach data protection principles or violate legitimate privacy preferences unless appropriate safeguards are in place. An important legal recognition of the “right to privacy” based on legitimate privacy preferences was the first, expressed in the

nineteenth century (Warren and Brandeis 1890). The stipulated “right to be let alone” as described by Warren and Brandeis was driven by a key technical innovation of the time, namely the ability to take photographs of individuals. This new technology raised concerns that had previously been immaterial when capturing the likeness of a person required them to sit down in front of a painter for extended periods.

Ever since the nineteenth century, data protection regulation and legislation have developed in tandem with new technical capabilities and resulting threats to privacy. The growing ability to process data through electronic computers led to much academic debate on the topic and the development of so-called principles of fair information practices (Severson 1997). These were originally developed in the US in 1973. They still underpin much of our thinking on data protection today. The principles include that

1. individuals should have the right to know how organizations use personal information and to inspect their records and correct any errors;
2. individuals should have the right to prevent secondary use of personal information if they object to such use; and
3. organizations that collect or use personal information must take reasonable precautions to prevent misuse of the information. (Culnan 1993: 344)

These principles have contributed to the creation of legislation and shaped its content since the 1970s and 1980s. At the European level, Directive 95/46/EC established a shared approach and visible data protection principles in 1995. It was superseded by the General Data Protection Regulation (GDPR) (European Parliament and Council of the EU 2016), which came into effect in 2018.

Given that AI is not the first potential threat to privacy or data protection, it is worth asking why the impacts of AI technologies on privacy are often seen as key ethical concerns. One part of the answer is that machine learning allows the development of fine-grained categories of data which, in turn, can be used to categorise and profile individuals. Such profiling may well be the intended result of AI use, for instance when an organisation seeks to identify potential customers to target with advertising campaigns. Such profiling may also have discriminatory effects as outlined in Chap. 2. It may also have other undesirable consequences for individuals or groups and open the way to misuse, such as when consumer profiles are used for political purposes (see Chap. 5).

AI uses of personal data can furthermore facilitate surveillance far beyond the capabilities that existed prior to AI. This includes automated surveillance of individuals using their biometric data, for example employing facial recognition, as developed in more detail in the cases below. There may be good reasons for the development and employment of such surveillance, as well as morally desirable outcomes, for instance the prevention of gender-based violence. But AI-based surveillance may also have undesired outcomes. The key challenge is that data protection is a moral value that must be balanced against other moral values. This is important to keep in mind from a moral perspective, especially because data protection is strongly regulated whereas other ethical issues and possible moral advantages are typically not

subject to the same level of regulation. The following cases of privacy violations that are enabled by AI demonstrate this point.

## 3.2 Cases of Privacy Violations Through AI

### 3.2.1 Case 1: Use of Personal Data by Authoritarian Regimes

China is one of the world's leading nations in AI development. It embraces the use of large amounts of data that it collects on its citizens, for instance in its social credit scoring system. This system uses a large number of data points, including social media data, local government data and citizens' activities, to calculate a trustworthiness score for every citizen. Several data platforms are used to integrate data into "a state surveillance infrastructure" (Liang et al. 2018). High scores lead to the allocation of benefits, such as lower utility rates and favourable booking conditions, whereas low scores can lead to the withdrawal of services (Raso et al. 2018). Within China, the system benefits from high levels of approval because Chinese citizens "interpret it through frames of benefit-generation and promoting honest dealings in society and the economy instead of privacy-violation" (Kostka 2019: 1565).

All states collect information about their citizens for a broad range of purposes. Some of these purposes may enjoy strong support from citizens, such as the allocation of financial support or healthcare, while others may be less popular, such as tax collection. Authoritarian governments can make additional use of data on their citizens to stabilise their power base (Liu 2019). A case in point is China, even though research has shown that Chinese citizens interpret the system from the perspective of its benefits.

It has also been argued that China has strong data protection laws. However, these do not apply to state bodies (Gal 2020), and government use of data for schemes such as social credit scoring are therefore not covered. This differs from the situation in Europe, where data protection law is binding on governments and state bodies as well. Social credit scoring is contentious. However, it is not always too different from activities such as "nudging" that democratic governments use, for example to encourage healthy behaviour such as giving up smoking or taking up exercise (Benartzi et al. 2017).

Both nudging and social credit scoring are contested, though one can see arguments in their favour. But the use of AI for the repression of citizens can go far beyond these. China, for example, has been reported to use AI to track behaviour that is deemed suspicious, such as religious speech by its Uighur population (Andersen 2020). Uighur community members who live in the Xianjiang area are subject to intrusive data collection and analysis that checks not only whether they exchange religious texts, but also where they live, their movement patterns, their pregnancy status and much more. The intention behind this data collection is ostensibly to

strengthen the state's control of the Uighur community. China's human rights record, in particular with regard to the Uighurs, suggests that this use of data and AI analysis is likely to result in further reduction of freedoms and limiting of human rights. By employing AI, authoritarian regimes may find it easier to analyse large amounts of data, such as social media posts, and to identify contributions that can trigger government responses.

### 3.2.2 Case 2: Genetic Privacy

Many genetic programmes are hailed for delivering medical breakthroughs via personalised medicine and the diagnosis of hereditary diseases. For instance, the Saudi Human Genome Program (SHGP), launched by the Saudi King in 2013, was announced with such aims (Alrefaei et al. 2022). Research showed that "90.7% of [Saudi] participants agreed that AI could be used in the SHGP" (*ibid*). However, the same research showed "a low level of knowledge ... regarding sharing and privacy of genetic data" (*ibid*), pointing to a potential mismatch of awareness of the benefits as opposed to the risks of genetic research supported by AI.

Genetic data is data that can provide deep insights into medical conditions, but also regarding possible risks and propensities for diseases that can go beyond other types of data. It thus has the properties of medical data and is therefore subject to stronger data protection regimes as part of a special category of data in many jurisdictions. Yet the importance and potential of genetic data goes beyond its medical uses. Genetic data of one person can provide information about their human heritage, their ancestors and their offspring. Access to genetic data can therefore present benefits as well as risks and entail a multitude of ethical issues. For instance, genomic datasets can improve research on cancer and rare diseases, while the reidentification of even anonymised data risks serious privacy concerns for the families involved (Takashima et al. 2018).

With the costs of gene sequencing continuing to fall, one can reasonably expect genetic data to become part of routine healthcare within a decade. This raises questions about data governance, storage, security etc. Such genetic data requires Big Data analytics approaches typically based on some sort of AI in order to be viable and provide relevant scientific or diagnostic insights.

In addition to the use of genetic data in healthcare, there is a growing number of private providers, such as 23andMe, Ancestry and Veritas Genetics (Rosenbaum 2018) that offer gene sequencing services commercially. This raises further questions around the ownership of data and the security of these companies, and creates uncertainty about the use of data should such a company go bankrupt or be bought out.

Addressing ethical concerns can lead to unpleasant surprises, for example when a genetic analysis contradicts assumed relationships in a family, proving that someone's

ancestry is not as had been supposed. In some cases this may be greeted with humour or mild embarrassment, but in others, where ancestry is crucial to the legitimacy of a social position, evidence of this kind may have manifestly negative consequences. Such consequences, it could be argued, are part of the nature of genetic data and should be dealt with via appropriate information and consent procedures. However, it is in the nature of genetic data that it pertains to more than one individual. If a sibling, for example, undertakes a genetic analysis, then many of the findings will be relevant to other family members. If such an analysis shows, for instance, that a parent is carrying a gene that contributes to a disease, other siblings' propensities to develop this disease would likely be increased as well, even though they did not take a genetic test themselves. This example demonstrates the possible conflicts arising from possessing and sharing such information.

AI analysis of genetic data may lead to medical insights. Indeed, this is the assumption that supports the business model of private gene-sequencing organisations. Their work is built on the assumption that collecting large amounts of genetic data in addition to other data that their customers provide will allow them to identify genetic patterns that can help predict or explain diseases. This, in turn, opens the way to medical research and finding cures, potentially a highly lucrative business.

From an ethical perspective this is problematic because the beneficiaries of this data analysis will normally be the companies, whereas the individual data subjects or donors will at best be notified of the insights their data has contributed to. Another concern is that the analysis may lead to the ability to predict disease trajectories without being able to intervene or cure (McCusker and Loy 2017), thus forcing patients to face difficult decisions involving complex probabilities that most non-experts are poorly equipped to deal with.

A further concern is that of mission creep, where the original purpose of the data collection is replaced by a changing or altogether different use. One obvious example is the growing interest from law enforcement agencies in gaining access to more genetic data so that they can, for example, identify culprits through genetic fingerprinting. The main point is that data, once it exists in digital form, is difficult to contain. Moor (2000) uses the metaphor of grease in an internal combustion engine. Data, once in an electronically accessible format, is very difficult to remove, just like grease in an engine. It may end up in unexpected places, and attempts to delete it may prove futile. In the case of genetic data this raises problems of possible future, and currently unpredicted, use, which, due to the very personal nature of the data, may have significant consequences that one currently cannot predict.

The Saudi case is predicated on the assumption of beneficial outcomes of the sharing of genetic data, and so far there is little data to demonstrate whether and in what way ethical issues have arisen or are likely to arise. A key concern here is that due to the tendency of data to leak easily, waiting until ethical concerns have materialised before addressing them is unlikely to be good enough. At that point the genie will be out of the bottle and the “greased” data may be impossible to contain.

### 3.2.3 Case 3: Biometric Surveillance

“Nijeer Parks is the third person known to be arrested for a crime he did not commit based on a bad face recognition match” (Hill 2020). Parks was falsely accused of stealing and trying to hit a police officer with his car based on facial recognition software – but he was 30 miles away at the time. “Facial recognition ... [is] very good with white men, very poor on Black women and not so great on white women, even” (Balli 2021). It becomes particularly problematic when “the police trust the facial recognition technology more than the individual” (*ibid*).

Biometric surveillance uses data about the human body to closely observe or follow an individual. The most prominent example of this is the use of facial features in order to track someone. In this broad sense of the term, any direct observation of a person, for example a suspected criminal, is an example of biometric surveillance. The main reason why biometric surveillance is included in the discussion of privacy concerns is that AI systems allow an enormous expansion of the scope of such surveillance. Whereas in the past one observer could only follow one individual, or maybe a few, the advent of machine learning and image recognition techniques, coupled with widespread image capture from closed-circuit television cameras, allows community surveillance. Automatic face recognition and tracking is not the only possible example of biometric surveillance, but it is the one that is probably most advanced and raises most public concern relating to privacy, as in the case described above.

There are numerous reasons why biometric surveillance is deemed to be ethically problematic. It can be done without the awareness of the data subject and thus lead to the possibility and the perception of pervasive surveillance. While some might welcome pervasive surveillance as a contribution to security and the reduction of crime, it has been strongly argued that being subject to it can lead to significant harm. Brown (2000), drawing on Giddens (1984) and others, argues that humans need a “protective cocoon” that shields from external scrutiny. This is needed to develop a sense of “ontological security”, a condition for psychological and mental health. Following this argument, pervasive surveillance is ethically problematic, simply for the psychological damage it can do through its very existence. Surveillance can lead to self-censoring and “social cooling” (Schep n.d.), that is, a modification of social interaction caused by fear of possible sanctions. AI-enabled large-scale biometric surveillance could reasonably be expected to lead to this effect.

## 3.3 Data Protection and Privacy

The above three case studies show that the analysis of personal data through AI systems can lead to significant harms. AI is by far not the only threat to privacy, but it adds new capabilities that can either exacerbate existing threats, for example by automating mass surveillance based on biometric data, or add new angles to privacy



**Fig. 3.1** Seven types of privacy

concerns, for example by exposing new types of data, such as genetic data, to the possibility of privacy violations.

Before we look at what is already being done to address these concerns and what else could be done, it is worth providing some more conceptual clarity. The title of this chapter and the headlines covering much of the public debate on the topics raised here refer to “privacy”. As suggested at the beginning of this chapter, however, privacy is a broad term that covers more than the specific aspects of AI-enabled analysis of personal data.

A frequently cited categorisation of privacy concepts (Finn et al. 2013: 7) proposes that there are seven types of privacy: privacy of the person, privacy of behaviour and action, privacy of personal communication, privacy of data and image, privacy of thoughts and feelings, privacy of location and space, and privacy of association (including group privacy) (see Fig. 3.1).

Most of these types of privacy have can be linked to data, but they go far beyond simple measures of data protection. Nissenbaum (2004) suggests that privacy can be understood as contextual integrity. This means that privacy protection must be context-specific and that information gathering needs to conform to the norms of the context. She uses this position to argue against public surveillance.

It should thus be clear that privacy issues cannot be comprehensively resolved by relying on formal mechanisms of data protection governance, regulation and/or legislation. However, data protection plays a crucial role in and is a necessary condition of privacy preservation. The application of data protection principles to AI raises several questions. One relates to the balance between the protection of personal data and the openness of data for novel business processes, where it has been argued that stronger data protection rules, such as the EU’s GDPR (European Parliament and Council of the EU 2016), can lead to the weakening of market positions in the race for AI dominance (Kaplan and Haenlein 2019). On the other hand, there are worries that current data protection regimes may not be sufficient in their coverage to deal with novel privacy threats arising from AI technologies and applications (Veale et al. 2018).

A core question which has long been discussed in the broader privacy debate is whether privacy is an intrinsic or an instrumental value. Intrinsic values are those values that are important in themselves and need no further justification. Instrumental values are important because they lead to something that is good (Moor 2000). The distinction is best known in environmental philosophy, where some argue that an

intact natural environment has an intrinsic value while others argue that it is solely needed for human survival or economic reasons (Piccolo 2017).

However, this distinction may be simplistic, and the evaluation of a value may require attention to both intrinsic and instrumental aspects (Sen 1988). For our purposes it is important to note that the question whether privacy is an intrinsic or instrumental value has a long tradition (Tavani 2000). The question is not widely discussed in the AI ethics discourse, but the answer to it is important in determining the extent to which AI-related privacy risks require attention. The recognition of privacy as a fundamental right, for example in the European Charter of Fundamental Rights (European Union 2012), settles this debate to some degree and posits privacy as a fundamental right worthy of protection (AI HLEG 2020). However, even assuming that privacy is an unchanging human right, technology will affect how respect for privacy is shown (Buttarelli 2017). AI can also raise novel threats to privacy, for example by making use of emotion data that do not fit existing remedies (Dignum 2019).

Finally, like most other fundamental rights, privacy is not an absolute right. Personal privacy finds its limits when it conflicts with other basic rights or obligations, for example when the state compiles data in order to collect taxes or prevent the spread of diseases. The balancing of privacy against other rights and obligations therefore plays an important role in finding appropriate mitigations for privacy threats.

### 3.4 Responses to AI-Related Privacy Threats

We propose two closely related responses to AI-related privacy threats: data protection impact assessments (DPIAs) and AI impact assessments (AI-IAs).

DPIAs developed from earlier privacy impact assessments (Clarke 2009; ICO 2009). They are predicated on the idea that it is possible to proactively identify possible issues and address them early in the development of a technology or a socio-technical system. This idea is widespread and there are numerous types of impact assessment, such as environmental impact assessments (Hartley and Wood 2005), social impact assessments (Becker and Vanclay 2003) and ethics impact assessments (Wright 2011). The choice of terminology for DPIAs indicates a recognition of the complexity of the concept of privacy and a consequently limited focus on data protection only. DPIAs are mandated in some cases under the GDPR. As a result of this legal requirement, DPIAs have been widely adopted and there are now well-established methods that data controllers can use.

#### Data Controllers and Data Processors

The concept of a data controller is closely linked to the GDPR, where it is defined as the organisation that determines the purposes for which and the means by which

personal data is processed. The data controller has important responsibilities with regard to the data they control and is normally liable when data protection rules are violated. The data processor is the organisation that processes personal data on behalf of the data controller. This means that data controller and data processor have clearly defined tasks, which are normally subject to a contractual agreement. An example might be a company that analyses personal data for training a machine learning system. This company, because it determines the purpose and means of processing, is the controller. It may store the data on a cloud storage system. The organisation running the cloud storage could then serve as data processor (EDPS 2019; EDPB 2020).

In practice, DPIAs are typically implemented in the form of a number of questions that a data controller or data processor has to answer, in order to identify the type of data and the purpose and legal basis of the data processing, and to explore whether the mechanisms in place to protect the data are appropriate to the risk of data breaches (ICO n.d.). The risk-based approach that underlies DPIAs, or at least those undertaken in response to the GDPR, shows that data protection is not a static requirement but must be amenable to the specifics of the context of data processing. This can raise questions when AI is used for data processing, as the exact uses of machine learning models may be difficult to predict, or where possible harms would not target the individual data subject but may occur at a social level, for instance when groups of the population are stigmatised because of characteristics that are manifest in their personal data. An example might be a healthcare system that identifies a correlation between membership of an ethnic group and propensity to a particular disease. Even though this says nothing about causality, it could nevertheless lead to prejudice against members of the ethnic group.

We have therefore suggested that a broader type of impact assessment is more appropriate for AI, one that includes questions of data protection but also looks at other possible ethical issues in a more structured way. Several such AI-IAs have been developed by various institutions. The most prominent was proposed by the EU's High-Level Expert Group in its Assessment List for Trustworthy AI, or ALTAI (AI HLEG 2020). Other examples are the AI Now Institute's algorithmic impact assessment (Reisman et al. 2018), the IEEE's recommended practice for assessing the impact of autonomous and intelligent systems on human wellbeing (IEEE 2020) and the ECP Platform's artificial intelligence impact assessment (ECP 2019).<sup>1</sup>

What all these examples have in common is that they broaden the idea of an impact assessment for AI to address various ethical issues. They all cover data protection questions but go beyond them. This means that they may deal with questions of long-term or large-scale use of AI, such as economic impact or changes in democratic norms, that go beyond the protection of individual personal data. In fact, there are several proposals that explicitly link AI-IAs and DPIAs or that focus in particular on the data protection aspect of an AI-IA (Government Digital Service 2020; Ivanova

<sup>1</sup> A collection of recent AI impact assessments can be accessed via a public Zotero group at [https://www.zotero.org/groups/4042832/ai\\_impact\\_assessments](https://www.zotero.org/groups/4042832/ai_impact_assessments).

2020; ICO 2021). An AI-IA should therefore not be seen as a way to replace a DPIA, but rather as supplementing and strengthening it.

### 3.5 Key Insights

Privacy remains a key concern in the AI ethics debate and this chapter has demonstrated several ways in which AI can cause harm, based on the violation of data protection principles. Unlike other aspects of the AI ethics debate, privacy is recognised as a human right, and data protection, as a means of supporting privacy, is extensively regulated. As a result of this high level of attention, there are well-established mechanisms, such as DPIAs, which can easily be extended to cover broader AI issues or incorporated into AI-IAs.

The link between DPIAs and AI-IAs can serve as an indication of the role of data and data protection as a foundational aspect of many other ethical issues. Not all of AI ethics can be reduced to data protection. However, many of the other issues discussed in this book have a strong link to personal data. Unfair discrimination, for example, typically requires and relies on personal data on the individuals who are discriminated against. Economic exploitation in surveillance capitalism is based on access to personal data that can be exploited for commercial purposes. Political and other types of manipulation require access to personal data to identify personal preferences and propensities to react to certain stimuli. Data protection is thus a key to many of the ethical issues of AI, and our suggested remedies are therefore likely to be relevant across a range of issues. Many of the responses to AI ethics discussed in this book will, in turn, touch on or incorporate aspects of data protection.

This does not imply, however, that dealing with privacy and data protection in AI is easy or straightforward. The responses that we suggest here, i.e. DPIAs and AI-IAs, are embedded in the European context, in which privacy is recognised as a human right and data protection has been codified in legislation. It might be challenging to address such issues in the absence of this societal and institutional support. Our Case 1 above, which describes the use of data by an authoritarian regime, is a reminder that state and government-level support for privacy and data protection cannot be taken for granted.

Another difficulty lies in the balancing of competing goods and the identification of the boundaries of what is appropriate and ethically defensible. We have mentioned the example of using AI to analyse social media to identify cases of religious speech that can be used to persecute religious minorities. The same technology can be used to search social media in a different institutional context to identify terrorist activities. These two activities may be technically identical, though they are subject to different interpretations. This raises non-trivial questions about who determines what constitutes an ethically legitimate use of AI and where the boundaries of that use are, and on what grounds such distinctions are drawn. This is a reminder that AI ethics can rarely be resolved simply, but needs to be interpreted from a broader perspective

that includes a systems view of the AI application and considers institutional and societal aspects when ethical issues are being evaluated and mitigated.

## References

- AI HLEG (2020) The assessment list for trustworthy AI (ALTAI). High-level expert group on artificial intelligence. European Commission, Brussels. [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=68342](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342). Accessed 10 Oct 2020
- Alrefaei AF, Hawsawi YM, Almaleki D et al (2022) Genetic data sharing and artificial intelligence in the era of personalized medicine based on a cross-sectional analysis of the Saudi human genome program. *Sci Rep* 12:1405. <https://doi.org/10.1038/s41598-022-05296-7>
- Andersen R (2020) The panopticon is already here. *The Atlantic*, Sept. <https://www.theatlantic.com/magazine/archive/2020/09/china-ai-surveillance/614197/>. Accessed 7 May 2022
- Balli E (2021) The ethical implications of facial recognition technology. ASU News, 17 Nov. <https://news.asu.edu/20211117-solutions-ethical-implications-facial-recognition-technology>. Accessed 7 May 2022
- Becker HA, Vanclay F (eds) (2003) The international handbook of social impact assessment: conceptual and methodological advances. Edward Elgar Publishing, Cheltenham
- Benartzi S, Besears J, Mlikman K et al (2017) Governments are trying to nudge us into better behavior. Is it working? *The Washington Post*, 11 Aug. <https://www.washingtonpost.com/news/wonk/wp/2017/08/11/governments-are-trying-to-nudge-us-into-better-behavior-is-it-working/>. Accessed 1 May 2022
- Brown WS (2000) Ontological security, existential anxiety and workplace privacy. *J Bus Ethics* 23:61–65. <https://doi.org/10.1023/A%3A1006223027879>
- Buttarelli G (2017) Privacy matters: updating human rights for the digital society. *Health Technol* 7:325–328. <https://doi.org/10.1007/s12553-017-0198-y>
- Clarke R (2009) Privacy impact assessment: its origins and development. *Comput Law Secur Rev* 25:123–135. <https://doi.org/10.1016/j.clsr.2009.02.002>
- Culnan M (1993) “How did they get my name?” An exploratory investigation of consumer attitudes toward secondary information use. *MIS Q* 17(3):341–363. <https://doi.org/10.2307/249775>
- Dignum V (2019) Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer Nature Switzerland AG, Cham
- ECP (2019) Artificial intelligence impact assessment. ECP Platform for the Information Society, The Hague. <https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf>. Accessed 1 May 2022
- EDPB (2020) Guidelines 07/2020 on the concepts of controller and processor in the GDPR. European Data Protection Board, Brussels. [https://edpb.europa.eu/sites/default/files/consultation/edpb\\_guidelines\\_202007\\_controllerprocessor\\_en.pdf](https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202007_controllerprocessor_en.pdf). Accessed 8 May 2022
- EDPS (2019) EDPS guidelines on the concepts of controller, processor and joint controllership under regulation (EU) 2018/1725. European Data Protection Supervisor, Brussels
- EDPS (2020) EDPS opinion on the European Commission’s white paper on artificial intelligence: a European approach to excellence and trust (Opinion 4/2020). European Data Protection Supervisor, Brussels. [https://edps.europa.eu/data-protection/our-work/publications/opinions/edps-opinion-european-commissions-white-paper\\_en](https://edps.europa.eu/data-protection/our-work/publications/opinions/edps-opinion-european-commissions-white-paper_en). Accessed 6 May 2022
- European Parliament, Council of the EU (2016) Regulation (EU) 2016/679 of the European Parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official J Eur Union L119(11):1–88. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN>. Accessed 1 May 2022.

- European Union (2012) Charter of fundamental rights of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:C2012/326/02&from=EN>. Accessed 1 May 2022
- Finn RL, Wright D, Friedewald M (2013) Seven types of privacy. In: Gutwirth S, Leenes R, de Hert P, Poulett Y (eds) European data protection: coming of age. Springer, Dordrecht, pp 3–32
- Gal D (2020) China's approach to AI ethics. In: Elliott H (ed) The AI powered state: China's approach to public sector innovation. Nesta, London, pp 53–62
- Giddens A (1984) The constitution of society: outline of the theory of structuration. Polity, Cambridge
- Government Digital Service (2020) Data ethics framework. Central Digital and Data Office, London. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/923108/Data\\_Ethics\\_Framework\\_2020.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/923108/Data_Ethics_Framework_2020.pdf). Accessed 1 May 2022
- Hartley N, Wood C (2005) Public participation in environmental impact assessment: implementing the Aarhus convention. Environ Impact Assess Rev 25:319–340. <https://doi.org/10.1016/j.eiar.2004.12.002>
- Hill K (2020) Another arrest, and jail time, due to a bad facial recognition match. The New York Times, 29 Dec. <https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html>. Accessed 7 May 2022
- ICO (2009) Privacy impact assessment handbook, v. 2.0. Information Commissioner's Office, Wilmslow. <https://www.huntonprivacyblog.com/wp-content/uploads/sites/28/2013/09/PIAhandbookV2.pdf>. Accessed 6 Oct 2020
- ICO (2021) AI and data protection risk toolkit beta. Information Commissioner's Office, Wilmslow
- ICO (n.d.) Data protection impact assessments. Information Commissioner's Office, Wilmslow. <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/>. Accessed 8 May 2022
- IEEE (2020) 7010-2020: IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being. IEEE Standards Association, Piscataway. <https://doi.org/10.1109/IEEEESTD.2020.9084219>
- Ivanova Y (2020) The data protection impact assessment as a tool to enforce non-discriminatory AI. In: Antunes L, Naldi M, Italiano GF et al (eds) Privacy technologies and policy. 8th Annual privacy forum, APF 2020, Lisbon, Portugal, 22–23 Oct. Springer Nature Switzerland, Cham, pp 3–24. [https://doi.org/10.1007/978-3-030-55196-4\\_1](https://doi.org/10.1007/978-3-030-55196-4_1)
- Kaplan A, Haenlein M (2019) Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz 62:15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>
- Kostka G (2019) China's social credit systems and public opinion: explaining high levels of approval. New Media Soc 21:1565–1593. <https://doi.org/10.1177/1461444819826402>
- Liang F, Das V, Kostyuk N, Hussain MM (2018) Constructing a data-driven society: China's social credit system as a state surveillance infrastructure. Policy Internet 10:415–453. <https://doi.org/10.1002/poi3.183>
- Liú C (2019) Multiple social credit systems in China. Social Science Research Network, Rochester
- McCusker EA, Loy CT (2017) Huntington disease: the complexities of making and disclosing a clinical diagnosis after premanifest genetic testing. Tremor Other Hyperkinet Mov (NY) 7:467. <https://doi.org/10.7916/D8PK0TDD>
- Moor JH (2000) Toward a theory of privacy in the information age. In: Baird RM, Ramsower RM, Rosenbaum SE (eds) Cyberethics: social and moral issues in the computer age. Prometheus, Amherst, pp 200–212
- Nissenbaum H (2004) Symposium: privacy as contextual integrity. Wash Law Rev 79:119–158. <https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=4450&context=wlr>. Accessed 2 May 2022
- Piccolo JJ (2017) Intrinsic values in nature: objective good or simply half of an unhelpful dichotomy? J Nat Conserv 37:8–11. <https://doi.org/10.1016/j.jnc.2017.02.007>

- Raso FA, Hilligoss H, Krishnamurthy V et al (2018) Artificial intelligence & human rights: opportunities & risks. Berkman Klein Center Research Publication No. 2018-6. <https://doi.org/10.2139/ssrn.3259344>
- Reisman D, Schultz J, Crawford K, Whittaker M (2018) Algorithmic impact assessments: a practical framework for public agency accountability. AI Now Institute, New York. <https://ainowinstitute.org/aiareport2018.pdf>. Accessed 18 June 2020
- Rosenbaum E (2018) 5 biggest risks of sharing your DNA with consumer genetic-testing companies. CNBC, 16 June. <https://www.cnbc.com/2018/06/16/5-biggest-risks-of-sharing-dna-with-consumer-genetic-testing-companies.html>. Accessed 7 May 2022
- Schep T (n.d.) Social cooling. <https://www.tijmenschep.com/socialcooling/>. Accessed 8 May 2022
- Sen A (1988) On ethics and economics, 1st edn. Wiley-Blackwell, Oxford
- Severson RW (1997) The principles for information ethics, 1st edn. Routledge, Armonk
- Takahshima K, Maru Y, Mori S et al (2018) Ethical concerns on sharing genomic data including patients' family members. BMC Med Ethics 19:61. <https://doi.org/10.1186/s12910-018-0310-5>
- Tavani H (2000) Privacy and security. In: Langford D (ed) Internet ethics, 2000th edn. Palgrave, Basingstoke, pp 65–95
- Veale M, Binns R, Edwards L (2018) Algorithms that remember: model inversion attacks and data protection law. Phil Trans R Soc A 376:20180083. <https://doi.org/10.1098/rsta.2018.0083>
- Warren SD, Brandeis LD (1890) The right to privacy. Harv Law Rev 4(5):193–220. <https://doi.org/10.2307/1321160>
- Wright D (2011) A framework for the ethical impact assessment of information technology. Ethics Inf Technol 13:199–226. <https://doi.org/10.1007/s10676-010-9242-6>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# Chapter 4

## Surveillance Capitalism



**Abstract** Surveillance capitalism hinges on the appropriation and commercialisation of personal data for profit-making. This chapter spotlights three cases connected to surveillance capitalism: data appropriation, monetisation of health data and the unfair commercial practice when “free” isn’t “free”. It discusses related ethical concerns of power inequality, privacy and data protection, and lack of transparency and explainability. The chapter identifies responses to address concerns about surveillance capitalism and discusses three key responses put forward in policy and academic literature and advocated for their impact and implementation potential in the current socio-economic system: antitrust regulation, data sharing and access, and strengthening of data ownership claims of consumers/individuals. A combination of active, working governance measures is required to stem the growth and ill-effects of surveillance capitalism and protect democracy.

**Keywords** Antitrust · Big Tech · Data appropriation · Data monetisation · Surveillance capitalism · Unfair commercial practices

### 4.1 Introduction

The flourishing of systems and products powered by artificial intelligence (AI) and increasing reliance on them fuels surveillance capitalism and “ruling by data” (Pistor 2020). The main beneficiaries, it is argued, are ‘Big Tech’ – including Apple, Amazon, Alphabet, Meta, Netflix, Tesla (Jolly 2021) – who are seen to be entrenching their power over individuals and society and affecting democracy (*ibid*).

Simply described, surveillance capitalism hinges on the appropriation and commercialisation of personal data for profit. Zuboff (2015) conceptualised and defined it as a “new form of information capitalism [which] aims to predict and modify human behavior as a means to produce revenue and market control”. She argues that surveillance capitalism “effectively exile[s] persons from their own behavior while producing new markets of behavioral prediction and modification” (*ibid*). It is underpinned by organisational use of behavioural data leading to asymmetries in knowledge and power (Zuboff 2019b). As a result, consumers often may

not realise the extent to which they are responding to prompts driven by commercial interests.

Doctorow (2021) elaborates that surveillance capitalists engage in segmenting (targeting based on behaviour/attitudes/choices), deception (by making fraudulent claims, replacing beliefs with false or inaccurate ones) and domination (e.g. Google's dominance on internet searches and the monopolisation of the market through mergers and acquisitions). He outlines three reasons why organisations continue to over-collect and over-retain personal data: first, that they are competing with people's ability to resist persuasion techniques once they are wise to them, and with competitors' abilities to target their customers; second, that the cheapness of data aggregation and storage facilitates the organisation's acquiring an asset for future sales; and third, that the penalties are imposed for leaking data are negligible (Wolff and Atallah 2021).

Surveillance capitalism manifests in both the private and public sectors; organisations in both sectors collect and create vast reserves of personal data under various guises. Examples for the private sector can be found in online marketplaces (e.g. Amazon—Nash 2021), the social media industry (e.g. Facebook—Zuboff 2019a) and the entertainment industry (Bellanova and González Fuster 2018). The most cited examples of surveillance capitalism are Google's (Zuboff 2020) and Facebook's (Zuboff 2019a) attempts to feed their systems and services. In the public sector, this is very notable in the healthcare and retail sectors. During the COVID-19 pandemic, attention has been drawn to health-related surveillance capitalism. For example, it has been argued that telehealth technologies were pushed too quickly during the COVID-19 pandemic (Cosgrove et al. 2020; Garrett 2021).

## 4.2 Cases of AI-Enabled Surveillance Capitalism

### 4.2.1 Case 1: Data Appropriation

Clearview AI is a software company headquartered in New York which specialises in facial recognition software, including for law enforcement. The company, which holds ten billion facial images, aims to obtain a further 90 billion, which would amount to 14 photos of each person on the planet (Harwell 2022). In May 2021, legal complaints were filed against Clearview AI in France, Austria, Greece, Italy and the United Kingdom. It was argued that photos were harvested from services such as Instagram, LinkedIn and YouTube in contravention of what users of these services were likely to expect or have agreed to (Campbell 2021). On 16 December 2021, the French Commission Nationale de l'Informatique et des Libertés announced that it had “ordered the company to cease this illegal processing and to delete the data within two months” (CNIL 2021).

The investigations carried out by the Commission Nationale de l’Informatique et des Libertés (CNIL) into Clearview AI revealed two breaches of the General Data Protection Regulation (GDPR) (European Parliament and Council of the EU 2016): first, an unlawful processing of personal data (*ibid*: Article 6), as the collection and use of biometric data had been carried out without a legal basis; and second, a failure to take into account the rights of individuals in an effective and satisfactory way, especially with regard to access to their data (*ibid*: Articles 12, 15 and 17).

CNIL ordered Clearview AI to stop the collection and use of data of persons on French territory in the absence of a legal basis, to facilitate the exercise of individuals’ rights and to comply with requests for erasure. The company was given two months to comply with the injunctions and justify its compliance or face sanctions.<sup>1</sup>

This case is an example of data appropriation (i.e. the illegal, unauthorised or unfair taking or collecting of personal data for known or unknown purposes, without consent or with coerced or uninformed consent). Organisations appropriating data in such a fashion do not offer data subjects “comparable compensation”, while the organisations themselves gain commercial profits and other benefits from the activity.

#### **4.2.2 Case 2: Monetisation of Health Data**

In 2021, the personal data of 61 million people became publicly available without password protection, due to data leaks at a New York-based provider of health tracking services. The data included personal information such as names, gender, geographic locations, dates of birth, weight and height (Scropton 2021). Security researcher Jeremiah Fowler, who discovered the database, traced its origin to a company that offered devices and apps to track health and wellbeing data. The service users whose personal data had been leaked were located all over the world. Fowler contacted the company, which thanked him and confirmed that the data had now been secured (Fowler n.d.).

This case highlights issues with the collection and storage of health data on a vast scale by companies using health and fitness tracing devices, and it reveals the vulnerability of such data to threats and exposure.

A related concern is the procurement of such data by companies such as Google through their acquisition of businesses such as Fitbit, a producer of fitness monitors and related software. Experts (Bourreau et al. 2020) have indicated that such acquisition is problematic for various reasons: major risks of “platform envelopment”, the extension of monopoly power (by undermining competition) and consumer exploitation. Their concerns also relate to the serious harms that might result from Google’s ability to combine its own data with Fitbit’s health data.

---

<sup>1</sup> As of May 2022, the CNIL had not published any update regarding the company’s compliance or otherwise.

The European Commission carried out an in-depth investigation (European Commission 2020a) into the acquisition of Fitbit by Google. The concerns that emerged related to advertising, in that the acquisition would increase the already extensive amount of data that Google was able to use for the personalisation of ads, and the resulting difficulty for rivals setting out to match Google's services in the markets for online search advertising. It was argued that the acquisition would raise barriers to entry and expansion by Google's competitors, to the detriment of advertisers, who would ultimately face higher prices and have less choice. The European Commission approved the acquisition of Fitbit by Google under the EU Merger Regulation, conditional on full compliance with a ten-year commitments package offered by Google (European Commission 2020b).

#### 4.2.3 Case 3: Unfair Commercial Practices

In 2021, Italy's Consiglio di Stato (Council of State) agreed with the Autorità Garante della Concorrenza e del Mercato (Italian Competition Authority) and the Tribunale Amministrativo Regionale (Regional Administrative Tribunal) of the Lazio region to sanction Facebook for an unfair commercial practice. Facebook was fined seven million euros for misleading its users by not explaining to them in a timely and adequate manner, during the activation of their accounts, that data would be collected with commercial intent (AGCM 2021).

This case spotlights how companies deceive users into believing they are getting a social media service free of charge when this is not true at all (Coraggio 2021). The problem is aggravated by the companies not communicating well enough to users that their data is the quid pro quo for the use of the service and that the service is only available to them conditional on their making their data available and accepting the terms of service. What is also fuzzily communicated is how and to what extent companies put such data into further commercial use and take advantage of it for targeted advertising purposes. Commentators have gone so far as to state that the people using such services have themselves become the product (Oremus 2018).

### 4.3 Ethical Questions About Surveillance Capitalism

One of the primary ethical concerns that arise in the context of all three case studies is that of *power inequality*. The power of Big Tech is significant; it has even been compared to that of nation states (Andal 2016; Apostolicas 2019) and is further strengthened by the development and/or acquisition of AI solutions. All three of the cases examined have served to further and enhance the control that AI owners hold. The concentrated power that rests with a handful of big tech companies and the

control and influence they have, for example on political decision-making (Dubhashi and Lappin 2021), market manipulation and digital lives, are disrupting economic processes (Fernandez et al 2020) and posing a threat to democracy (Fernandez 2021), freedoms of individuals and political and social life.

Another key ethical concern brought to the forefront with these cases is *privacy and data protection* (see Chap. 3). Privacy is critical to human autonomy and well-being and helps individuals protect themselves from interference in their lives (Nass et al 2009). For instance, leaked personal health data might be appropriated by employers or health insurers and used against the interests of the person concerned. Data protection requires that data be processed in a lawful, fair and transparent manner in addition to being purpose-limited, accurate and retained for a limited time. It also requires that such processing respect integrity, confidentiality and accountability principles.

*Lack of transparency and explainability* links to data appropriation, data monetisation and unfair commercial practices. While it may seem obvious from a data protection and societal point of view that transparency requirements should be followed by companies that acquire personal data, this imperative faces challenges. Transparency challenges are a result of the structure and operations of the data industry. Transparency is also challenged by the appropriation of transparency values in public relations efforts by data brokers (e.g. Acxiom, Experian and ChoicePoint) to water down government regulation (Crain 2016). Crain (2016) also highlights that transparency may only create an “illusion of reform” and not address basic power imbalances.

Other ethical concerns relate to *proportionality and do-no-harm*. The UNESCO Recommendation on the Ethics of Artificial Intelligence, as adopted (UNESCO 2021), suggests:

- The AI method chosen should be appropriate and proportional to achieve a given legitimate aim.
- The AI method chosen should not infringe upon foundational values; its use must not violate or abuse human rights.
- The AI method should be appropriate to the context and should be based on rigorous scientific foundations.

In the cases examined above, there are clear-cut failures to meet these checks. “Appropriateness” refers to whether the technological or AI solution used is the best (with regard to cost and quality justifying any invasions of privacy), whether there is a risk of human rights, such as that to privacy, being abused and data being reused, and whether the objectives can be satisfied using other means. The desirability of the use of AI solutions is also something that should be duly considered— with regard to the purpose, advantages and burden imposed by them on social values, justice and the public interest.

A key aspect of the perception of surveillance capitalism as being ethically problematic seems to be the encroaching of market mechanisms on areas of life that previously were not subject to financial exchange. To some degree this is linked to the perception of exploitation of the data producers. Many users of “free” online services are content to use services such as social media or online productivity tools

in exchange for the use of their data by application providers. There is also, nevertheless, a perception of unfairness, as the service providers have been able to make gigantic financial gains that are not shared with the individuals on whose data they rely to generate these gains. In addition, criticism of surveillance capitalism seems to be based on the perception that some parts of social life should be free of market exchange. A manifestation of this may be the use of the term “friend” in social media, where not only does the nature of friendship differ substantially from that in the offline world, but the number of friends and followers can lead to financial transactions that would be deemed inappropriate in the offline world.

There is no single clearly identifiable ethical issue that is at the base of surveillance capitalism. The term should be understood as signifying opposition to technically enabled social changes that concentrate economic and political power in the hands of a few high-profile organisations.

## 4.4 Responses to Surveillance Capitalism

Many types of responses have been put forward to address concerns about surveillance capitalism: legal or policy-based responses, market-based responses, and societal responses. Legal and policy measures include antitrust regulation, intergovernmental regulation, strengthening the data-ownership claims of consumers or individuals, socialising the ownership of evolving technologies (Garrett 2021), making big tech companies spend their monopoly profits on governance (Doctorow 2021), mandatory disclosure frameworks (Andrew et al. 2021) and greater data sharing and access.

Market-based responses include placing value on the information provided to surveillance capitalists, monopoly reductions (Doctorow 2021), defunding Big Tech and refunding community-oriented services (Barendregt et al. 2021) and users employing their market power by rejecting and avoiding companies with perceived unethical behaviour (Jensen 2019).

Societal responses include indignation (Lyon 2019), naming /public indignation (Kavenna 2019), personal data spaces or emerging intermediary services that allow users control over the sharing and use of their data (Lehtiniemi 2017), increasing data literacy and awareness of how transparent a company’s data policy is, and improving consumer education (Lin 2018).

This section examines three responses which present promising ways to curtail the impact of surveillance capitalism and the ethical questions studied in a variety of ways, though none on its own is a silver bullet. These responses have been discussed in policy and academic literature and advocated for their impact and implementation potential in the current socio-economic system. The challenges of surveillance capitalism arise from the socio-political environment in which AI is developed and used, and the responses we have spotlighted here are informed by this.

#### ***4.4.1 Antitrust Regulation***

“Antitrust” refers to actions to control monopolies, prevent companies from working together to unfairly control prices, and enhance fair business competition. Antitrust laws regulate monopolistic behaviour and prevent unlawful business practices and mergers. Courts review mergers case by case for illegality. Many calls and proposals have been made to counter the power of Big Tech (e.g. Warren 2019). Discussions have proliferated on the use of antitrust regulations to break up big tech companies (The Economist 2019; Waters 2019) and the appointment of regulators to reverse illegal and anti-competitive tech mergers (Rodrigues et al 2020).

Big Tech is regarded as problematic for its concentration of power and control over the economy, society and democracy to the detriment of competition and innovation in small business (Warren 2019). Grunes and Stucke (2015) emphasise the need for competition and antitrust’s “integral role to ensure that we capture the benefits of a data-driven economy while mitigating its associated risks”. However, the use of antitrust remedies to control dominant firms presents some problems, such as reducing competitive incentives (by forcing the sharing of information) and innovation, creating privacy concerns or resulting in stagnation and fear among platform providers (Sokol and Comerford 2016).

There are both upsides and downsides to the use of antitrust regulation as a measure to curb the power of Big Tech (Zuboff 2021). The upsides include delaying or frustrating acquisitions, generating better visibility, transparency and oversight, pushing Big Tech to improve their practices, and better prospects for small businesses (Warren 2019). One downside is that despite the antitrust actions taken thus far, some Big Tech companies continue to grow their power and dominance (Swartz 2021). Another downside is the implementation and enforcement burdens antitrust places on regulators (Rodrigues et al. 2020). Furthermore, antitrust actions are expensive and disruptive to business and might affect innovation (The Economist 2019).

Developments (legislative proposals, acquisition challenges, lawsuits and fines) in the USA and Europe show that Big Tech’s power is under deeper scrutiny than ever before (Reuters 2021; Council of the EU 2021).

#### ***4.4.2 Data Sharing and Access***

Another response to surveillance capitalism concerns is greater data sharing and access (subject to legal safeguards and restrictions). Making data open and freely available under a strict regulatory environment is suggested as having the potential to better address the limitations of antitrust legislation (Leblond 2020). In similar vein, Kang (2020) suggests that data-sharing mandates (securely enforced through privacy-enhancing technologies) “have become an essential prerequisite for competition and innovation to thrive”; to counter the “monopolistic power derived from data, Big Tech should share what they know – and make this information widely usable for current and potential competitors” (*ibid*).

At the European Union level, the proposal for the Data Governance Act (European Commission 2020d) is seen as a “first building block for establishing a solid and fair data-driven economy” and “setting up the right conditions for trustful data sharing in line with our European values and fundamental rights” (European Commission 2021).

The Data Governance Act aims to foster the availability of data for more widespread use by increasing trust in data intermediaries and by strengthening data-sharing mechanisms across the EU. It specifies conditions for the reuse, within the European Union, of certain categories of data held by public sector bodies; a notification and supervisory framework for the provision of data sharing services; and a framework for the voluntary registration of entities which collect and process data made available for altruistic purposes.

The European Commission will also set out “a second major legislative initiative, the Data Act, to maximise the value of data for the economy and society” and “to foster data sharing among businesses, and between businesses and governments” (*ibid*). The proposed Digital Markets Act aims to lay down harmonised rules ensuring contestable and fair markets in the digital sector across the European Union. Gatekeepers will be present, and it is expected that business access to certain data will go through gatekeepers (European Commission 2020c).

#### ***4.4.3 Strengthening of Data Ownership Claims of Consumers/Individuals***

Another response to surveillance capitalism is to strengthen the data ownership claims of consumers and individuals. Jurcys et al. (2021) argue that even if user-held data is intangible, it meets all the requirements of an “asset” in property laws and that such “data is specifically defined, has independent economic value to the individual, and can be freely alienated” (*ibid*).

[T]he economic benefits property law type of entitlements over user-held data are superior over the set of data rights that are afforded by public law instruments (such as the GDPR or the CCPA) to individuals vis-a-vis third-party service providers who hold and benefit enormously from the information about individuals. (*ibid*)

Fadler and Legner (2021) also suggest that data ownership remains a key concept to clarify rights and responsibilities but should be revisited in the Big Data and analytics context. They identify three distinct types of data ownership – *data*, *data platform* and *data product ownership* – which may guide the definition of governance mechanisms and serve as the basis for more comprehensive data governance roles and frameworks.

As Hummel, Braun and Dabrock outline (2021), the commonality in calls for data ownership relates to modes of controlling how data is used and the ability to channel, constrain and facilitate the flow of data. They also suggest that with regard to the marketisation and commodification of data, ownership has turned out to be a

double-edged sword, and that using this concept requires reflection on how data subjects can protect their data and share appropriately. They furthermore outline that “even if legal frameworks preclude genuine ownership in data, there remains room to debate whether they can and should accommodate such forms of quasi-ownership” (Hummel et al. 2021).

Challenges that affect this response include the ambiguousness of the concept of ownership, the complexity of the data value cycle and the involvement of multiple stakeholders, as well as difficulty in determining who could or would be entitled to claim ownership in data (Van Asbroeck et al. 2019).

## 4.5 Key Insights

A combination of active, working governance measures is required to stem the growth and ill effects of surveillance capitalism and protect democracy. As we move forward, there are some key points that should be considered.

### *Breaking Up with Antitrust Regulation is Hard to Do*

While breaking up Big Tech using antitrust regulation might seem like a very attractive proposition, it is challenging and complex (Matsakis 2019). Moss (2019) assesses the potential consequences of breakup proposals and highlights the following three issues:

- Size thresholds could lead to broad restructuring and regulation.
- Breakup proposals do not appear to consider the broader dynamics created by prohibition on ownership of a platform and affiliated businesses.
- New regulatory regimes for platform utilities will require significant thought.

A report from the EU-funded SHERPA project (Rodrigues et al. 2020) also highlights the implementation burdens imposed on legislators (who must define the letter and scope of the law) and on enforcement authorities (who must select appropriate targets for enforcement action and make enforcement decisions.)

Further challenges include the limitations in antitrust enforcement officials’ knowledge and the potential impact of ill-advised investigations and prosecutions on markets (Cass 2013), never-ending processes, defining what conduct contravenes antitrust law (*ibid*), business and growth disruption, and high costs (The Economist 2019; Waters 2019), including the impact on innovation (Sokol and Comerford 2016).

### *Are “Big Tech’s Employees One of the Biggest Checks on Its Power”?*

Among the most significant actions that have changed the way digital companies behave and operate has been action taken by employees of such organisations to hold their employers to account over ethical concerns and illegal practices, while in the process risking career, reputation, credibility and even life. Ghaffary (2021) points out that tech employees are uniquely positioned (with their “behind the scenes” understanding of algorithms and company policies) to provide checks and enable the



**Fig. 4.1** Unveiled by whistleblowers

scrutiny needed to influence Big Tech. In the AI context, given issues of lack of transparency, this is significant for its potential to penetrate corporate veils.

A variety of issues have been brought to light by tech whistleblowers: misuse or illegal use of data (Cambridge Analytica) (Perrigo 2019), institutional racism, research suppression (Simonite 2021), suppression of the right to organise (Clayton 2021), the falsification of data, a lack of safety controls and the endangerment of life through hosting hate speech and illegal activity (Milmo 2021) (see Fig. 4.1).

Whistleblowing is now seen in the digital and AI context as a positive corporate governance tool (Brand 2020). Laws have been and are being amended to increase whistleblower protections, for instance in New York and the European Union. Reporting by whistleblowers to enforcement bodies is expected to increase as regulators improve enforcement and oversight over AI. This might provide a necessary check on Big Tech. However, whistleblowing comes with its own price, especially for the people brave enough to take this step, and by itself is not enough, given the resources of Big Tech and the high human and financial costs to the individuals who are forced to undertake such activity (Bridle 2018).

Surveillance capitalism may be here to stay, at least for a while, and its effects might be strong and hard in the short to medium term (and longer if not addressed), but as shown above, there are a plethora of mechanisms and tools to address it. In addition to the responses discussed in this chapter, it is important that other measures be duly reviewed for their potential to support ethical AI and used as required – be they market-based, policy- or law-based or societal interventions. Even more important is the need to educate and inform the public about the implications for and adverse effects on society of surveillance capitalism. This is a role that civil society organisations and the media are well placed to support.

## References

- AGCM (2021) IP330—Sanzione a Facebook per 7 milioni. Autorità Garante della Concorrenza e del Mercato, Rome. Press release, 17 Feb. <https://www.agcm.it/media/comunicati-stampa/2021/2/IP330->. Accessed 10 May 2022
- Andal S (2016) Tech giants' powers rival those of nation states. The Interpreter, 6 Apr. The Lowly Institute. <https://www.lowyinstitute.org/the-interpreter/tech-giants-powers-rival-those-nation-states>. Accessed 15 May 2022

- Andrew J, Baker M, Huang C (2021) Data breaches in the age of surveillance capitalism: do disclosures have a new role to play? *Crit Perspect Account*. <https://doi.org/10.1016/j.cpa.2021.102396>
- Apostolica P (2019) Silicon states: how tech titans are acquiring state-like powers. *Harv Int Rev* 40(4):18–21. <https://www.jstor.org/stable/26917261>. Accessed 3 May 2022
- Barendregt W, Becker C, Cheon E et al (2021) Defund Big Tech, refund community. *Tech Otherwise*, 5 Feb. <https://techotherwise.pubpub.org/pub/dakcci1r/release/1>. Accessed 15 May 2022
- Bellanova R, González Fuster G (2018) No (big) data, no fiction? Thinking surveillance with/against Netflix. In: Saetnan AR, Schneider I, Green N (eds) *The politics and policies of Big Data: Big Data Big Brother?* Routledge, London (forthcoming). <https://ssrn.com/abstract=3120038>. Accessed 3 May 2022
- Bourreau M, Caffarra C, Chen Z et al (2020) Google/Fitbit will monetise health data and harm consumers CEPR Policy Insight No. 107, 30 Sept. <https://voxeu.org/article/googlefitbit-will-monetise-health-data-and-harm-consumers>. Accessed 3 May 2022
- Brand V (2020) Corporate whistleblowing, smart regulation and regtech: the coming of the whistlebot? *Univ NSW Law J* 43(3):801–826. <https://doi.org/10.2139/ssrn.3698446>
- Bridle J (2018) Whistleblowers are a terrible answer to the problems of Big Tech. *Wired*, 11 June. <https://www.wired.co.uk/article/silicon-valley-whistleblowers-james-bridle-book-new-dark-age>. Accessed 3 May 2022
- Campbell IC (2021) Clearview AI hit with sweeping legal complaints over controversial face scraping in Europe. *The Verge*, 27 May. <https://www.theverge.com/2021/5/27/22455446/clearview-ai-legal-privacy-complaint-privacy-international-facial-recognition-eu>. Accessed 10 May 2022
- Cass RA (2013) Antitrust for high-tech and low: regulation, innovation, and risk. *J Law Econ Policy* 9(2): 169–200. <https://ssrn.com/abstract=2138254>. Accessed 3 May 2022
- Clayton J (2021) Silenced no more: a new era of tech whistleblowing? BBC News, 11 Oct. <https://www.bbc.com/news/technology-58850064>. Accessed 15 May 2022
- CNIL (2021) Facial recognition: the CNIL orders Clearview AI to stop reusing photographs available on the Internet. Commission Nationale de l'Informatique et des Libertés, Paris. <https://www.cnil.fr/en/facial-recognition-cnil-orders-clearview-ai-stop-reusing-photographs-available-internet>. Accessed 10 May 2022
- Coraggio G (2021) Facebook is NOT free and users shall be made aware of paying with their personal data. GamingTechLaw, 12 Apr. <https://www.gamingtechlaw.com/2021/04/facebook-not-free-personal-data-italian-court.html>. Accessed 3 May 2022
- Cosgrove L, Karter JM, Morrill Z, McGinley, M (2020) Psychology and surveillance capitalism: the risk of pushing mental health apps during the COVID-19 pandemic. *J Humanistic Psychol* 60(5):611–625. <https://doi.org/10.1177/0022167820937498>. Accessed 3 May 2022
- Council of the EU (2021) Regulating “Big Tech”: council agrees on enhancing competition in the digital sphere. Press release, 25 Nov. <https://www.consilium.europa.eu/en/press/press-releases/2021/11/25/regulating-big-tech-council-agrees-on-enhancing-competition-in-the-digital-sphere/>. Accessed 3 May 2022
- Crain M (2016) The limits of transparency: data brokers and commodification. *New Media Soc* 20(1):88–104. <https://doi.org/10.1177/1461444816657096>
- Doctorow C (2021) How to destroy surveillance capitalism. Medium Editions. <https://onezero.medium.com/how-to-destroy-surveillance-capitalism-8135e6744d59>. Accessed 3 May 2022
- Dubhashi D, Lappin S (2021) Scared about the threat of AI? It's the big tech giants that need reining in. *The Guardian*, 16 Dec. <https://www.theguardian.com/commentisfree/2021/dec/16/scared-about-the-threat-of-ai-its-the-big-tech-giants-that-need-reining-in>. Accessed 15 May 2022
- European Commission (2020a) Case M.9660—Google/Fitbit: public version. DG Competition, European Commission, Brussels. [https://ec.europa.eu/competition/mergers/cases/1/202120/m9660\\_3314\\_3.pdf](https://ec.europa.eu/competition/mergers/cases/1/202120/m9660_3314_3.pdf). Accessed 10 May 2022

- European Commission (2020b) Mergers: commission clears acquisition of Fitbit by Google, subject to conditions. Press release, 17 Dec. European Commission, Brussels. [https://ec.europa.eu/commission/presscorner/detail/en/ip\\_20\\_2484](https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2484). Accessed 10 May 2022
- European Commission (2020c) Proposal for a regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act). European Commission, Brussels. <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020c%3A842%3AFIN>. Accessed 11 May 2022
- European Commission (2020d) Proposal for a regulation of the European Parliament and of the Council on European data governance (Data Governance Act). European Commission, Brussels. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020dPC0767>. Accessed 11 May 2022
- European Commission (2021) Commission welcomes political agreement to boost data sharing and support European data spaces. Press release, 30 Nov. [https://ec.europa.eu/commission/presscorner/detail/en/IP\\_21\\_6428](https://ec.europa.eu/commission/presscorner/detail/en/IP_21_6428). Accessed 11 May 2022
- European Parliament, Council of the EU (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official J Eur Union L119(1):1–88. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN>. Accessed 1 May 2022
- Fadler M, Legner C (2021) Data ownership revisited: clarifying data accountabilities in times of Big Data and analytics. *J Bus Anal.* <https://doi.org/10.1080/2573234X.2021.1945961>
- Fernandez R (2021) How Big Tech is becoming the government. SOMO, 5 Feb. <https://www.somo.nl/how-big-tech-is-becoming-the-government/>. Accessed 15 May 2022
- Fernandez R, Adriaans I, Klinge TJ, Hendrikse R (2020) The financialisation of Big Tech. SOMO (Centre for Research on Multinational Corporations), Amsterdam. [https://www.somo.nl/nl/wp-content/uploads/sites/2/2020/12/Engineering\\_Financial-BigTech.pdf](https://www.somo.nl/nl/wp-content/uploads/sites/2/2020/12/Engineering_Financial-BigTech.pdf). Accessed 3 May 2022
- Fowler J (n.d.) Report: fitness tracker data breach exposed 61 million records and user data online. Website Planet. <https://www.websiteplanet.com/blog/gethealth-leak-report/>. Accessed 10 May 2022
- Garrett PM (2021) “Surveillance capitalism, COVID-19 and social work”: a note on uncertain future (s). *Br J Soc Work* 52(3):1747–1764. <https://doi.org/10.1093/bjsw/bcab099>
- Ghaffary S (2021) Big Tech’s employees are one of the biggest checks on its power. Vox-Recode, 29 Dec. <https://www.vox.com/recode/22848750/wistleblower-facebook-google-apple-employees>. Accessed 3 May
- Grunes AP, Stucke ME (2015) No mistake about it: the important role of antitrust in the era of Big Data. Antitrust Source, Online, University of Tennessee Legal Studies Research Paper 269. [https://papers.ssrn.com/sol3/Delivery.cfm SSRN\\_ID2608540\\_code869490.pdf?abstractid=2600051&mirid=1](https://papers.ssrn.com/sol3/Delivery.cfm SSRN_ID2608540_code869490.pdf?abstractid=2600051&mirid=1). Accessed 3 May 2022
- Harwell D (2022) Facial recognition firm Clearview AI tells investors it’s seeking massive expansion beyond law enforcement. The Washington Post, 16 Feb. <https://www.washingtonpost.com/technology/2022/02/16/clearview-expansion-facial-recognition/>. Accessed 10 May 2022
- Hummel P, Braun M, Dabrock P (2021) Own data? Ethical reflections on data ownership. *Philos Technol* 34:545–572. <https://doi.org/10.1007/s13347-020-00404-9>
- Jensen J (2019) Ethical aspects of surveillance capitalism. LinkedIn, 7 Nov. <https://www.linkedin.com/pulse/ethical-aspects-surveillance-capitalism-jostein-jensen/>. Accessed 15 May 2022
- Jolly J (2021) Is Big Tech now just too big to stomach? The Guardian, 6 Feb. <https://www.theguardian.com/business/2021/feb/06/is-big-tech-now-just-too-big-to-stomach>. Accessed 9 May 2022
- Jurcys P, Donewald C, Fenwick M et al (2020) Ownership of user-held data: why property law is the right approach. Harv J Law Technol Digest. <https://jolt.law.harvard.edu/assets/digestImages/ Paulius-Jurcys-Feb-19-article-PJ.pdf>. Accessed 3 May 2022

- Kang SS, (2020) Don't blame privacy for Big Tech's monopoly on information. Just Security, 18 Sept. <https://www.justsecurity.org/72439/dont-blame-privacy-for-big-techs-monopoly-on-information/>. Accessed 3 May 2022
- Kavenna J (2019) Interview. Shoshana Zuboff: 'Surveillance capitalism is an assault on human autonomy'. The Guardian, 4 Oct. <https://www.theguardian.com/books/2019/oct/04/shoshana-zuboff-surveillance-capitalism-assault-human-autonomy-digital-privacy>. Accessed 16 May 2022
- Leblond P (2020) How open data could tame Big Tech's power and avoid a breakup. The Conversation, 5 Aug. <https://theconversation.com/how-open-data-could-tame-big-techs-power-and-avoid-a-breakup-143962>. Accessed 3 May 2022
- Lehtiöniemi T (2017) Personal data spaces: an intervention in surveillance capitalism? *Surveill Soc* 15(5):626–639. <https://doi.org/10.24908/ss.v15i5.6424>
- Lin Y (2018) #DeleteFacebook is still feeding the beast—but there are ways to overcome surveillance capitalism. The Conversation, 26 Mar. <https://theconversation.com/deletefacebook-is-still-feeding-the-beast-but-there-are-ways-to-overcome-surveillance-capitalism-93874>. Accessed 3 May 2022
- Lyon D (2019) Surveillance capitalism, surveillance culture and data politics. In: Bigo D, Isin E, Ruppert E (eds) *Data politics: worlds, subjects, rights*. Routledge, Oxford, pp 64–77
- Matsakis L (2019) Break up Big Tech? Some say not so fast. Wired, 7 June. <https://www.wired.com/story/break-up-big-tech-antitrust-laws/>. Accessed 3 May 2022
- Milmo D (2021) Frances Haugen: "I never wanted to be a whistleblower. But lives were in danger". The Observer, 24 Oct. <https://www.theguardian.com/technology/2021/oct/24/frances-haugen-i-never-wanted-to-be-a-whistleblower-but-lives-were-in-danger>. Accessed 15 May 2022
- Moss DL (2019) Breaking up is hard to do: the implications of restructuring and regulating digital technology markets. *The Antitrust Source* 19(2). [https://www.americanbar.org/content/dam/aba/publishing/antitrust-magazine-online/2018-2019/atsource-october2019/oct19\\_full\\_source.pdf](https://www.americanbar.org/content/dam/aba/publishing/antitrust-magazine-online/2018-2019/atsource-october2019/oct19_full_source.pdf). Accessed 3 May 2022
- Nash, J (2021) Amazon sees profit in your palmprint. Opponents see harmful surveillance capitalism. *Biometric Update*, 3 Aug. <https://www.biometricupdate.com/202108/amazon-sees-profit-in-your-palmprint-opponents-see-harmful-surveillance-capitalism>. Accessed 10 May 2022
- Nass SJ, Levit LA, Gostin LO (eds) (2009) *Beyond the HIPAA privacy rule: enhancing privacy, improving health through research*. National Academies Press, Washington DC. <https://www.ncbi.nlm.nih.gov/books/NBK9579/>. Accessed 10 May 2022
- Oremus W (2018) Are you really the product? The history of a dangerous idea. *Slate*, 27 Apr. <https://slate.com/technology/2018/04/are-you-really-facesbooks-product-the-history-of-a-dangerous-idea.html>. Accessed 10 May 2022
- Perrigo B (2019) "The capabilities are still there." Why Cambridge Analytica whistleblower Christopher Wylie is still worried. *Time*, 8 Oct. <https://time.com/5695252/christopher-wylie-cambridge-analytica-book/>. Accessed 15 May 2022
- Pistor K (2020) Rule by data: the end of markets? *Law Contemp Prob* 83:101–124. <https://scholarship.law.duke.edu/lcp/vol83/iss2/6>. Accessed 4 May 2022
- Reuters (2021) Factbox: how Big Tech is faring against U.S. lawsuits and probes. *Reuters*, 7 Dec. <https://www.reuters.com/technology/big-tech-wins-two-battles-fight-with-us-antitrust-enforcers-2021-06-29/>. Accessed 4 May 2022
- Rodrigues R, Panagiotopoulos A, Lundgren B et al (2020) SHERPA deliverable 3.3 Report on regulatory options. <https://doi.org/10.21253/DMU.11618211.v7>
- Scoroxton A (2021) Mass health tracker data breach has UK impact. *Computer Weekly*, 14 Sept. <https://www.computerweekly.com/news/252506664/Mass-health-tracker-data-breach-has-UK-impact>. Accessed 10 May 2022
- Simonite T (2021) What really happened when Google ousted Timnit Gebru. *Wired*, 8 June. <https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/>. Accessed 15 May 2022
- Sokol DD, Comerford R (2016) Antitrust and regulating Big Data. *Geo Mason Law Rev* 23:1129–1161. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2834611](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2834611). Accessed 4 May 2022

- Swartz J (2021) Big Tech heads for “a year of thousands of tiny tech papercuts,” but what antitrust efforts could make them bleed? MarketWatch, 27 Dec (updated 1 Jan 2022). <https://www.marketwatch.com/story/big-tech-heads-for-a-year-of-thousands-of-tiny-tech-papercuts-but-what-antitrust-efforts-could-make-them-bleed-11640640776>. Accessed 4 May 2022
- The Economist (2019) Breaking up is hard to do: dismembering Big Tech. The Economist, 24 Oct. <https://www.economist.com/business/2019/10/24/dismembering-big-tech>. Accessed 4 May 2022
- UNESCO (2021) Recommendation on the ethics of artificial intelligence. SHS/BIO/PI/2021/. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>. Accessed 18 Oct 2022
- Van Asbroeck B, Debussche J, César J (2019) Big Data & issues & opportunities: data ownership. Bird & Bird, 25 Mar. <https://www.twobirds.com/en/news/articles/2019/global/big-data-and-issues-and-opportunities-data-ownership>. Accessed 3 May 2022
- Warren E (2019) Here’s how we can break up Big Tech. Medium, 8 Mar. <https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c>. Accessed 4 May 2022
- Waters R (2019) Three ways that Big Tech could be broken up. Financial Times, 7 June. <https://www.ft.com/content/cb8b707c-88ca-11e9-a028-86cea8523dc2>. Accessed 4 May 2022
- Wolff J, Atallah N (2021) Early GDPR penalties: analysis of implementation and fines through May 2020. J Inf Policy 11(1):63–103. <https://doi.org/10.5325/jinfopoli.11.2021.0063>
- Zuboff S (2015) Big other: surveillance capitalism and the prospects of an information civilization. J Inf Technol 30(1):75–89. <https://doi.org/10.1057/jit.2015.5>
- Zuboff S (2019a) Facebook, Google and a dark age of surveillance capitalism. Financial Times, 25 Jan. <https://www.ft.com/content/7fafec06-1ea2-11e9-b126-46fc3ad87c65>. Accessed 10 May 2022
- Zuboff S (2019b) The age of surveillance capitalism: the fight for a human future at the new frontier of power. Public Affairs, New York
- Zuboff S (2020) Surveillance capitalism. Project Syndicate, 3 Jan. <https://www.project-syndicate.org/onpoint/surveillance-capitalism-exploiting-behavioral-data-by-shoshana-zuboff-2020-01>. Accessed 10 May 2022
- Zuboff S (2021) The coup we are not talking about. The New York Times, 29 Jan. <https://www.nytimes.com/2021/01/29/opinion/sunday/facebook-surveillance-society-technology.html>. Accessed 4 May 2022

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# Chapter 5

## Manipulation



**Abstract** The concern that artificial intelligence (AI) can be used to manipulate individuals, with undesirable consequences for the manipulated individual as well as society as a whole, plays a key role in the debate on the ethics of AI. This chapter uses the case of the political manipulation of voters and that of the manipulation of vulnerable consumers as studies to explore how AI can contribute to and facilitate manipulation and how such manipulation can be evaluated from an ethical perspective. The chapter presents some proposed ways of dealing with the ethics of manipulation with reference to data protection, privacy and transparency in the use of data. Manipulation is thus an ethical issue of AI that is closely related to other issues discussed in this book.

**Keywords** Right to life · Safety · Security · Self-driving cars · Smart homes · Adversarial attacks · Responsibility · Liability · Quality management · Adversarial robustness

### 5.1 Introduction

In the wake of the 2016 US presidential election and the 2016 Brexit referendum it became clear that AI had been used to target undecided voters and persuade them to vote in a particular direction. Both polls were close, and a change of mind by a single-digit percentage of the voter population would have been enough to change the outcome. It is therefore reasonable to state that these interventions led by artificial intelligence (AI) played a causal role in the ascent of Donald Trump to the American presidency and the success of the Brexit campaign.

These examples of the potential manipulation of elections are probably the most high-profile cases of human action being influenced using AI. They are not the only ones, however, and they point to the possibility of much further-reaching manipulation activities that may be happening already, but are currently undetected.

## 5.2 Cases of AI-Enabled Manipulation

### 5.2.1 Case 1: Election Manipulation

The 2008 US presidential election has been described as the first that “relied on large-scale analysis of social media data, which was used to improve fundraising efforts and to coordinate volunteers” (Polonski 2017). The increasing availability of large data sets and AI-enabled algorithms led to the recognition of new possibilities of technology use in elections. In the early 2010s, Cambridge Analytica, a voter-profiling company, wanted to become active in the 2014 US midterm election (Rosenberg et al. 2018). The company attracted a \$15 million investment from Robert Mercer, a Republican donor, and engaged Stephen Bannon, who later played a key role in President Trump’s 2016 campaign and was an important early member of the Trump cabinet. Cambridge Analytica lacked the data required for voter profiling, so it solved this problem with Facebook data (Cadwalladr and Graham-Harrison 2018). Using a permission to harvest data for academic research purposes that Facebook had granted to Aleksandr Kogan, a researcher with links to Cambridge University, the company harvested not just the data of people who had been paid to take a personality quiz, but also that of their friends. This allowed Cambridge Analytica to harvest in total 50 million Facebook profiles, which allowed the delivery of personalised messages to the profile holders and also – importantly – a wider analysis of voter behaviour.

The Cambridge Analytica case led to a broader discussion of the permissible and appropriate uses of technology in Western democracies. Analysing large datasets with a view to classifying demographics into small subsets and tailoring individual messages designed to curry favour with the individuals requires data analytics techniques that are part of the family of technologies typically called AI.

We will return to the question of the ethical evaluation of manipulation below. The questions that are raised by manipulation will become clearer when we look at a second example, this one in the commercial sphere.

### 5.2.2 Case 2: Pushing Sales During “Prime Vulnerability Moments”

Human beings do not feel and behave the same way all of the time; they have ups and downs, times when they feel more resilient and times when they feel less so. A 2013 marketing study suggests that one can identify typical times when people feel more vulnerable than usual. US women across different demographic categories, for example, have been found to feel least attractive on Mondays, and therefore possibly more open to buying beauty products (PHD Media 2013). This study goes on to suggest that such insights can be used to develop bespoke marketing strategies. While the original study couches this approach in positive terms such as “encourage”

and “empower”, independent observers have suggested that it may be the “grossest advertising strategy of all time” (Rosen 2013).

Large internet companies such as Google and Amazon use data they collect about potential customers to promote goods and services that their algorithms suggest searchers are in need of or looking for. This approach could easily be combined with the concept of “prime vulnerability moments”, where real-time data analysis is used to identify such moments in much more detail than the initial study.

The potential manipulation described in this second case study is already so widespread that it may not be noticeable any more. Most internet users are used to being targeted in advertising.

The angle of the case that is interesting here is the use of the “prime vulnerability moment”, which is not yet a concept widely referred to in AI-driven personal marketing. The absence of a word for this concept does not mean, however, that the underlying approach is not used. As indicated, the company undertaking the original study couched the approach in positive and supportive terms. The outcome of such a marketing strategy may in fact be positive for the target audience. If a person has a vulnerable moment due to fatigue, suggestions of relevant health and wellbeing products might help combat that state. This leads us to the question we will now discuss: whether and in what circumstances manipulation arises, and how it can be evaluated from an ethical position.

### 5.3 The Ethics of Manipulation

An ethical analysis of the concept of manipulation should start with an acknowledgement that the term carries moral connotations. The Cambridge online dictionary offers the following definition: “controlling someone or something to your own advantage, often unfairly or dishonestly” (Cambridge Dictionary n.d.) and adds that it is used mainly in a disapproving way. The definition thus offers several pointers to why manipulation is seen as ethically problematic. The act of controlling others may be regarded as concerning, especially the fact that it is done for someone’s advantage, which is exacerbated if it is done unfairly or dishonestly. In traditional philosophical terms, it is Kant’s prominent categorical imperative that prohibits such manipulation on ethical grounds, because one person is being used solely as a means to another person’s ends (Kant 1998: 37 [4:428]).

One aspect of the discussion that is pertinent to the first case study is that the manipulation of the electorate through AI can damage democracy.

AI can have (and likely already has) an adverse impact on democracy, in particular where it comes to: (i) social and political discourse, access to information and voter influence, (ii) inequality and segregation and (iii) systemic failure or disruption. (Muller 2020: 12)

Manipulation of voters using AI techniques can fall under heading (i) as voter influence. However, it is not clear under which circumstances such influence on voters would be illegitimate. After all, election campaigns explicitly aim to influence voters and doing so is the daily work of politicians. The issue seems to be not so much the fact that voters are influenced, but that this happens without their knowledge and maybe in ways that sidestep their ability to critically reflect on election messages. An added concern is the fact that AI is mostly held and made use of by large companies, and that these are already perceived to have an outsized influence on policy decisions, which can be further extended through their ability to influence voters. This contributes to the “concentration of technological, economic and political power among a few mega corporations [that] could allow them undue influence over governments” (European Parliament 2020: 16).

Another answer to the question why AI-enabled manipulation is ethically problematic is that it is based on privacy infringements and constitutes surveillance. This is certainly a key aspect of the Cambridge Analytica case, where the data of Facebook users was harvested in many cases without their consent or awareness. This interpretation would render the manipulation problem a subproblem of the broader discussion of privacy, data protection and surveillance as discussed in Chap. 3.

However, the issue of manipulation, while potentially linked with privacy and other concerns, seems to point to a different fundamental ethical concern. In being manipulated, the objects of manipulation, whether citizens and voters or consumers, seem to be deprived of their basic freedom to make informed decisions.

Freedom is a well-established ethical value that finds its expressions in many aspects of liberal democracy and forms a basis of human rights. It is also a very complex concept that has been discussed intensively by moral philosophers and others over millennia (Mill 1859; Berlin 2002). While it may sound intuitively plausible to say that manipulating individuals using AI-based tools reduces their freedom to act as they normally would, it is more difficult to determine whether or how this is the case. There are numerous interventions which claim that AI can influence human behaviour (Whittle 2021), for example by understanding cognitive biases and using them to further one’s own ends (Maynard 2019). In particular the collecting of data from social media seems to provide a plausible basis for this claim, where manipulation (Mind Matters 2018) is used to increase corporate profits (Yearsley 2017). However, any such interventions look different from other threats to our freedom to act or to decide, such as incarceration and brainwashing.

Facebook users in the Cambridge Analytica case were not forced to vote in a particular way but received input that influenced their voting behaviour. Of course, this is the intended outcome of election campaigns. Clearly the argument cannot be that one should never attempt to influence other people’s behaviour. This is what the law and, to some extent, ethics do as a matter of course. Governments, companies and also special interest groups all try to influence, often for good moral reasons. If a government institutes a campaign to limit smoking by displaying gruesome pictures of cancer patients on cigarette packets, then this has the explicit intention of dissuading people from smoking without ostensibly interfering with their basic right to freedom. We mentioned the idea of nudging in Chap. 3, in the context of

privacy (Benartzi et al. 2017), which constitutes a similar type of intervention. While nudging is contentious, certainly when done by governments, it is not always and fundamentally unethical.

So perhaps the reference to freedom or liberty as the cause of ethical concerns in the case of manipulation is not fruitful in the discussion of the Cambridge Analytica case. A related alternative that is well established as a mid-level principle from biomedical ethics (Childress and Beauchamp 1979) is that of autonomy. Given that biomedical principles including autonomy have been widely adopted in the AI ethics debate, this may be a more promising starting point. Respect for autonomy is, for example, one of the four ethical principles that the EU's High-Level Expert Group bases its ethics guidelines for trustworthy AI on (AI HLEG 2019). The definition of this principle makes explicit reference to the ability to partake in the democratic process and states that “AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans” This suggests that manipulation is detrimental to autonomy as it reduces “meaningful opportunity for human choice” (ibid: 12).

This position supports the contention that the problem with manipulation is its detrimental influence on autonomy. A list of requirements for trustworthy AI starts with “human agency and oversight” (ibid: 15). This requirement includes the statement that human autonomy may be threatened when AI systems are “deployed to shape and influence human behaviour through mechanisms that may be difficult to detect, since they may harness sub-conscious processes” (ibid: 17). The core of the problem, then, is that people are not aware of the influence that they are subjected to, rather than the fact that their decisions or actions are influenced in a particular way.

This allows an interesting question to be raised about the first case study (Facebook and Cambridge Analytica). Those targeted were not aware that their data had been harvested from Facebook, but they may have been aware that they were being subjected to attempts to sway their political opinion – or conceivably might have been, if they had read the terms and conditions of Facebook and third-party apps they were using. In this interpretation the problem of manipulation has a close connection to the question of informed consent, a problem that has been highlighted with regard to possible manipulation of Facebook users prior to the Cambridge Analytica event (Flick 2016).

The second case (pushing sales during “prime vulnerability moments”) therefore presents an even stronger example of manipulation, because the individuals subjected to AI-enabled interventions may not have been aware of this at all. A key challenge, then, is that technology may be used to fundamentally alter the space of perceived available options, thereby clearly violating autonomy.

Coeckelbergh (2019) uses the metaphor of the theatre, with a director who sets the stage and thereby determines what options are possible in a play. AI can similarly be used to reveal or hide possible options for people in the real world. In this case manipulation would be undetectable by the people who are manipulated, precisely because they do not know that they have further options. It is not always possible to fully answer the question: when does an acceptable attempt to influence someone

turn into an unacceptable case of manipulation? But it does point to possible ways of addressing the problem.

## 5.4 Responses to Manipulation

An ethical evaluation of manipulation is of crucial importance in determining which interventions may be suitable to ensure that AI use is acceptable. If the core of the problem is that political processes are disrupted and power dynamics are affected in an unacceptable manner, then the response could be sought at the political level. This may call for changes to electoral systems or maybe the breaking up of inappropriately powerful large tech companies that threaten existing power balances, as proposed by the US senator and former presidential candidate Warren (2019) and others (Yglesias 2019). Similarly, if the core of the ethical concern is the breach of data protection and privacy, then strengthening or enforcing data protection rules is likely to be the way forward.

While such interventions may be called for, the uniqueness of the ethical issue of manipulation seems to reside in the hidden way in which people are influenced. There are various ways in which this could be addressed. On one hand, one could outlaw certain uses of personal data, for example its use for political persuasion. As political persuasion is neither immoral in principle nor illegal, such an attempt to regulate the use of personal data would likely meet justified resistance and be difficult to define and enforce legally.

A more promising approach would be to increase the transparency of data use. If citizens and consumers understood better how AI technologies are used to shape their views, decisions and actions, they would be in a better position to consciously agree or disagree with these interventions, thereby removing the ethical challenge of manipulation.

Creating such transparency would require work at several levels. At all of these levels, there is the need to understand and explain how AI systems work. Machine learning is currently the most prominent AI application that has given rise to much of the ethical discussion of AI. One of the characteristics of machine learning approaches using neural networks and deep learning (Bengio et al. 2021) is the opacity of the resulting model. A research stream on explainable AI has developed in response to this problem of technical opacity. While it remains a matter of debate whether explainability will benefit AI, or to what degree the internal states of an AI system can be subject to explanation (Gunning et al. 2019), much technical work has been undertaken to provide ways in which humans can make sense of AI and AI outputs. For instance, there have been contributions to the debate highlighting the need for humans to be able to relate to it (Miller 2019; Mittelstadt et al. 2019). Such work could, for example, make it clear to individual voters why they have been selected as targets for a specific political message, or to consumers why they are deemed to be suitable potential customers for a particular product or service.

Technical explainability will not suffice to address the problem. The ubiquity of AI applications means that individuals, even if highly technology-savvy, will not have the time and resources to follow up all AI decisions that affect them and even less to intervene, should these be wrong or inappropriate. There will thus need to be a social and political side to transparency and explainability. This can include the inclusion of stakeholders in the design, development and implementation of AI, which is an intention that one can see in various political AI strategies (Presidency of the Council of the EU 2020; HM Government 2021).

Stakeholder involvement is likely to address some of the problems of opacity, but it is not without problems, as it poses the perennial question: who should have a seat at the table (Borenstein et al. 2021)? It will therefore need to be supplemented with processes that allow for the promotion of meaningful transparency. This requires the creation of conditions where adversarial transparency is possible, for instance where critical civil society groups such as Privacy International<sup>1</sup> are given access to AI systems in order to scrutinise those systems as well as their uses and social consequences. To be successful, this type of social transparency will need a suitable regulatory environment. This may include direct legislation that would force organisations to share data about their systems; a specific regulator with the power to grant access to systems or undertake independent scrutiny; and/or novel standards or processes, such as AI impact assessments, whose findings are required to be published (see Sect. 2.4.1).

## 5.5 Key Insights

This chapter has shown that concerns about manipulation as an ethical problem arising from AI are closely related to other ethical concerns. Manipulation is directly connected to data protection and privacy. It has links to broader societal structures and the justice of our socio-economic systems and thus relates to the problem of surveillance capitalism. By manipulating humans, AI can reduce their autonomy.

The ethical issue of manipulation can therefore best be seen using the systems-theoretical lens proposed by Stahl (2021, 2022). Manipulation is not a unique feature that arises from particular uses of a specific AI technology; it is a pervasive capability of the AI ecosystem(s). Consequently what is called for is not one particular solution, but rather the array of approaches discussed in this book. In the present chapter we have focused on transparency and explainable AI as key aspects of a successful mitigation strategy. However, these need to be embedded in a larger regulatory framework and are likely to draw on other mitigation proposals ranging from standardisation to ethics-by-design methodologies.

---

<sup>1</sup> <https://privacyinternational.org/>.

## References

- AI HLEG (2019) Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence. European Commission, Brussels. [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=60419](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419). Accessed 25 Sept 2020
- Benartzi S, Besears J, Mlikman K et al (2017) Governments are trying to nudge us into better behavior. Is it working? The Washington Post, 11 Aug. <https://www.washingtonpost.com/news/wonk/wp/2017/08/11/governments-are-trying-to-nudge-us-into-better-behavior-is-it-working/>. Accessed 1 May 2022
- Bengio Y, Lecun Y, Hinton G (2021) Deep learning for AI. Commun ACM 64:58–65. <https://doi.org/10.1145/3448250>
- Berlin I (2002) Liberty. Oxford University Press, Oxford
- Borenstein J, Grodzinsky FS, Howard A et al (2021) AI ethics: a long history and a recent burst of attention. Computer 54:96–102. <https://doi.org/10.1109/MC.2020.3034950>
- Cadwalladr C, Graham-Harrison E (2018) How Cambridge analytica turned Facebook ‘likes’ into a lucrative political tool. The Guardian, 17 Mar. <https://www.theguardian.com/technology/2018/mar/17/facebook-cambridge-analytica-kogan-data-algorithm>. Accessed 1 May 2022
- Cambridge Dictionary (n.d.) Manipulation. <https://dictionary.cambridge.org/dictionary/english/manipulation>. Accessed 11 May 2022
- Childress JF, Beauchamp TL (1979) Principles of biomedical ethics. Oxford University Press, New York
- Coeckelbergh M (2019) Technology, narrative and performance in the social theatre. In: Kreps D (ed) Understanding digital events: Bergson, Whitehead, and the experience of the digital, 1st edn. Routledge, New York, pp 13–27
- European Parliament (2020) The ethics of artificial intelligence: issues and initiatives. European Parliamentary Research Service, Brussels. [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS\\_STU\(2020\)634452\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf). Accessed 1 May 2022
- Flick C (2016) Informed consent and the Facebook emotional manipulation study. Res Ethics 12. <https://doi.org/10.1177/1747016115599568>
- Gunning D, Stefik M, Choi J et al (2019) XAI: explainable artificial intelligence. Sci Robot 4(37). <https://doi.org/10.1126/scirobotics.aay7120>
- HM Government (2021) National AI strategy. Office for Artificial Intelligence, London. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/1020402/National\\_AI\\_Strategy\\_-\\_PDF\\_version.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf)
- Kant I (1998) Groundwork of the metaphysics of morals. Cambridge University Press, Cambridge
- Maynard A (2019) AI and the art of manipulation. Medium, 18 Nov. <https://medium.com/edge-of-innovation/ai-and-the-art-of-manipulation-3834026017d5>. Accessed 15 May 2022
- Mill JS (1859) On liberty and other essays. Kindle edition, 2010. Digireads.com
- Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Mind Matters (2018) AI social media could totally manipulate you, 26 Nov. <https://mindmatters.ai/2018/11/ai-social-media-could-totally-manipulate-you/>. Accessed 15 May 2022
- Mittelstadt B, Russell C, Wachter S (2019) Explaining explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency (FAT\*’19). Association for Computing Machinery, New York, pp 279–288. <https://doi.org/10.1145/3287560.3287574>
- Muller C (2020) The impact of artificial intelligence on human rights, democracy and the rule of law. Ad Hoc Committee on Artificial Intelligence (CAHAI), Council of Europe, Strasbourg. <https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da>. Accessed 2 May 2022
- PHD Media (2013) New beauty study reveals days, times and occasions when U.S. women feel least attractive, 2 Oct. <https://www.prnewswire.com/news-releases/new-beauty-study-reveals-days-times-and-occasions-when-us-women-feel-least-attractive-226131921.html>. Accessed 11 May 2022

- Polonski V (2017) The good, the bad and the ugly uses of machine learning in election campaigns, 30 Aug. Centre for Public Impact, London. <https://www.centreforpublicimpact.org/insights/good-bad-ugly-uses-machine-learning-election-campaigns>. Accessed 11 May 2022
- Presidency of the Council of the EU (2020) Presidency conclusions: the Charter of Fundamental Rights in the context of artificial intelligence and digital change. Council of the European Union, Brussels. <https://www.consilium.europa.eu/media/46496/st11481-en20.pdf>. Accessed 1 May 2022
- Rosen RJ (2013) Is this the grossest advertising strategy of all time? The Atlantic, 3 Oct. <https://www.theatlantic.com/technology/archive/2013/10/is-this-the-grossest-advertising-strategy-of-all-time/280242/>. Accessed 11 May 2022
- Rosenberg M, Confessore N, Cadwalladr C (2018) How Trump consultants exploited the Facebook data of millions. The New York Times, 17 Mar. <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html>. Accessed 11 May 2022
- Stahl BC (2021) From computer ethics and the ethics of AI towards an ethics of digital ecosystems. *AI Ethics*. <https://doi.org/10.1007/s43681-021-00080-1>
- Stahl BC (2022) Responsible innovation ecosystems: ethical implications of the application of the ecosystem concept to artificial intelligence. *Int J Inf Manage* 62:102441. <https://doi.org/10.1016/j.ijinfomgt.2021.102441>
- Warren E (2019) Here's how we can break up Big Tech. Medium, 8 Mar. <https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c>. Accessed 15 May 2022
- Whittle J (2021) AI can now learn to manipulate human behaviour. The Conversation, 11 Feb. <https://theconversation.com/ai-can-now-learn-to-manipulate-human-behaviour-155031>. Accessed 15 May 2022
- Yearsley Y (2017) We need to talk about the power of AI to manipulate humans. MIT Technology Review, 5 June. <https://www.technologyreview.com/2017/06/05/105817/we-need-to-talk-about-the-power-of-ai-to-manipulate-humans/>. Accessed 15 May 2022
- Yglesias M (2019) The push to break up Big Tech, explained. Vox-Recode, 3 May. <https://www.vox.com/recode/2019/5/3/18520703/big-tech-break-up-explained>. Accessed 15 May 2022

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# Chapter 6

## Right to Life, Liberty and Security of Persons



**Abstract** Artificial intelligence (AI) can support individuals' enjoyment of life, liberty and security, but it can also have adverse effects on them in a variety of ways. This chapter covers three cases affecting human life, liberty and security: one in transportation (self-driving cars), one in the home (smart security systems) and one in healthcare services (adversarial attacks). The chapter discusses ethical questions and three potential solutions to address AI human rights issues related to life, liberty and security of persons: defining and strengthening liability regimes, implementing quality management systems and adversarial robustness. AI developers, deployers and users must respect the sanctity of human life and embed, value and respect this principle in the design, development and use of their products and/or services. Critically, AI systems should *not* be programmed to kill or injure humans.

**Keywords** Right to life · Safety · Security · Self-driving cars · Smart homes · Adversarial attacks

### 6.1 Introduction

All humans enjoy the right to life, liberty and security of the person. The right to life is also included as a core right in 77% of the world's constitutions (UN 2018), is the cornerstone of other rights and is enshrined in international human rights instruments (Table 6.1).

State parties who are signatories to the human rights instruments enshrining the right to life have a duty to take necessary measures to ensure individuals are protected from its violation: i.e. its loss, deprivation or removal.

Artificial intelligence (AI) can support an individual's enjoyment of life, liberty and security by, for example, supporting the diagnosis and treatment of medical conditions. Raso et al. (2018) outline how criminal justice risk assessment tools could benefit low-risk individuals through increased pre-trial releases and shorter sentences. Reports suggest that AI tools could help identify and mitigate human security risks and lower crime rates (Deloitte n.d., Muggah 2017).

**Table 6.1** Right to life in international human rights instruments

Provision	Human Rights Instrument
Right to life, liberty and security of person	Universal Declaration of Human Rights (UN 1948: art. 3)
Right to life	International Covenant on Civil and Political Rights (UN 1966: art. 6)
Right to life, survival and development	Convention on the Rights of the Child (UN 1989: art. 6)
Right to life	International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (UN 1990: art. 9)
Right to life	Convention on the Rights of Persons with Disabilities (UN 2006: art. 10)
Right to life	American Convention on Human Rights (Pact of San José) (UN 1969: art. 4)
Right to life	African Charter on Human and Peoples' Rights (Banjul Charter) (ACHPR 1981: art. 4)
Right to life and dignity in old age	Inter-American Convention on Protecting the Human Rights of Older Persons (OAS 2015: art. 6)
Right to life	European Convention on Human Rights (ECHR 1950: art. 2)

AI can have adverse effects on human life, liberty and security in a variety of ways (Vasic and Billard 2013; Leslie 2019), as elaborated in this chapter. Human rights issues around life, liberty and security of persons are particularly serious, and risks from the use of AI need to be weighed up against the risks incurred when not using AI, in comparison with other innovations. AI systems identified as high-risk (European Commission 2021) include those used in critical infrastructure (e.g. transportation) that could put the life and health of people at risk; in educational or vocational training that determine access to education and the professional course of someone's life (e.g. the scoring of exams); in the safety components of products (e.g. AI applications in robot-assisted surgery); in employment, the management of workers and access to self-employment (e.g. CV-sorting software for recruitment procedures); in essential private and public services (e.g. when credit scoring denies citizens the opportunity to obtain a loan); in law enforcement that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence); in migration, asylum and border control management (e.g. verification of the authenticity of travel documents); and in the administration of justice and democratic processes (e.g. applying the law to a concrete set of facts). These categories of high-risk AI systems have the potential to impact the right to life, liberty and security (some in more direct ways than others, but nonetheless relevant).

Life-threatening issues have been raised regarding the use of robot-assisted medical procedures and robotics systems in surgery (Alemzadeh et al. 2016), robot

accidents and malfunctions in manufacturing, law enforcement (Boyd 2016), retail and entertainment settings (Jiang and Gainer 1987), security vulnerabilities in smart home hubs (Fránik and Čermák 2020), self-driving and autonomous vehicles (AP and Reuters 2021), and lethal attacks by AI-armed drone swarms and autonomous weapons (Safi 2019). We look at three different cases affecting human life, liberty and security, one in the transportation context (self-driving cars), one related to the home (smart home security), and one in the healthcare service setting (adversarial attacks).

## 6.2 Cases of AI Adversely Affecting the Right to Life, Liberty and Security of Persons

### 6.2.1 Case 1: Fatal Crash Involving a Self-driving Car

In May 2016, a Tesla car was the first known self-driving car to be involved in a fatal crash. The 42-year-old passenger/driver died instantly after colliding with a tractor-trailer. The tractor driver was not injured. “According to Tesla’s account of the crash, the car’s sensor system, against a bright spring sky, failed to distinguish a large white 18-wheel truck and trailer crossing the highway” (Levin and Woolf 2016). An examination by the Florida Highway Patrol concluded that the Tesla driver had not been attentive and had failed to take evasive action. At the same time, the tractor driver had failed, during a left turn, to give right of way, according to the report. (Golson 2017)

In this case, the driver had put his car into Tesla’s autopilot mode, which was able to control the car. According to Tesla, its autopilot is “an advanced driver assistance system that enhances safety and convenience behind the wheel” and, “[w]hen used properly” is meant to reduce a driver’s “overall workload” (Tesla n.d.). While Tesla clarified that the underlying autonomous software was designed to nudge consumers to keep their hands on the wheels to make sure they were paying attention, that does not seem to have happened in this case and resulted in a fatality. According to Tesla, “the currently enabled Autopilot and Full Self-Driving features require active driver supervision and do not make the vehicle autonomous” (*ibid*).

In 2018, an Uber test driver in charge of monitoring one of the company’s self-driving cars was charged with negligent homicide when it hit and killed a pedestrian. An investigation by the National Transportation Safety Board (NSTB) concluded that the crash had been caused by the Uber test driver being distracted by her phone and implicated Uber’s inadequate safety culture (McFarland 2019). The NSTB also found that Uber’s system could not correctly classify and predict the path of a pedestrian crossing midblock.

In 2021, two men were killed in Texas after the Tesla vehicle they were in, which was going at a high speed, went off the road and hit a tree. The news report also

mentioned that the men been discussing the autopilot feature before they drove off ([Pietsch 2021](#)). Evidence is believed to show that no one was driving the vehicle when it crashed.

While drivers seem to expect self-driving cars, as marketed to them, to give them more independence and freedom, self-driving cars are not yet, as stated by Tesla, for example, “autonomous”. The autopilot function and the “Full Self-Driving” capability are intended for use with a fully attentive driver with hands on the wheel and ready to take over at any moment.

While some research ([Kalra and Groves 2017](#); [Teoh and Kidd 2017](#)) seems to suggest that self-driving cars may be safer than those driven by the average human driver, the main case and the further examples cited here point to human safety challenges from different angles: the safety of the drivers, passengers and other road users (e.g. cyclists, pedestrians and animals) and objects that encounter self-driving cars.

Other standard issues raised about self-driving cars, as outlined by [Jansen et al. \(2020\)](#), relate to *security* (the potential for their hacking leading to the compromising of personal and sensitive data) and *responsibility*, that is, where does responsibility for harms caused lie: with the manufacturer, the system programmer or software engineer, the driver/passenger, or the insurers? A responsibility gap could also occur, as pointed out by the Council of Europe’s Committee on Legal Affairs and Human Rights, “where the human in the vehicle—the ‘user-in-charge’, even if not actually engaged in driving—cannot be held liable for criminal acts and the vehicle itself was operating according to the manufacturer’s design and applicable regulations.” ([Council of Europe 2020](#)). There is also the challenge of shared driving responsibilities between the human driver and the system ([BBC News 2020](#)).

The underlying causes that require addressing in these cases include software/system vulnerabilities, inadequate safety risk assessment procedures and oversight of vehicle operators, as well as human error and driver distractions (including a false sense of security) ([Clifford Law 2021](#)).

### **6.2.2 Case 2: Smart Home Hubs Security Vulnerabilities**

A smart home hub is a control centre for home automation systems, such as those operating the heating, blinds, lights and internet-enabled electronic appliances. Such systems allow the user to interact remotely with the hub using, for instance, a smartphone. A user who is equipped to activate appliances remotely can arrive at home with the networked gas fire burning and supper ready in the networked oven. However, it is not only the users themselves who can access their smart home hubs, but also external entities, if there are security vulnerabilities, as was the case for three companies operating across Europe. ([Fránič and Čermák 2020](#))

Smart home security vulnerabilities directly affect all aspects of the right to life, liberty and security of the person. E.g., Man-in-the-middle attacks that interrupt or spoof communication between smart home devices and denial-of-service attacks could disrupt or shut devices down and compromise user well-being, safety and security.

Such vulnerabilities and attacks exploiting them can threaten a home, together with the peaceful enjoyment of life and human health within it. Unauthorised access could also result in threats to human life and health. For example, as outlined in a report from the European Union Agency for Cybersecurity (ENISA), safety might be compromised and human life thus endangered by the breach, or loss of control, of a thermostat, a smoke detector, a CO<sub>2</sub> detector or smart locks (Lévy-Bencheton et al. 2015).

When smart home security is exposed to vulnerabilities and threats, these can facilitate criminal actions and intrusions, or could themselves be a form of crime (e.g. physical damage, theft or unauthorised access to smart home assets) (Barnard-Wills et al. 2014).

While there are many other ethical issues that concern smart homes (e.g. access, autonomy, freedom of association, freedom of movement, human touch, informed consent, usability), this case study also further underlines two critical issues connected to the right to life: *security* and *privacy* (Marikyan et al. 2019; Chang et al. 2021). Hackers could spy on people, get access to very personal information and misuse smart-home-connected devices in a harmful manner (Laughlin 2021). Nefarious uses could include the perpetration of identity theft, location tracking, home intrusions and access lock-outs.

The responsibilities for ensuring that smart home devices and services do not suffer from vulnerabilities or attacks are manifold, and lie largely with the manufacturers and service providers, and with users. Users of smart-home-connected devices must carry out their due diligence when purchasing smart devices (by buying from reputable companies with good security track records and ensuring that security is up to the task).

### 6.2.3 Case 3: Adversarial Attacks in Medical Diagnosis

Medical diagnosis, particularly in radiology, often relies on images. Adversarial attacks on medical image analysis systems are a problem (Bortsova et al. 2021) that can put lives at risk. This applies whether the AI system is tasked with the medical diagnosis or whether the task falls to radiologists, as an experiment with mammogram images has shown. Zhou et al. used a generative adversarial network (GAN) model to make intentional modifications to radiology images taken to detect breast cancer (Zhou et al. 2021). The resulting fake images were then analysed by an AI model and by radiologists. The adversarial samples “fool the AI-CAD model to output a wrong diagnosis on 69.1% of the cases that are initially correctly classified by the AI-CAD

model. Five breast imaging radiologists visually identify 29–71% of the adversarial samples” (*ibid*). In both cases, a wrong cancer diagnosis could lead to risks to health and life.

Adversarial attacks are “advanced techniques to subvert otherwise-reliable machine-learning systems” (Finlayson et al. 2019). These techniques, for example by making tiny image manipulations (adversarial noise) to images that might help confirm a diagnosis, guarantee positive trial results or control the rates of medical interventions to the advantage of those carrying out such attacks (Finlayson et al. 2018).

To raise awareness of adversarial attacks, Rahman et al. (2021) tested COVID-19 deep learning applications and found that they were vulnerable to adversarial example attacks. They report that due to the wide availability of COVID-19 data sets, and because some data sets included both COVID-19 patients’ public data and their attributes, they could poison data and launch classified inference attacks. They were able to inject fake audio, images and other types of media into the training data set. Based on this, Rahman et al. (2021) call for further research and the use of appropriate defence mechanisms and safeguards.

The case study and examples mentioned in this section expose the problem of machine and deep learning application vulnerabilities in the healthcare setting. They show that a lack of appropriate defence mechanisms, safeguards and controls would cause serious harm by changing results to detrimental effect.

## 6.3 Ethical Questions

All the case studies raise several ethical issues. Here we discuss some of the core ones.

### 6.3.1 *Human Safety*

To safeguard human safety, which has come to the fore in all three case studies, unwanted harms, risks and vulnerabilities to attack need to be addressed, prevented and eliminated throughout the life cycle of the AI product or service (UNESCO 2021). Human safety is rooted in the value of human life and wellbeing. Safety requires that AI systems and applications should not cause harm through misuse, questionable or defective design and unintended negative consequences. Safety, in the context of AI systems, is connected to ensuring their accuracy, reliability, security and robustness (Leslie 2019). Accuracy refers to the ability of an AI system to make correct judgements, predictions, recommendations or decisions based on data or

models (AI HLEG 2019). Inaccurate AI predictions may result in serious and adverse effects on human life. *Reliability* refers to the ability of a system to work properly using different inputs and in a range of situations, a feature that is deemed critical for both scrutiny and harms prevention (*ibid*). *Security* calls for protective measures against vulnerabilities, exploitation and attacks at all levels: data, models, hardware and software (*ibid*). *Robustness* requires that AI systems use a preventative approach to risk. The systems should behave reliably while minimising unintentional and unexpected harm and preventing unacceptable harm, and at the same time ensuring the physical and mental integrity of humans (*ibid*).

### 6.3.2 Privacy

As another responsible AI principle, privacy (see Chap. 3) is also particularly implicated in the first and second case studies. Privacy, while an ethical principle and human right in itself, intersects with the right to life, liberty and security, and supports it with protective mechanisms in the technological context. This principle, in the AI context, includes respect for the privacy, quality and integrity of data, and access to data (AI HLEG 2019). Privacy vulnerabilities manifest themselves in data leakages which are often used in attacks (Denko 2017). Encryption by itself is not seen to provide “adequate privacy protection” (Apthorpe et al. 2017). AI systems must have appropriate levels of security to prevent unauthorised or unlawful processing, accidental loss, destruction or damage (ICO 2020). They must also ensure that privacy and data protection are safeguarded throughout the system’s lifecycle, and data access protocols must be in place (AI HLEG 2019). Furthermore, the quality and integrity of data are critical, and processes and data sets used require testing at all stages.

### 6.3.3 Responsibility and Accountability

When anything goes wrong, we look for *who* is responsible for making decisions about liability and accountability. Responsibility is seen in terms of ownership and/or answerability. In the cases examined here, responsibility might lie with different entities, depending on their role and/or culpability in the harms caused. The cases furthermore suggest that the allocation of responsibility may not be simple or straightforward. In the case of an intentional attack on an AI system, it may be possible to identify the individual orchestrating it. However, in the case of the autonomous vehicle or that of the smart home, the combination of many contributions and the dynamic nature of the system may render the attempt to attribute the actions of the system difficult, if not impossible.

Responsibility lies not only at the point of harm but goes to the point of inception of an AI system. As the ethics guidelines of the European Commission’s High-Level Expert Group on Artificial Intelligence outline (AI HLEG 2019), companies must

identify the impacts of their AI systems and take steps to mitigate adverse impacts. They must also comply with technical requirements and legal obligations. Where a provider (a natural or legal person) puts a high-risk AI system on the market or into service, they bear the responsibility for it, whether or not they designed or developed it (European Commission 2021).

Responsibility faces many challenges in the socio-technical and AI context (Council of Europe 2019). The first, the challenge of “many hands” (Van de Poel et al. 2012) results as the “development and operation of AI systems typically entails contributions from multiple individuals, organisations, machine components, software algorithms and human users, often in complex and dynamic environments”. (Council of Europe 2019). A second challenge relates to how humans placed in the loop are made responsible for harms, despite having only partial control of an AI system, in an attempt by other connected entities to shirk responsibility and liability. A third challenge highlighted is the unpredictable nature of interactions between multiple algorithmic systems that generate novel and potentially catastrophic risks which are difficult to understand (Council of Europe 2019).

For now, responsibility for acts and omissions in relation to an AI product or service and system-related harms lies with humans. The Montreal Declaration for a Responsible Development of AI (2018) states that the development and use of AI “must not contribute to lessening the responsibility of human beings when decisions must be made”. However, it also provides that “when damage or harm has been inflicted by an AIS [AI system], and the AIS is proven to be reliable and to have been used as intended, it is not reasonable to place blame on the people involved in its development or use”.

Accountability, as outlined by the OECD, refers to.

the expectation that organisations or individuals will ensure the proper functioning, throughout their lifecycle, of the AI systems that they design, develop, operate or deploy, in accordance with their roles and applicable regulatory frameworks, and for demonstrating this through their actions and decision-making process (for example, by providing documentation on key decisions throughout the AI system lifecycle or conducting or allowing auditing where justified). (OECD n.d.)

Accountability, in the AI context, is linked to auditability (assessment of algorithms, data and design processes), minimisation and reporting of negative impacts, addressing trade-offs and conflicts in a rational and methodological manner within the state of the art, and having accessible redress mechanisms (AI HLEG 2019).

But accountability in the AI context is also not without its challenges, as Busuioc (2021) explains. Algorithm use creates deficits that affect accountability: the compounding of informational problems, the absence of adequate explanation or justification of algorithm functioning (limits on questioning this), and ensuing difficulties with diagnosing failure and securing redress. Various regulatory tools have thus become important to boost AI accountability.

## 6.4 Responses

Given the above issues and concerns, it is important to put considerable effort into preventing AI human rights issues arising around life, liberty and security of persons, for which the following tools will be particularly helpful.

### 6.4.1 *Defining and Strengthening Liability Regimes*

An effective liability regime offers incentives that help reduce risks of harm and provide means to compensate the victims of such harms. “Liability” may be defined by contractual requirements, fault or negligence-based liability, or no-fault or strict liability. With regard to self-driving cars, liability might arise from tort for drivers and insurers and from product liability for manufacturers. Different approaches are adopted to reduce risks depending on the type of product or service.

Are current liability regimes adequate for AI? As of 1 April 2022, there were no AI-specific legal liability regimes in the European Union or United States, though there have been some attempts to define and strengthen existing liability regimes to take into account harms from AI (Karner et al. 2021).

The European Parliament’s resolution of 20 October 2020 with recommendations to the European Commission on a civil liability regime for AI (European Parliament 2020) outlined that there was no need for a complete revision of the well-functioning liability regimes in the European Union. However, the capacity for self-learning, the potential autonomy of AI systems and the multitude of actors involved presented a significant challenge to the effectiveness of European Union and national liability framework provisions. The European Parliament recognised that specific and co-ordinated adjustments to the liability regimes were necessary to compensate persons who suffered harm or property damage, but did not favour giving legal personality to AI systems. It stated that while physical or virtual activities, devices or processes that were driven by AI systems might technically be the direct or indirect cause of harm or damage, this was nearly always the result of someone building, deploying or interfering with the systems (European Parliament 2020). Parliament recognised, though, that the Product Liability Directive (PLD), while applicable to civil liability claims relating to defective AI systems, should be revised (along with an update of the Product Safety Directive) to adapt it to the digital world and address the challenges posed by emerging digital technologies. This would ensure a high level of effective consumer protection and legal certainty for consumers and businesses and minimise high costs and risks for small and medium-sized enterprises and start-ups. The European Commission is taking steps to revise sectoral product legislation (Ragonnaud 2022; Šajn 2022) and undertake initiatives that address liability issues related to new technologies, including AI systems.

A comparative law study on civil liability for artificial intelligence (Karner et al. 2021) questioned whether the liability regimes in European Union Member States

provide for an adequate distribution of all risks, and whether victims will be indemnified or remain undercompensated if harmed by the operation of AI technology, even though tort law principles would favour remedying the harm. The study also highlights that there are some strict liabilities in place in all European jurisdictions, but that many AI systems would not fall under such risk-based regimes, leaving victims to pursue compensation via fault liability.

With particular respect to self-driving vehicles, existing legal liability frameworks are being reviewed and new measures have been or are being proposed (e.g. Automated and Electric Vehicles Act 2018; Dentons 2021). These will need to deal with issues that arise from the shifts of control from humans to automated driver assistance systems, and to address conflicts of interest, responsibility gaps (who is responsible and in what conditions, i.e. the human driver/passengers, system operator, insurer or manufacturer) and the remedies applicable.

A mixture of approaches is required to address harms by AI, as different liability approaches serve different purposes: these could include fault- or negligence-based liability, strict liability and contractual liability. The strengthening of provisions for strict liability (liability that arises irrespective of fault or of a defect, malperformance or non-compliance with the law) is highly recommended for high-risk AI products and services (New Technologies Formation 2019), especially where such products and services may cause serious and/or significant and frequent harms, e.g. death, personal injury, financial loss or social unrest (Wendehorst 2020).

#### ***6.4.2 Quality Management for AI Systems***

Given the risks shown in the case studies presented, it is critical that AI system providers have a good quality management system in place. As outlined in detail in the proposal for the Artificial Intelligence Act (European Commission 2021), this should cover the following aspects:

1. a strategy for regulatory compliance ...
2. techniques, procedures and systematic actions to be used for the design, design control and design verification of the high-risk AI system;
3. techniques, procedures and systematic actions to be used for the development, quality control and quality assurance of the high-risk AI system;
4. examination, test and validation procedures to be carried out before, during and after the development of the high-risk AI system, and the frequency with which they have to be carried out,
5. technical specifications, including standards, to be applied and, where the relevant harmonised standards are not applied in full, the means to be used to ensure that the high-risk AI system complies with the requirements set out [in this law];
6. systems and procedures for data management, including data collection, data analysis, data labelling, data storage, data filtration, data mining, data aggregation, data retention and any other operation regarding the data that is performed

- before and for the purposes of the placing on the market or putting into service of high-risk AI systems;
- 7. the risk management system ...
  - 8. the setting-up, implementation and maintenance of a post-market monitoring system ...
  - 9. procedures related to the reporting of serious incidents and of malfunctioning ...
  - 10. the handling of communication with national competent authorities, competent authorities, including sectoral ones, providing or supporting the access to data, notified bodies, other operators, customers or other interested parties;
  - 11. systems and procedures for record keeping of all relevant documentation and information,
  - 12. resource management, including security of supply related measures,
  - 13. an accountability framework setting out the responsibilities of the management and other staff ...

### 6.4.3 *Adversarial Robustness*

Case 3 demonstrates the need to make AI models more robust to adversarial attacks. As an IBM researcher puts it, “Adversarial robustness refers to a model’s ability to resist being fooled” (Chen 2021). This calls for the adoption of various measures, such as the simulation and mitigation of new attacks, via, for example, reverse engineering to recover private data, adversarial training (Tramèr et al. 2018; Bai et al 2021, University of Pittsburgh 2021), using pre-generating adversarial images and teaching the model that these images are manipulated, and designing robust models and algorithms (Dhawale et al. 2022). The onus is clearly on developers to prepare for and anticipate AI model vulnerabilities and threats.

Examples abound of efforts to increase adversarial robustness (Gorsline et al. 2021). Li et al. (2021) have proposed an enhanced defence technique called Attention and Adversarial Logit Pairing (AT + ALP), which, when applied to clean examples and their adversarial counterparts, would help improve accuracy on adversarial examples over adversarial training. Tian et al. (2021) have proposed what they call “detect and suppress the potential outliers” (DSPO), a defence against data poisoning attacks in federated learning scenarios.

## 6.5 Key Insights

The right to life is the baseline of all rights: the first among other human rights. It is closely related to other human rights, including some that are discussed elsewhere in this book, such as privacy (see Chap. 3) or dignity (see Chap. 7).

In the AI context, this right requires AI developers, deployers and users to respect the sanctity of human life and embed, value and respect this principle in the design,

development and use of their products and/or services. Critically, AI systems should *not* be programmed to kill or injure humans.

Where there is a high likelihood of harms being caused, even if accidental, additional precautions must be taken and safeguards set up to avoid them, for example the use of standards, safety-based design, adequate monitoring of the AI system (Anderson 2020), training, and improved accident investigation and reporting (Alemzadeh et al. 2016).

While the technology may have exceeded human expectations, AI must support human life, *not* undermine it. The sanctity of human life must be preserved. What is furthermore required is *sensitivity* to the value of human life, liberty and security. It is insensitivity to harms and impacts that leads to change-resistant problematic actions. Sensitivity requires the ability to understand what is needed and the taking of helpful actions to fulfil that need. It also means remembering that AI can influence, change and damage human life in many ways. This sensitivity is required at all levels: development, deployment and use. It requires continuous learning on the adverse impacts that an AI system may have on human life, liberty and security and avoiding and/or mitigating such impacts to the fullest extent possible.

## References

- ACHPR (1981) African (Banjul) Charter on Human and Peoples' Rights. Adopted 27 June. African Commission on Human and Peoples' Rights, Banjul. [https://www.achpr.org/public/Document/file/English/banjul\\_charter.pdf](https://www.achpr.org/public/Document/file/English/banjul_charter.pdf). Accessed 24 May 2022
- AI HLEG (2019) Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence, European Commission, Brussels. [https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=60419](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419). Accessed 25 Sept 2020
- Alemzadeh H, Raman J, Leveson N et al (2016) Adverse events in robotic surgery: a retrospective study of 14 years of FDA data. PLoS ONE 11(4):e0151470. <https://doi.org/10.1371/journal.pone.0151470>
- Anderson B (2020) Tesla autopilot blamed on Fatal Japanese Model X crash. Carscoops, 30 April. <https://www.carscoops.com/2020/04/tesla-autopilot-blamed-on-fatal-japanese-model-x-crash/>. Accessed 24 May 2022
- AP, Reuters (2021) US regulators probe deadly Tesla crash in Texas. DW, 19 April. <https://p.dw.com/p/3sFbD>. Accessed 22 May 2022
- Apthorpe NJ, Reisman D, Feamster N (2017) A smart home is no castle: privacy vulnerabilities of encrypted IoT traffic. ArXiv, abs/1705.06805. <https://doi.org/10.48550/arXiv.1705.06805>
- Automated and Electric Vehicles Act (2018) c18. HMSO, London. <https://www.legislation.gov.uk/ukpga/2018/18/contents>. Accessed 24 May 2022
- Bai T, Luo J, Zhao J et al (2021) Recent advances in adversarial training for adversarial robustness. In: Zhou Z-H (ed) Proceedings of the thirtieth international joint conference on artificial intelligence (IJCAI-21), International Joint Conferences on Artificial Intelligence, pp 4312–4321. <https://doi.org/10.24963/ijcai.2021/591>
- Barnard-Wills D, Marinos L, Portesi S (2014). Threat landscape and good practice guide for smart home and converged media. European Union Agency for Network and Information Security (ENISA). <https://www.enisa.europa.eu/publications/threat-landscape-for-smart-home-and-media-convergence>. Accessed 25 May 2022

- BBC News (2020) Uber's self-driving operator charged over fatal crash. 16 September. <https://www.bbc.com/news/technology-54175359>. Accessed 23 May 2022
- Bortsova G, González-Gonzalo C, Wetstein SC et al (2021) Adversarial attack vulnerability of medical image analysis systems: unexplored factors. *Med Image Anal* 73:102141. <https://doi.org/10.1016/j.media.2021.102141>
- Boyd EB (2016) Is police use of force about to get worse—with robots? *POLITICO Magazine*, 22 September. <https://www.politico.com/magazine/story/2016/09/police-robots-ethics-debate-214273/>. Accessed 22 May 2022
- Busuioc M (2021) Accountable artificial intelligence: holding algorithms to account. *Public Adm Rev* 81(5):825–836. <https://doi.org/10.1111/puar.13293>
- Chang V, Wang Z, Xu QA et al (2021) Smart home based on internet of things and ethical issues. In: Proceedings of the 3rd international conference on finance, economics, management and IT business (FEMIB), pp 57–64. <https://doi.org/10.5220/0010178100570064>
- Chen P-Y (2021) Securing AI systems with adversarial robustness. *IBM Research*. <https://research.ibm.com/blog/securing-ai-workflows-with-adversarial-robustness>. Accessed 15 May 2022
- Clifford Law (2021) The dangers of driverless cars. *The National Law Review*, 5 May. <https://www.natlawreview.com/article/dangers-driverless-cars>. Accessed 23 May 2022
- Council of Europe (2019) Responsibility and AI: a study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Prepared by the Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT). <https://rm.coe.int/respondability-and-ai-en/168097d9c5>. Accessed 25 May 2022
- Council of Europe (2020) Legal aspects of “autonomous” vehicles. Report Committee on Legal Affairs and Human Rights, Parliamentary Assembly, Council of Europe. <https://assembly.coe.int/LifeRay/JUR/Pdf/DocsAndDecs/2020/AS-JUR-2020-20-EN.pdf>. Accessed 25 May 2022
- Deloitte (n.d.) Urban future with a purpose: 12 trends shaping the future of cities by 2030. <https://www2.deloitte.com/global/en/pages/public-sector/articles/urban-future-with-a-purpose.html>.
- Denko MW (2017) A privacy vulnerability in smart home IoT devices. Dissertation, University of Michigan-Dearborn. [https://deepblue.lib.umich.edu/bitstream/handle/2027.42/139706/49698122\\_ECE\\_699\\_Masters\\_Thesis\\_Denko\\_Michael.pdf](https://deepblue.lib.umich.edu/bitstream/handle/2027.42/139706/49698122_ECE_699_Masters_Thesis_Denko_Michael.pdf). Accessed 25 May 2022
- Dentons (2021) Global guide to autonomous vehicles 2021. <http://www.thedriverlesscommute.com/wp-content/uploads/2021/02/Global-Guide-to-Autonomous-Vehicles-2021.pdf>. Accessed 24 May 2022
- Dhawale K, Gupta P, Kumar Jain T (2022) AI approach for autonomous vehicles to defend from adversarial attacks. In: Agarwal B, Rahman A, Patnaik S et al (eds) Proceedings of international conference on intelligent cyber-physical systems. Springer Nature, Singapore, pp 207–221. [https://doi.org/10.1007/978-981-16-7136-4\\_17](https://doi.org/10.1007/978-981-16-7136-4_17)
- ECHR (1950) European Convention on Human Rights. 5 November. European Court of Human Rights, Strasbourg. [https://www.echr.coe.int/documents/convention\\_eng.pdf](https://www.echr.coe.int/documents/convention_eng.pdf). Accessed 25 May 2022
- European Commission (2021) Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. European Commission, Brussels. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>. Accessed 1 May 2022
- European Parliament (2020) Resolution of 20 October 2020 with recommendations to the commission on a civil liability regime for artificial intelligence (2020/2014(INL)). [https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276\\_EN.pdf](https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.pdf). Accessed 24 May 2022
- Finlayson SG, Bowers JD, Ito J et al (2019) Adversarial attacks on medical machine learning. *Science* 363(6433):1287–1289. <https://doi.org/10.1126/science.aaw4399>
- Finlayson SG, Chung HW, Kohane IS, Beam AL (2018) Adversarial attacks against medical deep learning systems. ArXiv preprint. <https://doi.org/10.48550/arXiv.1804.05296>

- Fránik M, Čermák M (2020) Serious flaws found in multiple smart home hubs: is your device among them? WeLiveSecurity, 22 April. <https://www.welivesecurity.com/2020/04/22/serious-flaws-smart-home-hubs-is-your-device-among-them/>. Accessed 22 May 2022
- Golson J (2017) Read the Florida Highway Patrol's full investigation into the fatal Tesla crash. The Verge, 1 February. <https://www.theverge.com/2017/2/1/14458662/tesla-autopilot-crash-accident-florida-fatal-highway-patrol-report>. Accessed 23 Msay 2022
- Gorsline M, Smith J, Merkel C (2021) On the adversarial robustness of quantized neural networks. In: Proceedings of the 2021 Great Lakes symposium on VLSI (GLSVLSI '21), 22–25 June 2021, virtual event. Association for Computing Machinery, New York, pp 189–194. <https://doi.org/10.1145/3453688.3461755>
- ICO (2020) Guidance on AI and data protection. Information Commissioner's Office, Wilmslow, UK. <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/>. Accessed 25 May 2022
- Jansen P, Brey P, Fox A et al (2020). SIENNA D4.4: Ethical analysis of AI and robotics technologies V1.<https://doi.org/10.5281/zenodo.4068083>
- Jiang BC, Gainer CA Jr (1987) A cause-and-effect analysis of robot accidents. *J Occup Accid* 9(1):27–45. [https://doi.org/10.1016/0376-6349\(87\)90023-X](https://doi.org/10.1016/0376-6349(87)90023-X)
- Kalra N, Groves DG (2017) The enemy of good: estimating the cost of waiting for nearly perfect automated vehicles. Rand Corporation, Santa Monica CA
- Karner E, Koch BA, Geistfeld MA (2021) Comparative law study on civil liability for artificial intelligence. Directorate-General for Justice and Consumers, European Commission, Brussels. <https://data.europa.eu/doi/10.2838/77360>. Accessed 24 May 2022
- Laughlin A (2021) How a smart home could be at risk from hackers. Which?, 2 July. <https://www.which.co.uk/news/article/how-the-smart-home-could-be-at-risk-from-hackers-akeR18s9eBHU>. Accessed 23 May 2022
- 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>
- Levin S, Woolf N (2016) Tesla driver killed while using autopilot was watching Harry Potter, witness says. The Guardian, 1 July. <https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter>. Accessed 23 May 2022
- Lévy-Bencheton C, Darra E, Tétu G et al (2015) Security and resilience of smart home environments. good practices and recommendations. European Union Agency for Network and Information Security (ENISA). <https://www.enisa.europa.eu/publications/security-resilience-good-practices>. Accessed 25 May 2022
- Li X Goodman D Liu J et al (2021) Improving adversarial robustness via attention and adversarial logit pairing. *Front ArtifIntell* 4. <https://doi.org/10.3389/frai.2021.752831>
- Marikyan D, Papagiannidis S, Alamanos E (2019) A systematic review of the smart home literature: a user perspective. *Technol Forecast Soc Change* 138:139–154. <https://doi.org/10.1016/j.techfore.2018.08.015>
- McFarland M (2019) Feds blame distracted test driver in Uber self-driving car death. CNN Business, 20 November. <https://edition.cnn.com/2019/11/19/tech/uber-crash-ntsb/index.html>. Accessed 23 May 2022
- Montreal Declaration (2018) Montréal declaration for a responsible development of artificial intelligence. Université de Montréal, Montreal. <https://www.montrealdeclaration-responsibleai.com/the-declaration>. Accessed 21 Sept 2020
- Muggah R (2017) What happens when we can predict crimes before they happen? World Economic Forum, 2 February. <https://www.weforum.org/agenda/2017/02/what-happens-when-we-can-predict-crimes-before-they-happen/>. Accessed 16 May 2022
- New Technologies Formation (2019) Liability for artificial intelligence and other emerging digital technologies. Expert Group on Liability and New Technologies, Directorate-General for Justice and Consumers, European Commission, Brussels. <https://data.europa.eu/doi/10.2838/573689>. Accessed 24 May 2022

- OAS (2015) Inter-American Convention on Protecting the Human Rights of Older Persons. Forty-fifth regular session of the OAS General Assembly, 15 June. [http://www.oas.org/en/sla/dil/docs/inter\\_american\\_treaties\\_A-70\\_human\\_rights\\_older\\_persons.pdf](http://www.oas.org/en/sla/dil/docs/inter_american_treaties_A-70_human_rights_older_persons.pdf). Accessed 25 May 2022
- OECD (n.d.) Accountability (Principle 1.5). OECD AI Policy Observatory. <https://oecd.ai/en/dashboards/ai-principles/P9>. Accessed 23 May 2022
- Pietsch B (2021) 2 killed in driverless Tesla car crash, officials say. The New York Times, 18 April. <https://www.nytimes.com/2021/04/18/business/tesla-fatal-crash-texas.html>. Accessed 23 May 2022
- Ragonnaud G (2022) Legislative train schedule: revision of the machinery directive (REFIT). European Parliament. <https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-revision-of-the-machinery-directive>. Accessed 24 May 2022
- Rahman A, Hossain MS, Alrajeh NA, Alsolami F (2021) Adversarial examples: security threats to COVID-19 deep learning systems in medical IoT devices. IEEE Internet Things J 8(12):9603–9610. <https://doi.org/10.1109/JIOT.2020.3013710>
- Raso F, Hilligoss H, Krishnamurthy V et al (2018). Artificial intelligence and human rights: opportunities and risks. Berkman Klein Center for Internet and Society Research, Harvard University, Cambridge MA. <http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439>. Accessed 25 May 2022
- Safi M (2019) Are drone swarms the future of aerial warfare? The Guardian, 4 December. <https://www.theguardian.com/news/2019/dec/04/are-drone-swarms-the-future-of-aerial-warfare>. Accessed 22 May 2022
- Šajn N (2022) Legislative train schedule: general product safety regulation. European Parliament. <https://www.europarl.europa.eu/legislative-train/theme-a-new-push-for-european-democracy/file-revision-of-the-general-product-safety-directive>. Accessed 24 May 2022
- Teoh ER, Kidd DG (2017) Rage against the machine? Google's self-driving cars versus human drivers. J Safety Rs 63:57–60. <https://doi.org/10.1016/j.jsr.2017.08.008>
- Tesla (n.d.) Support: autopilot and full self-driving capability. <https://www.tesla.com/support/autopilot>. Accessed 23 May 2022
- Tian Y, Zhang W, Simpson A, et al (2021). Defending against data poisoning attacks: from distributed learning to federated learning. Computer J bxab192. <https://doi.org/10.1093/comjnl/bxab192>
- Tramèr F, Kurakin A, Papernot N et al (2018) Ensemble adversarial training: attacks and defenses. Paper presented at 6th international conference on learning representations, Vancouver, 30 April – 3 May. <https://doi.org/10.48550/arXiv.1705.07204>
- UN (1948) Universal Declaration of Human Rights. <http://www.un.org/en/universal-declaration-human-rights/>. Accessed 4 May 2022
- UN (1966) International Covenant on Civil and Political Rights. General Assembly resolution 2200A (XXI), 16 December. <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>. Accessed 24 May 2022
- UN (1969) American Convention on Human Rights: “Pact of San José, Costa Rica”. Signed at San José, Costa Rica, 22 November. <https://treaties.un.org/doc/publication/unts/volume%201144/volume-1144-i-17955-english.pdf>. Accessed 24 May 2022
- UN (1989) Convention on the Rights of the Child. General Assembly resolution 44/25, 20 November. <https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child>. Accessed 24 May 2022
- UN (1990) International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families. General Assembly resolution 45/158, 18 December. <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-protection-rights-all-migrant-workers>. Accessed 24 May 2022
- UN (2006) Convention on the Rights of Persons with Disabilities. General Assembly resolution A/RES/61/106, 13 December. <https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities>. Accessed 24 May 2005
- UN (2018) Universal Declaration of Human Rights at 70: 30 articles on 30 articles – article 3. Press release, 12 November. Office of the High Commissioner for Human Rights,

- United Nations. <https://www.ohchr.org/en/press-releases/2018/11/universal-declaration-human-rights-70-30-articles-30-articles-article-3>. 24 May 2022
- UNESCO (2021) Recommendation on the ethics of artificial intelligence. SHS/BIO/RECAIETHICS/2021. General Conference, 41st, 23 November. <https://unesdoc.unesco.org/ark:/48223/pf0000380455>. Accessed 25 May 2022
- University of Pittsburgh (2021) Cancer-spotting AI and human experts can be fooled by image-tampering attacks. Science Daily, 14 December. <https://www.sciencedaily.com/releases/2021/12/211214084541.htm>. Accessed 24 May 2022.
- Vasic M, Billard A (2013) Safety issues in human-robot interactions. In: Proceedings of the 2013 IEEE international conference on robotics and automation, Karlsruhe, 6–10 May, pp 197–204. <https://doi.org/10.1109/ICRA.2013.6630576>
- Van de Poel I, Fahlquist JN, Doorn N et al (2012) The problem of many hands: climate change as an example. *Sci Eng Ethics* 18(1):49–67. <https://doi.org/10.1007/s11948-011-9276-0>
- Wendehorst C (2020) Strict liability for AI and other emerging technologies. *JETL* 11(2):150–180. <https://doi.org/10.1515/jetl-2020-0140>
- Zhou Q, Zuley M, Guo Y et al (2021) (2021) A machine and human reader study on AI diagnosis model safety under attacks of adversarial images. *Nat Commun* 12:7281. <https://doi.org/10.1038/s41467-021-27577-x>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# Chapter 7

## Dignity



**Abstract** Dignity is a very prominent concept in human rights instruments, in particular constitutions. It is also a concept that has many critics, including those who argue that it is *useless* in ethical debates. How useful or not dignity can be in artificial intelligence (AI) ethics discussions is the question of this chapter. Is it a conversation stopper, or can it help explain or even resolve some of the ethical dilemmas related to AI? The three cases in this chapter deal with groundless dismissal by an automated system, sex robots and care robots. The conclusion argues that it makes perfect sense for human rights proponents to treat dignity as a prime value, which takes precedence over others in the case of extreme dignity violations such as torture, human trafficking, slavery and reproductive manipulation. However, in AI ethics debates, it is better seen as an equal among equals, so that the full spectrum of potential benefits and harms are considered for AI technologies using all relevant ethical values.

**Keywords** Dignity · AI ethics · Sex robots · Care robots

### 7.1 Introduction

Most human rights instruments protect the inherent dignity of human beings. For instance, the opening of the Universal Declaration of Human Rights states that “recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world” (UN 1948). And the first article of the German constitution (*Grundgesetz*) reads: “(1) Human dignity shall be inviolable. To respect and protect it shall be the duty of all state authority” (Germany 1949).

Given the focus of this book on artificial intelligence (AI) ethics cases that relate to potential human rights infringements, one might think that the concept of dignity would be very useful. And indeed, as will be seen below, dignity has made an entrance into AI ethics discussions.

However, it is also important to note that the meaning of the term “dignity” is highly contested (Schroeder and Bani-Sadr 2017). The Canadian Supreme Court



**Fig. 7.1** Dignity classification

even decided that dignity was no longer to be used in anti-discrimination cases as it was too confusing and difficult to apply.

[H]uman dignity is an abstract and subjective notion that ... cannot only become confusing and difficult to apply; it has also proven to be an additional burden on equality claimants. (Kapp 2008: 22)

To give specific meaning to the concept of dignity in the context of three AI ethics cases, the classification model developed by Schroeder and Bani-Sadr (2018) is summarised below.

Three broad types of dignity can be distinguished: the dignity associated with specific conduct or roles, the intrinsic dignity of all human beings and the critical interpretation, which sees dignity as nothing but a slogan to stop debate (*ibid*: 53) (see Fig. 7.1).

*Aspirational* dignity, associated with specific conduct or roles, is not available to all human beings, and it is possible to distinguish three main varieties.

**Dignity as an expression of virtue** According to Beyleveld and Brownsword (2001: 139), Nelson Mandela exemplifies the personification of dignity as virtue. His fortitude in the face of adversity—throughout decades of imprisonment—deserves almost universal admiration.

**Dignity through rank and position** The original, historical meaning of dignity is related to rank and position within hierarchies. For instance, Machiavelli (2015) believed that “dignity [is conferred] by antiquity of blood” rather than through actions individuals can take, such as being virtuous.

**Dignity of comportment** In Dostoevsky’s *Crime and Punishment* (1917), two impoverished ladies are described whose gloves “were not merely shabby but had holes in them, and yet this evident poverty gave the two ladies an air of special dignity, which is always found in people who know how to wear poor clothes.” Independently of virtue or rank, these two ladies display dignified comportment, which Dostoevsky singles out for praise in the name of dignity.

In contrast to these varieties of aspirational dignity, *intrinsic* dignity is available to *all* human beings and is described in two main ways within Western philosophy and Christian thinking.

**Dignity as intrinsic worth** The most prominent understanding of dignity in Western philosophy today is based on Immanuel Kant’s interpretation of dignity

as intrinsic worth, which is not selective but belongs to all human beings.<sup>1</sup> It cannot be denied even to a vicious man, according to Immanuel Kant (1990: 110 [463]), our translation). Hence, dignity as intrinsic worth is unrelated to virtue and moral conduct. From Kantian-type dignity stems the prohibition against actions that dehumanise and objectify human beings—enshrined in the Universal Declaration of Human Rights (UN 1948)—in the worst cases through slavery, torture or degrading treatment.

**Dignity as being created in the image of God** The Catholic Church also promotes the idea that *all* human beings have dignity, because all human beings, they maintain, “are created in the image and likeness of God” (Markwell 2005: 1132).

These two groups of dignity interpretations (aspirational and intrinsic) are joined by a third, highly critical position. Ruth Macklin (2003) famously argued that dignity “is a useless concept … [that] can be eliminated [from ethics debates] without any loss of content”. Harvard professor Steven Pinker even claimed that dignity is “a squishy, subjective notion” used mostly “to condemn anything that gives someone the creeps” (Pinker 2008). Thinkers in this group believe that dignity is often used as a “conversation stopper” to avoid having an in-depth dialogue about challenging issues (Birnbacher 1996).

How useful dignity might be in AI ethics discussions is the question for this chapter. Is it a conversation stopper, or is it useful in helping understand or even resolve some of the ethical dilemmas related to AI?

## 7.2 Cases of AI in Potential Conflict with Human Dignity

### 7.2.1 Case 1: Unfair Dismissal

“Automation can be an asset to a company, but there needs to be a way for humans to take over if the machine makes a mistake,” says Ibrahim Diallo (Wakefield 2018). Diallo was jobless for three weeks in 2017 (Diallo 2018) after being dismissed by an automated system for no reason his line manager could ascertain. It started with an inoperable access card, which no longer worked for his Los Angeles office, and led to him being escorted from the building “like a thief” (Wakefield 2018) by security staff following a barrage of system-generated messages. Diallo said the message that made him jobless was “soulless and written in red as it gave orders that dictated my fate. Disable this, disable that, revoke access here, revoke access there, escort out of premises … The system was out for blood and I was its very first victim” (*ibid*). After three weeks, his line manager identified the problem (an employee who had left the company had omitted to approve an action) and reinstated Mr Diallo’s contractual rights.

<sup>1</sup> One of us has dealt elsewhere with the challenge that Immanuel Kant bestows dignity only upon *rational* beings (Schroeder and Bani-Sadr 2017).

Commenting on the case, AI expert Dave Coplin noted: “It’s another example of a failure of human thinking where they allow it to be humans versus machines rather than humans plus machines” (*ibid*). Another commentator used the term “dignity” in connection with this case, noting that “the dignity of human beings and their ‘diminishing value’ [is at stake] as we approach the confluence of efficiencies gained from the increasing implementation of artificial intelligence and robotics” (Diallo 2018).

Being escorted from the building like a thief—that is, a criminal—as Diallo put it, and without any wrongdoing on his part, can indeed be interpreted in terms of a loss of dignity. Psychological research has shown that being wrongly accused of criminal offences can have severe consequences for the accused, including for their sense of self and their sense of dignity.

Along with changes in personality, participants also experienced various other losses related to their sense of self, for example loss of *dignity* and credibility ... and loss of hope and purpose for the future [*italics added*]. (Brooks and Greenberg 2020)

Interpreted this way, the dignity lost by Diallo would be aspirational dignity, dignity that is conduct-related. Being publicly suspected of unlawful or immoral behaviour can, then, lead to a sense of having lost dignity. However, the reason why dignity is not helpful in this AI ethics case is that it is unnecessary to the argument. A dignity interpretation does not move the case forward. There is no moral dilemma to be solved. It is obvious that a human being should not be treated like a criminal and made redundant for the sole reason that an opaque system is unresponsive—in this example, to Diallo’s line manager. AI designed to assist with human resources decisions should be understandable, and, as Dave Coplin has noted, should operate on the basis of humans *plus* machines, not humans *versus* machines.

This problem is not unique to AI, a fact that can easily be verified with reference to the success rates of unfair dismissal cases brought by employees. Useful figures are available from Australia, for instance, which radically reformed its dismissal regulations in 2006. Freyens and Oslington (2021) put the success rate of employees claiming unfair dismissal at 47–48% for the period from 2001 to 2015 (a time when cases are unlikely to have been influenced significantly by AI decision-making). Hence, almost half of the employees who challenged employers about their dismissal were deemed to have been unfairly dismissed, as Diallo was, yet most likely without the involvement of AI systems.

The challenge is summarised by Goodman and Kilgallan (2021) from the point of view of employers:

AI will continue to develop and will likely outperform humans in some aspects of working life. However, the technology has wide implications and employers should be cautious. AI has two notable flaws: the human error of thinking AI is infallible and the lack of transparency in its outcomes.

From the point of view of employees like Ibrahim Diallo, it is also vital for the system that made him jobless to significantly increase its transparency.

### 7.2.2 Case 2: Sex Robots

One of the first sex robots available to buy is called Harmony (Boran 2018). The female body of this sex doll is combined with a robotic head, which can turn; the mouth can smile and the eyes can blink. The robot's AI element is steered through an app on the owner's phone. While Harmony cannot stand up, conversation is possible as the app stores information about the owner. The machine doll has been described as "a more sophisticated, sexy Alexa" built on "a very reductive stereotype of femininity: narrow waist, big breasts, curvaceous hips, long blonde hair" (Cosmopolitan 2021). Harmony and other sex robots have been at the centre of highly controversial ethics debates.

The concept of dignity plays a big role in ethics debates about sex robots. Sex robots have been promoted as a way of "achieving 'dignity' ... [by enabling] physical touch, intimacy, and sexual pleasure ... [for] disabled people" and for "men and women rejected sexually by other men, or women" (Zardashvili and Fosch-Villaronga 2020). However, it is also argued that "the everlasting availability as well as the possibility to perform any sexual activity violates gender dignity and equality, causing harm that is understood as objectification and commodification" (Rigotti 2020).

The type of dignity both of these discourses appeal to is *intrinsic dignity*, which requires all human beings to be treated with respect and not objectified. Or, as Immanuel Kant (1998: 110 [4:428]) put it in his principle of humanity,

the human being ... exists as an end in itself, *not merely as a means* to be used by this or that will at its discretion; instead [the human being] ... must ... always be regarded *at the same time as an end*.

Two main routes are available for discussing the potential moral dilemma of sex robots and their impact on the intrinsic dignity of human beings.

The first route links human rights to sexual rights and sexual autonomy. This route can start, for instance, with the World Health Organization's (WHO) strategy to count "sexual ... well-being" as "fundamental to the overall health and well-being of individuals" and therefore related to article 25 of the Universal Declaration of Human Rights (WHO n.d.b) (see box).

#### Universal Declaration of Human Rights, Article 25

"Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family." (UN 1948)

Historically, the sexual rights movement focused on the prevention of harm, in particular emphasising the rights of girls and women to be free from sexual violence.

Later, the same movement focused on the rights to self-expression of sexual inclination without fear of discrimination for lesbian, gay and transgender people (Miller 2009).

In 2002, going beyond the focus on harm and discrimination, a WHO-commissioned report defined sexual rights as “human rights that are already recognized in national laws, international human rights documents and other consensus statements. They include the right of all persons ... to ... pursue a satisfying, safe and pleasurable sexual life” (WHO n.d.a). This sexual right to a pleasurable sexual life is limited by the injunction not to infringe the rights of others (*ibid*). The step from this type of sexual autonomy to sex robots is short.

If sexual freedom is an integral part of personal autonomy and interference is illegitimate whenever consent and private acts are involved, then robotic sexual intercourse will take place at home, coming to no direct harm to others and falling within the buyer’s right to privacy. (Rigotti 2019)

Despite some highly optimistic predictions about the benefits of sex robots—e.g. they will allegedly fill the void in the lives of people who have no-one and therefore provide a “terrific service” to humankind (Wiseman 2015)—the linchpin in this discussion will be the consideration of potential harm. And this is the starting point of the second route for discussing the potential moral dilemma of sex robots and their impact on the intrinsic dignity of human beings.

This second route assumes as its *starting* point that the harm from sex robots is inevitable. For instance, Professor Kathleen Richardson’s Campaign Against Sex Robots warns against “reinforcing female dehumanisation” and seeks “to defend the dignity and humanity of women and girls” (CASR n.d.). In her view, sex robots will strengthen a relationship model where buyers of sex can turn off empathy towards the sellers of sex and cement relations of power where one party is not recognised as a human being, but simply as a needs fulfills framed according to male desires (Richardson 2015): a sole means, not simultaneously an end in herself, in Kantian terminology.

Richardson’s views are strengthened by the fact that sex robots are almost exclusively female (Rigotti 2019), that one can speak of “a market by men for men” (Cosmopolitan 2021) and that “the tech’s development is largely ... focused on the fulfillment of straight male desire” (Edwards 2016). Sex robots are thus seen as an extension of sex work (Richardson uses the term “prostitution”), where “the buyer of sex is at liberty to ignore the state of the other person as a human subject who is turned into a thing” (Richardson 2015).

At this point, most commentators will note that sex robots *are* things, they do not have to be objectified *into* things. Somewhat cynically, one could ask: what harm does it do if a fanatical Scarlett Johansson admirer builds a sex robot in her image, which winks and smiles at him (Pascoe 2017)? It’s only a thing, and it might even do some good. For instance, it has been suggested that sex robots, used within controlled environments, could help redirect “the sexual behavior of high-risk child molesters ... without endangering real children” (Zara et al. 2022).



**Fig. 7.2** Golden mean on sex robot positions?

The two routes to discussing the ethical issues of sex robots sketched above could also be illustrated as a continuum with two opposing poles which meet where sexual autonomy that does not create harm might be the Aristotelian golden mean (Aristotle 2000: 102 [1138b]). (Aristotle argued that one should always strive to find the middle between excess and deficiency: courage, for instance, lying between recklessness and cowardice (Aristotle 2000: 49 [1115b]) (see Fig. 7.2).

Ensuring that sex robots are used without creating harm might pose serious challenges, which are discussed further below. First, we look at care robots.

### 7.2.3 Case 3: Care Robots

An old lady sits alone in her sheltered accommodation stroking her pet robot seal. She has not had any human visitors for days. A humanoid robot enters the room, delivers a tray of food, and leaves after attempting some conversation about the weather, and encouraging her to eat it all up. The old lady sighs, and reluctantly complies with the robot's suggestions. When she finishes eating, she goes back to stroking the pet robot seal: 'At least you give my life some meaning' she says. (Sharkey 2014)

Amanda Sharkey paints this picture of an old lady and her care robot in her paper "Robots and Human Dignity: A Consideration of the Effects of Robot Care on the Dignity of Older People".

Despite the emphasis on dignity in the title, she then delivers a very even-handed risk and benefit analysis of care robots, with only limited reference to potential dignity violations. In particular, she distinguishes three types of care robots for the elderly: first, assistive robots, which can, for instance, help with feeding and bathing, or moving a person with limited mobility from a bed to a wheelchair; second, monitoring robots, which can, for instance, detect falls, manage diaries or provide reminders about taking medication; and third, companion robots, which often take the form of pets such as the robot seal in the case description.

In her discussion of the ethical challenges of care robots, Sharkey establishes only three connections with dignity. *Assistive* robots can increase human dignity for elderly people, she believes, in particular by increasing mobility and access to social

interaction. *Monitoring* robots, likewise, can increase human dignity, she argues, by enabling elderly people to live independently for longer than otherwise possible. The only ethical challenge related to human dignity Sharkey identifies is related to *companion* robots, which could infantilise elderly people in the eyes of carers or undermine their self-respect, if offered as a sole replacement for human interaction.

Arguably, the concept of dignity does no important work in the first two cases. If a technology can increase mobility and access to social interaction, and enable elderly people to live independently for longer than otherwise possible, the positive contributions to wellbeing are easily understood without reference to the contested concept of dignity (Stahl and Coeckelberg 2016). A similar point could be made using research by Robinson et al. (2014):

[O]lder people indicated that they would like a robot for detecting falls, controlling appliances, cleaning, medication alerts, making calls and monitoring location. Most of these tasks point towards maintaining independence and dignity.

If one removed the term “dignity” in this quote, which is a technique suggested by dignity critics to ascertain whether the concept is useful or not (Macklin 2003), then the potential benefit of robotic technology in elderly care remains, namely maintaining independence.

Perhaps unsurprisingly, then, when Stahl and Coeckelbergh (2016) examine exactly the same problem (the ethical challenges of health and care robots) they make no reference to dignity at all. Where they align with Sharkey, but use different wording, is on the question of “cold and mechanical” machine care, which might be seen as abandoning elderly people and handing them “over to robots devoid of human contact” (*ibid*). Stahl and Coeckelberg (*ibid*) ask whether this might be an objectification of care receivers. In other words, they employ the Kantian concept of dignity, also used by Richardson to justify her Campaign Against Sex Robots, to ask whether elderly people who are cared for by robots are objectified—in other words, turned from subjects into things.

### 7.3 Ethical Questions Concerning AI and Dignity

There is something behind all three AI cases that is not easy to describe and warrants ethical attention. It is linked to how people are seen by others and how this relates to their own self-respect. Jean-Paul Sartre (1958: 222) used the concept of the *gaze* to describe this situation.

[T]he Other is the indispensable mediator between myself and me. ... By the mere appearance of the Other, I am put in the position of passing judgement on myself as an object, for it is as an object that I appear to the other.

It was the fact that Diallo was seen by others and possibly judged as a wrongdoer that made his forced removal from his workplace a potential dignity issue. The concern that sex robots may reinforce human relationships that see women and girls

as mere needs fulfillers of male sexual desires and not as human subjects in their own right is what seems to drive Richardson's campaign to ban sex robots. And it is the fact that elderly people in care who engage with their robot pets are possibly seen as infantilised in the eyes of carers that is one of Sharkey's main concerns.

The judgement-filled gaze of the other and the link to self-respect as one passes judgement on oneself through the eyes of the other is what makes the three AI cases above relevant to dignity debates. "Dignity" therefore seems a suitable word to describe at least some of the moral dilemmas involved in the cases. One might hence argue that dignity is not a mere conversation stopper, but a helpful concept in the context of AI ethics. Let us examine this further.

It is suggested above that the concept of dignity is not necessary in drawing useful ethical conclusions from the first case, that of unfair dismissal due to an opaque AI system. This position would maintain that there is no moral dilemma as there are no proponents of competing claims who have to find common ground. Unfair dismissal due to opaque AI ought to be avoided.<sup>2</sup> Any technical or organisational measures (e.g. ethics by design, see Sect. 2.4.2) that can reasonably be used, should be used to achieve this goal. At the same time, it is essential that remedies be available to employees and workers who find themselves in a situation similar to Diallo's. Thus, a report presented to the UK Trades Union Congress (Allen and Masters 2021: 77) stressed the following:

Unfair dismissal legislation should protect employees ... from dismissal decisions that are factually inaccurate or opaque in the usual way. The use of AI-powered tools to support such decisions does not make any difference to this important legal protection.

Cases 2 and 3 are more complex in terms of their dignity angle, especially Case 2, sex robots. We shall first discuss Case 3 to provide additional leads for Case 2.

In "Oh, Dignity Too?" Said the Robot: Human Dignity as the Basis for the Governance of Robotics", Zardiashvili and Fosch-Villaronga (2020) identify eight major ethical concerns in employing care robots for the elderly: safe human–robot interaction, the allocation of responsibility, privacy and data protection loss, autonomy restriction, deception and infantilisation, objectification and loss of control, human–human interaction decrease, and long-term consequences. While the authors recognise that "[r]obots might be the solution to bridge the loneliness that the elderly often experience; they may help wheelchair users walk again, or may help navigate the blind" (*ibid*), they also "acknowledge that human contact is an essential aspect of personal care and that ... robots for healthcare applications can challenge the dignity of users" (*ibid*).

We have themed seven of the above concerns into five headings and removed the eighth as it applies to all emerging technologies ("Technology ... may have long-term consequences that might be difficult to foresee"). For our interpretation of Zardiashvili and Fosch-Villaronga's (2020) ethical concerns in relation to care robots and the elderly see Fig. 7.3.

---

<sup>2</sup> Unfair dismissal due to performance being measured by inaccurate algorithms is not covered here, as it is more relevant to Chap. 2 on discrimination than this one on dignity. For relevant literature, see De Stefano (2018).



**Fig. 7.3** Ethical concerns about care robots and the elderly

As Fig. 7.3 shows, dignity is only one of five main ethical concerns identified in relation to care robots and the elderly. This is not unusual. It is often the case that several ethical values or principles need to be protected to achieve an optimal ethical outcome. These values may even conflict. For instance, it may be that the safety of a device could be increased through a more invasive collection of personal data. Stakeholders then have to weigh up the relative importance of safety versus the relative importance of privacy. It is in this weighing-up process that dignity may be an outlier value.

As the founding principle of many constitutions around the world, dignity is often given precedence over all other values. For instance, in the Daschner case (Schroeder 2006), the German constitutional court ruled in 2004 that a threat of duress by police forces to extract information was an unacceptable violation of the dignity of a detainee, even though the police forces were as certain as they could be that the child kidnapper in their custody was the only person who could reveal the whereabouts of an 11-year old who might be starving in an undetected location.<sup>3</sup> The presiding judge noted: “Human dignity is inviolable. Nobody must be made into an object, a bundle of fear” (Rückert 2004, our translation). The inviolability of the value of dignity meant it took precedence over all other values, including what the police forces thought was the right to life of a child.

If the power given to dignity in constitutional courts were to colour other moral debates, such as those on care robots and the elderly, important ethical factors might be ignored. Dignity could then indeed become a conversation stopper, overriding safety, privacy, autonomy and liability issues and moving discussions away from potential benefits. It has already been noted by Sharkey (2014) that (assistive) care robots can increase the wellbeing of elderly people by improving mobility and access to social interaction. Likewise, (monitoring) care robots can increase human wellbeing by enabling elderly people to live independently for longer than otherwise possible.

In the case of care robots, potential dignity issues therefore have to be treated as one of several ethical challenges, without being given privileged importance. Only then can the ethical risks for the technology be addressed proportionally in the context of serious social care staff shortages. Taking the UK as an example, around ten per cent of social care posts were vacant in 2020 and an additional need for 650,000 to

<sup>3</sup> In fact, the child had already been killed by the kidnapper.

950,000 new adult social care jobs is anticipated by 2035 (Macdonald 2020). It is in this context that the development of care robots for the elderly might become an ethical goal in its own right.

As a geriatric nurse commented, “Care robots don’t substitute for the human being—they help when no one else is there to help” (Wachsmuth 2018).

Two aspects of the care robot discussion are useful when turning to sex robots. First, unless dealing with extreme dignity violations, such as torture, human trafficking, slavery or reproductive manipulation (Bourcard 2004), the ethical value of dignity should be treated as an equal among equals and not as an ethical value that automatically overrides all others. For instance, those who demand a ban on sex robots to promote the dignity of women and girls (Richardson 2015) seem to make a categorical claim without considering the ethical value of sexual autonomy. This position has been criticised from a range of angles, e.g. that it relies on unjustified parallels to sex work or that it ignores the demand for male sex dolls and toys (Hancock 2020).

The position we want to take here seeks to achieve the Aristotelian mean between two extremes that would, on the one hand, ban sex robots and, on the other, declare them an important service to humankind (see Fig. 7.2). Instead, the right to “pursue a satisfying, safe and pleasurable sexual life” (WHO n.d.a) may include the use of sex robots to foster sexual autonomy, while researchers and regulators should monitor both potential harms *and* potential benefits.

Benefits could, for instance, involve improving “the satisfaction of the sexual needs of a user” (Fosch-Vilaronga and Poulsen 2020) who might have difficulties accessing alternatives, thereby contributing to their health and wellbeing.

Harms could range from very practical considerations, such as sexually transmitted infections from sex robots employed by multiple users in commercial sex work settings (Hancock 2020), to seeking responses to actions that might be criminalised if a human person were involved. One of the most controversially discussed topics in this regard is child sex robots for use by paedophiles.

In July 2014, the roboticist Ronald Arkin suggested that child sex robots could be used to treat those with paedophilic predilections in the same way that methadone is used to treat heroin addicts. ... But most people seem to disagree with this idea, with legal authorities in both the UK and US taking steps to outlaw such devices. (Danaher 2019b)

Using the UK example, since 2017, the Crown Prosecution Service has outlawed the import of child sex dolls, referring to the 1876 Customs Consolidation Act, which forbids the importation of obscene items (Danaher 2019a). This approach is applicable to child sex robots, but other legal avenues have been suggested, in particular using the UK child protection framework or the UK’s 2003 Sexual Offences Act to forbid the use of child sex robots (Chatterjee 2020). On the other hand, “proponents of love and sex with robots would argue that a CSB [childlike sexbot] could have a twofold interest: protecting children from sexual predators and by the same token, treating the latter” (Behrendt 2018). One solution between the two extremes is to restrict the use of child sex robots to cases requiring medical authorisation and under strict medical supervision (*ibid*).

It is noteworthy that the concept of harm rather than that of dignity is usually evoked in the case of child sex robots, which also leads to our conclusion. All ethical aspects that would normally be considered with an emerging technology have to be considered in the process so that dignity is not used as a conversation stopper in assessing this case.

Robot technology may have moral implications, contribute to the loss of human contact, reinforce existing socio-economic inequalities or fail in delivering good care. (Fosch-Villaronga and Poulsen 2020)

We therefore argue that the use of sex robots should not be ruled out categorically based on dignity claims alone.

## 7.4 Key Insights

From the early days of drafting human rights instruments, dignity seems to be the concept that has succeeded in achieving consensus between highly diverse negotiators (Schroeder 2012). One negotiator may interpret dignity from a religious perspective, another from a philosophical perspective and yet another from a pragmatic perspective (Tiedemann 2006). This is possible because dignity does not seem to be ideologically fixed in its meaning, and thus allows a basic consensus between different world views.

This advantage could, however, become a problem in AI ethics debates that are concerned with human rights, if dignity considerations are given the power to override all other ethical values. While this would make perfect sense to human rights proponents in the case of extreme dignity violations such as torture, human trafficking, slavery and reproductive manipulation (Bourcarte 2004), dignity is better seen as an equal among equals in AI ethics debates, especially given the risk of losing it altogether, an approach recommended by those who believe dignity is a mere slogan.

The dignity of the elderly is an important consideration in the design and employment of care robots, but it should not be a conversation stopper in the case of sex robots. As with other ethical dilemmas, the full spectrum of potential benefits and harms needs to be considered using all relevant ethical values. In the case of sex robots this can range from the empowerment of “persons with disabilities and older adults to exercise their sexual rights, which are too often disregarded in society” (Fosch-Villaronga and Poulsen 2020) to restrictive regulation for sex robots that enable behaviour that would be criminal in sex work (Danaher 2019b).

## References

- Allen R, Masters D (2021) Technology managing people: the legal implications. Trades Union Congress, London. [https://www.tuc.org.uk/sites/default/files/Technology\\_Managing\\_People\\_2021\\_Report\\_AW\\_0.pdf](https://www.tuc.org.uk/sites/default/files/Technology_Managing_People_2021_Report_AW_0.pdf). Accessed 13 May 2022
- Aristotle (2000) Nicomachean Ethics (trans: Crisp R). Cambridge University Press, Cambridge
- Behrendt M (2018) Reflections on moral challenges posed by a therapeutic childlike sexbot. In: Cheok AD, Levy D (eds) Love and sex with robots. Springer Nature Switzerland, Cham, pp 96–113. [https://doi.org/10.1007/978-3-319-76369-9\\_8](https://doi.org/10.1007/978-3-319-76369-9_8)
- Beyveld D, Brownsword R (2001) Human dignity in bioethics and biolaw. Oxford University Press, Oxford
- Birnbacher D (1996) Ambiguities in the concept of Menschenwürde. In: Bayertz K (ed) Sanctity of life and human dignity. Kluwer Academic, Dordrecht, pp 107–121. [https://doi.org/10.1007/978-94-009-1590-9\\_7](https://doi.org/10.1007/978-94-009-1590-9_7)
- Boran M (2018) Robot love: the race to create the ultimate AI sex partner. The Irish Times, 1 November. <https://www.irishtimes.com/business/technology/robot-love-the-race-to-create-the-ultimate-ai-sex-partner-1.3674387>. Accessed 13 May 2022
- Bourcard K (2004) Folter im Rechtsstaat? Die Bundesrepublik nach dem Entführungsfall Jakob von Metzler. Self-published, Gießen, Germany. [http://www.bourcard.eu/texte/folter\\_im\\_rechtsstaat.pdf](http://www.bourcard.eu/texte/folter_im_rechtsstaat.pdf). Accessed 16 June 2021
- Brooks SK, Greenberg N (2020) Psychological impact of being wrongfully accused of criminal offences: a systematic literature review. *Med Sci Law* 61(1): 44–54. <https://doi.org/10.1177/0025802420949069>
- CASR (n.d.) Our story. Campaign Against Sex Robots. <https://campaignagainstsexrobots.org/our-story/>. Accessed 13 May 2022
- Chatterjee BB (2020) Child sex dolls and robots: challenging the boundaries of the child protection framework. *Int Rev Law Comput Technol* 34(1):22–43. <https://doi.org/10.1080/13600869.2019.1600870>
- Cosmopolitan (2021) Sex robots: how do sex robots work and can you buy a sex robot? Cosmopolitan, 12 July. <https://www.cosmopolitan.com/uk/love-sex/sex/a36480612/sex-robots/>. Accessed 13 May 2022
- Danaher J (2019a) How should we regulate child sex robots: restriction or experimentation? 4 February (blog). *BMJ Sex Reprod Health*. <https://blogs.bmjjournals.com/bmjsrh/2020/02/04/child-sex-robots/>. Accessed 13 May 2022
- Danaher J (2019b) Regulating child sex robots: restriction or experimentation? *Med Law Rev* 27(4):553–575. <https://doi.org/10.1093/medlaw/fwz002>
- De Stefano V (2018) “Negotiating the algorithm”: automation, artificial intelligence and labour protection. Employment Working Paper No 246. International Labour Office, Geneva. [https://www.ilo.org/wcms5/groups/public/---ed\\_emp/---emp\\_policy/documents/publication/wcms\\_634157.pdf](https://www.ilo.org/wcms5/groups/public/---ed_emp/---emp_policy/documents/publication/wcms_634157.pdf). Accessed 14 May 2022
- Diallo I (2018) The machine fired me: no human could do a thing about it! <https://idiallo.com/blog/when-a-machine-fired-me>. Accessed 12 May 2022
- Dostoevsky F (1917) Crime and punishment (trans: Garnett C). PF Collier & Son, New York. <http://www.bartleby.com/318/32.html>. Accessed 12 May 2022
- Edwards S (2016) Are sex robots unethical or just unimaginative as hell? Jezebel, 7 April. <https://jezebel.com/are-sex-robots-unethical-or-just-unimaginative-as-hell-1769358748>. 13 May 2022
- Fosch-Villaronga E, Poulsen A (2020) Sex care robots: exploring the potential use of sexual robot technologies for disabled and elder care. *Paladyn* 11:1–18. <https://doi.org/10.1515/pjbr-2020-0001>
- Freyens BP, Oslington P (2021) The impact of unfair dismissal regulation: evidence from an Australian natural experiment. *Labour* 35(2):264–290. <https://doi.org/10.1111/labr.12193>

- Germany (1949) Basic Law for the Federal Republic of Germany. Federal Ministry of Justice and Federal Office of Justice, Berlin. [https://www.gesetze-im-internet.de/englisch\\_gg/englisch\\_gg.pdf](https://www.gesetze-im-internet.de/englisch_gg/englisch_gg.pdf). Accessed 12 May 2022
- Goodman T, Kilgallon P (2021) The risks of using AI in employment processes. People Management, 28 September. <https://www.peoplemanagement.co.uk/article/1741566/the-risks-using-ai-employment-processes>. Accessed 13 May 2022
- Hancock E (2020) Should society accept sex robots? Changing my perspective on sex robots through researching the future of intimacy. Paladyn 11:428–442. <https://doi.org/10.1515/pjbr-2020-0025>
- Kant I (1990) Metaphysische Anfangsgründe der Tugendlehre. Felix Meiner Verlag, Hamburg
- Kant I (1998) Groundwork of the metaphysics of morals. Cambridge University Press, Cambridge MA
- Kapp RV (2008) 2 SCR 483. <https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/5696/index.do>. Accessed 12 May 2022
- Macdonald M (2020) The health and social care workforce gap. Insight, 10 January. House of Commons Library, London. <https://commonslibrary.parliament.uk/the-health-and-social-care-workforce-gap/>. Accessed 13 May 2022
- Machiavelli N (2015) The prince (trans: Marriott WK). Wisehouse Classics, Sweden
- Macklin R (2003) Dignity is a useless concept. BMJ 327:1419–1420. <https://doi.org/10.1136/bmj.327.7429.1419>
- Markwell H (2005) End-of-life: a Catholic view. Lancet 366:1132–1135. [https://doi.org/10.1016/s0140-6736\(05\)67425-9](https://doi.org/10.1016/s0140-6736(05)67425-9)
- Miller A (2009) Sexuality and human rights: discussion paper. International Council on Human Rights Policy, Versoix. [https://biblioteca.cejamericas.org/bitstream/handle/2015/654/Sexuality\\_Human\\_Rights.pdf](https://biblioteca.cejamericas.org/bitstream/handle/2015/654/Sexuality_Human_Rights.pdf). Accessed 13 May 2022
- Pascoe A (2017) This Scarlett Johansson Robot is uncomfortably realistic. Marie Claire Australia, 10 May. <https://www.marieclaire.com.au/scarlett-johansson-robot-sex-doll>. Accessed 13 May 2022
- Pinker S (2008) The stupidity of dignity. The New Republic, 28 May. <https://newrepublic.com/article/64674/the-stupidity-dignity>. Accessed 12 May 2022
- Richardson K (2015) The asymmetrical ‘relationship’: parallels between prostitution and the development of sex robots. ACM SIGCAS Comput Soc 45(3):290–293. <https://doi.org/10.1145/287429.2874281>
- Rigotti C (2020) How to apply Asimov’s first law to sex robots. Paladyn 11:161–170. <https://doi.org/10.1515/pjbr-2020-0032>
- Rigotti C (2019) Sex robots: a human rights discourse? OpenGlobalRights, 2 May. <https://www.openglobalrights.org/sex-robots-a-human-rights-discourse/>. Accessed 13 May 2022
- Robinson H, MacDonald B, Broadbent E (2014) The role of healthcare robots for older people at home: a review. Int J of Soc Robot 6:575–591. <https://doi.org/10.1007/s12369-014-0242-2>
- Rückert S (2004) Straflos schuldig. Die Zeit, 22 December. [https://www.zeit.de/2004/53/01\\_Leiter\\_2](https://www.zeit.de/2004/53/01_Leiter_2). Accessed 13 May 2022
- Sartre J-P (1958) Being and nothingness (trans: Barnes HE). Methuen & Co, London
- Schroeder D (2006) A child’s life or a “little bit of torture”? State-sanctioned violence and dignity. Camb Q Healthc Ethics 15(2):188–201. <https://doi.org/10.1017/S0963180106060233>
- Schroeder D (2012) Human rights and human dignity. Ethical Theory Moral Pract 15:323–335. <https://doi.org/10.1007/s10677-011-9326-3>
- Schroeder D, Bani-Sadr A-H (2017) Dignity in the 21st century: middle east and west. Springer Int Publishin AG, Cham. <https://doi.org/10.1007/978-3-319-58020-3>
- Sharkey A (2014) Robots and human dignity: a consideration of the effects of robot care on the dignity of older people. Ethics Inf Technol 16:63. <https://doi.org/10.1007/s10676-014-9338-5>
- Stahl BC, Coeckelberg M (2016) Ethics of healthcare robotics: towards responsible research and innovation. Robot Auton Syst 86:152–161. <https://doi.org/10.1016/j.robot.2016.08.018>
- Tiedemann P (2006) Was ist Menschenwürde? Wissenschaftliche Buchgesellschaft, Darmstadt
- UN (1948) Universal declaration of human rights. <http://www.un.org/en/universal-declaration-human-rights/>. Accessed 4 May 2022

- Wachsmuth I (2018) Robots like me: challenges and ethical issues in aged care. *Front Psychol* 9:432. <https://doi.org/10.3389/fpsyg.2018.00432>
- Wakefield J (2018) The man who was fired by a machine. BBC News, 21 June. <https://www.bbc.com/news/technology-44561838>. Accessed 12 May 2022
- WHO (n.d.b) Sexual health: overview. World Health Organization. [https://www.who.int/health-topics/sexual-health#tab=tab\\_1](https://www.who.int/health-topics/sexual-health#tab=tab_1). Accessed 13 May 2022
- WHO (n.d.a) Sexual and reproductive health: gender and human rights. World Health Organization. [https://www.who.int/reproductivehealth/topics/gender\\_rights/sexual\\_health/en/](https://www.who.int/reproductivehealth/topics/gender_rights/sexual_health/en/). Accessed 25 November 2021
- Wiseman E (2015) Sex, love and robots: is this the end of intimacy? *The Observer*, 13 December. <https://www.theguardian.com/technology/2015/dec/13/sex-love-and-robots-the-end-of-intimacy>. Accessed 13 May 2022
- Zara G, Veggi S, Farrington DP (2022) *Int J Soc Robot* 14:479–498. <https://doi.org/10.1007/s12369-021-00797-3>
- Zardiashvili L, Fosch-Villaronga E (2020) “Oh, dignity too?” said the robot: human dignity as the basis for the governance of robotics. *Minds Mach* 30:121–143. <https://doi.org/10.1007/s11023-019-09514-6>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# Chapter 8

## AI for Good and the SDGs



**Abstract** In 2015, 193 nations came together to agree Agenda 2030: 17 goals ranging from the elimination of poverty to the building of partnerships to achieve those goals. The spirit of the UN Sustainable Development Goals (SDGs) is to leave no one behind. Artificial intelligence (AI) has a great potential to assist in reaching the SDGs. For instance, using algorithms on new and vast agricultural data sets can improve the efficiency of agriculture practices and thereby contribute to SDG 1, “Zero hunger”. However, the high energy consumption, computational resources and levels of expertise required for AI can exacerbate existing inequalities. At the same time, potentially useful AI applications such as seasonal climate forecasting have led to the accelerated laying off of workers in Peru and credit denial to poor farmers in Zimbabwe and Brazil. If AI for Good is to be truly realised, AI’s potential to worsen inequality, to overexploit resources, to be undertaken through “helicopter research” and to focus on SDG issues relevant mainly to high-income countries must be overcome, ideally in close collaboration and engagement with potential beneficiaries in resource-limited settings.

**Keywords** AI for good · AI ethics · SDGs · Helicopter research

### 8.1 Introduction

Artificial intelligence (AI) is one of the few emerging technologies that are very prominently linked to the UN Sustainable Development Goals (SDGs) (UN n.d.a). Through 17 goals and 169 targets, 193 nations resolved “to end poverty and hunger everywhere; to combat inequalities within and among countries; [and] to build peaceful, just and inclusive societies” by 2030 (UN 2015). One could perhaps even argue that AI has been linked directly to international justice and sustainability through the SDGs.

“AI for Good” is a UN-led digital platform<sup>1</sup> that identifies AI solutions to problems relevant to the SDGs. The site offers mostly information about big data sets

---

<sup>1</sup> <https://ai4good.org/>.

provided as links to other sites. For instance, data sets on primary energy production and consumption as well as renewable energy data are provided in relation to SDG 7, “Affordable and clean energy”. More unusual links from the AI for Good platform include AI-generated photographs designed to increase empathy with distant strangers. To achieve this increase in empathy, AI calculations transformed pictures of a Boston neighbourhood into images reminiscent of a war-ravaged Syrian city. The results of this DeepEmpathy project were linked to the AI for Good site under SDG 1, “Zero poverty” (Scalable Cooperation n.d.).

A similar initiative, AI for SDGs,<sup>2</sup> led by the Chinese Academy of Sciences, also collates projects from around the world, mapped onto individual SDGs. For instance, an Irish project using remote sensing data, Microsoft Geo AI Data Science Virtual Machines and GIS mapping “to develop machine learning models that can identify agricultural practices” leading to a decline of bees was linked to SDG 2, “Zero hunger”, and SDG 15, “Life on land” (AI for SDGs Think Tank 2019).

Many philosophers and ethicists make a distinction between doing no harm and doing good. Prominently, Immanuel Kant distinguished the two by referring to perfect and imperfect duties. According to Kant, certain actions such as lying, stealing and making false promises can never be justified, and it is a perfect duty not to commit those acts (Kant 1965: 14f [397f]). Even without understanding the complicated Kantian justification for perfect duties (categorical imperatives, ibid 42f [421f]), one should find complying with this ethical requirement straightforward, by doing no intentional harm. Imperfect duties, on the other hand, are more difficult to comply with, as they are open-ended. Kant also calls them virtue-based (Kant 1990: 28f [394f]). How much help to offer to the needy is a typical example. Until one has exhausted one’s own resources? Or by giving 10% of one’s own wealth?

This Kantian distinction is also prominent in the law and everyday morality; as “in both law and ordinary moral reasoning, the avoidance of harm has priority over the provision of benefit” (Keating 2018). AI for Good would then fall into the second category, providing benefits, and by implication become an area of ethics and morality, which is more difficult to assess.

Both case studies are examples of trying to provide benefits. However, as they are drawn from the real world, they blur the lines of the Kantian distinctions. The first case study also illustrates direct harm to vulnerable populations, and the second illustrates a high likelihood of potential harm and a lack of care and equity in international collaboration.

While the intentionality of harm is decisive for Kant when assessing moral actions, lack of due diligence or care has long been identified as a shortcoming in ethical action (Bonnitcha and McCorquodale 2017). (Kant famously said that there is nothing in the world that is ethical or good per se other than a good will—Kant 1965: 10 [393].) Similarly, the 2000-year-old “*Primum non nocere, secundum cavere, tertium sanare*” (Do no harm, then act cautiously, then heal) (Weckbecker 2018) has been employed in the twenty-first century to describe responsible leadership in business and innovation (Leisinger 2018: 120–122). Technologies can often be used for purposes that were not

---

<sup>2</sup> <https://ai-for-sdgs.academy/>.

originally foreseen or intended, which is why responsiveness and care are required in responsible innovation today (Owen et al. 2013: 35).

## 8.2 Cases of AI for Good or Not?

“Farming 4.0”, “precision agriculture” and “precision farming” (Auernhammer 2001) are all terms used to express, among other things, the employment of big data and AI in agriculture. The International Society of Precision Agriculture has defined precision agriculture as follows:

Precision agriculture is a management strategy that gathers, processes and analyzes temporal, spatial and individual data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production. (ISPA n.d.)

Precision agriculture even has its own academic journal, which covers topics from machine learning methods for crop yield prediction (Burdett and Weller 2022) to neural networks for irrigation management (Jimenez et al. 2021).

In the service of precision agriculture, AI is useful, for example, in processing vast data amounts for weather forecasting, climate monitoring and decadal predictions (climate predictions of up to a decade) with the ultimate aim of increasing forecast quality (Dewitte et al. 2021). Examples of the benefits of increased forecast quality could be earlier evacuation in the case of severe weather incidents such as tornados and reduced irrigation if future rainfall could be forecast with high precision.

### 8.2.1 Case 1: Seasonal Climate Forecasting in Resource-Limited Settings

Seasonal climate forecasting (SCF) is used to predict severe weather, such as droughts and floods, in order to provide policymakers and farmers with the means to address problems in an anticipatory rather than a reactive manner (Klemm and McPherson 2017). Lemos and Dilling (2007) have argued that the benefits of SCF mostly reach those “that are already more resilient, or more resource-rich … in terms of … ability to cope with hazards and disasters”. By contrast, those who are most at risk of being pushed below the poverty line by severe weather have been harmed in cases in Zimbabwe, Brazil and Peru. In Zimbabwe and Brazil, poor farmers were denied credit after SCF results predicted a drought (*ibid*). In Zimbabwe, “misinterpretation of the probabilistic nature of the forecast by the banking sector” might have played a role in decision-making about credits (Hammer et al. 2001). SCF forecasting in Peru also led to accelerated layoffs of workers in the fishing industry due to “a forecast of El Niño and the prospect of a weak season.” (Lemos and Dilling 2007)

Agenda 2030, the underlying framework for the SDGs, makes the following commitment: “Leave no one behind” (UNSDG n.d.). The case above shows that some of those most in need of not being left behind have suffered as a result of new seasonal climate forecasting techniques.

SDG 9 focuses on fostering innovation and notes in its first target that affordable and equitable access to innovation *for all* should be aimed for (UN n.d.a). While the above cases from Zimbabwe, Brazil and Peru precede the SDGs, the potential for AI to “exacerbate inequality” has since been identified as a major concern for Agenda 2030 (Vinuesa et al. 2020). We will return to this problem after the second case.

### 8.2.2 Case 2: “Helicopter Research”

In 2014, a research team from higher-income countries requested access to vast amounts of mobile phone data from users in Sierra Leone, Guinea and Liberia to track population movements during the Ebola crisis. They argued that the value of such data was undeniable in the public health context of the Ebola crisis (Wesolowski 2014). Other researchers disagreed and maintained that quantified population movements would not reveal how the Ebola virus spread (Maxmen 2019). As no ethics guidelines on providing access to mobile phone data existed in Sierra Leone, Guinea and Liberia, government time was spent deliberating whether to provide such access. This time expended on debating access rights, it was argued, “could have been better spent handling the escalating crisis” (*ibid*). Liberia decided to deny access owing to privacy concerns (*ibid*) and the research was undertaken on mobile phone data from Sierra Leone. The data showed that fewer people travelled during the Ebola travel ban, but it did not assist in tracking Ebola. (*ibid*)

One could analyse this case ethically from a harm perspective too, if valuable government time was indeed lost that could have been used to handle the Ebola crisis, as one case commentator argued. One could also analyse it in the context of *potential* harm from privacy breaches when researchers obtain big data sets from countries that have limited means to ensure privacy, especially during a crisis. So-called data-for-good projects have “analysed calls from tens of millions of phone owners in Pakistan, Bangladesh, Kenya and at least two dozen other low- and middle-income nations” (Maxmen 2019) and it has been argued that

concerns are rising over the lack of consent involved; the potential for breaches of privacy, even from anonymized data sets; and the possibility of misuse by commercial or government entities interested in surveillance. (*ibid*)

However, we will analyse the case from the perspective of “helicopter research”, defined thus:

The practice of Global North … researchers making roundtrips to the Global South … to collect materials and then process, analyze, and publish results with little to no involvement from local collaborators is referred to as “helicopter research” or “parachute research”. (Haelewaters et al. 2021)

Helicopter research thrives in crisis. For instance, during the same 2014 Ebola crisis a social scientist from the North collected social science data without obtaining ethics approval for his research, taking undue advantage of the fragile national regulatory framework for overseeing research (Tegli 2017). Before the publication of his results, the researcher realised that he would need research ethics approval to publish. He had already left the country and asked a research assistant to make the case for retrospective approval. The approval was denied by the relevant research ethics committee (*ibid*).

One of the main problems of helicopter research is the lack of involvement of local researchers, potentially leading to colonial assumptions about what will help another country best. Often benefits for researchers from the Global North are clear (e.g. access to data, publications, research grants), while benefits might not materialise at all locally, in the Global South (Schroeder et al. 2021). We will return to this in the next section, but here, in the context of obtaining large-scale phone data during a crisis, we can cite a news feature in *Nature* reporting that

researchers ... say they have witnessed the roll-out of too many technological experiments during crises that don't help the people who most need it. ... [A] digital-governance researcher ... cautions that crises can be used as an excuse to rapidly analyse call records without frameworks first being used to evaluate their worth or to assess potential harms. (Maxmen 2019)

### 8.3 Ethical Questions Concerning AI for Good and the SDGs

At first sight, AI for Good seems to deserve celebration, especially when linked to the SDGs. And it is likely that praise is warranted for many efforts, possibly most (Caine 2020). However, the spectre of inequities and unintended harm due to helicopter research or a lack of due diligence looms large. AI for Good may be reminiscent of other efforts where technological solutions have been given precedence over alternatives and where local collaborators have not been consulted, or have even been excluded from contributing.

Another similarly named movement is called GM for Good,<sup>3</sup> and examples of helicopter research on the application of genetically modified (GM) technologies in resource-limited settings are not hard to find.

In 2014, a US university aimed to produce a transgenic banana containing beta-carotene to address vitamin A deficiency in Uganda. Later the research was abandoned for ethical reasons. During the human food trials conducted among US-based students, safety issues and undue inducement concerns materialised. However, the study also raised concerns in Uganda, in particular about the potential release of a transgenic fruit, the risks of undermining local food and cultural systems, and the risks of reducing banana agrobiodiversity. Uganda is home to non-modified banana

---

<sup>3</sup> <https://gm4good.org/>.

varieties that are already higher in beta-carotene than the proposed transgenic variety. Uninvited intrusions into local food systems that were not matched to local needs were unwelcome and considered inappropriate (Van Niekerk and Wynberg 2018).

Analysing the problems of building GM solutions for populations on the poverty line, Kettenburg et al. (2018) made the following suggestion in the context of Golden Rice, another contentious example (Kettenburg et al. 2018):

To transcend the reductionism of regarding rice as mere nutrient provider, neglecting its place in the eco- and cultural system ... and of describing vitamin A-deficient populations as passive victims ... we propose to reframe the question: from “how do we create a rice plant producing beta-carotene?” ... to “how do we foster the well-being of people affected by malnutrition, both in short and long terms?”

AI for Good can also be susceptible to the weaknesses of helicopter research and reductionism for the following five reasons.

### ***8.3.1 The Data Desert or the Uneven Distribution of Data Availability***

AI relies on data. Machine learning and neural networks are only possible with the input of data. Data is also a highly valuable resource (see Chap. 4 on surveillance capitalism). In this context, a South African report speaks of the “data desert”, with worrying figures such as that statistical capacity has *decreased* over the past 15 years in 11 out of 48 African countries (University of Pretoria 2018: 31). This is highly relevant to the use of AI in the context of SDGs. For instance, Case 1 used the records of mobile phone calls during a crisis to track population movements. “However, vulnerable populations are less likely to have access to mobile devices” (Rosman and Carman 2021).

The data desert has at least two implications. First, if local capacity is not available to generate a sufficient amount of data for AI applications in resource-limited settings, it might have to be generated by outsiders, for example researchers from the Global North “helicoptering” into the region. Second, such helicopter research has then the potential to increase the digital divide, as local capacities are left undeveloped. (See below for more on the digital divide.) In this context, Shamika N Sirimanne, Director of Technology and Logistics for the UN Conference on Trade and Development, says, “As the digital economy grows, a data-related divide is compounding the digital divide” (UNCTAD 2021).

### ***8.3.2 The Application of Double Standards***

Helicopter research can in effect be research that is *only* carried out in lower-income settings, as it would not be permitted, or would be severely restricted, in higher-income settings, for instance due to the potential for privacy breaches from the large-scale processing of mobile phone records. For example, there is no evidence in the literature of any phone tracking research having been used during the catastrophic 2021 German floods in the Ahr district, even though almost 200 people died and it took weeks to track all the deceased and all the survivors (Fitzgerald et al. 2021). One could speculate that it would have been very hard to obtain consent to gather mobile phone data, even anonymised data, for research from a German population, even in a crisis setting.

### ***8.3.3 Ignoring the Social Determinants of the Problems the SDGs Try to Solve***

SDG 2 “Zero hunger” refers to a long-standing problem that Nobel economics laureate Amartya Sen ascribed to entitlement failure rather than a shortage of food availability (Sen 1983). He used the Bengal famine of 1943 to show that the region had more food in 1943 than in 1941, when no famine was experienced. To simplify the argument, the first case study above could be called a study of how the social determinants of hunger were ignored. By trying to improve the forecasting of severe weather in order to give policymakers and farmers options for action in anticipation of failed crops, SCF overlooks the fact that this information, in the hands of banks and employers, could make matters even worse for small-scale farmers and seasonal labourers. That is because the latter have no resilience or resources for addressing food shortages (Lemos and Dilling 2007), the social determinants of hunger.

Another example. An AI application has been developed that identifies potential candidates for pre-exposure prophylaxis in the case of HIV (Marcus et al. 2020). Pre-exposure prophylaxis refers to the intake of medication to prevent infection with HIV. However, those who might need the prophylaxis the most can experience major adherence problems related to SDG 2 “Zero hunger”, such as this patient explained.

When you take these drugs, you feel so hungry. They are so powerful. If you take them on an empty stomach they just burn. I found that sometimes I would just skip the drugs, but not tell anyone. These are some of the things that make it difficult to survive. (Nagata et al. 2012)

An AI solution on its own, without reference to the social determinants of health such as local food security, might therefore not succeed for the most vulnerable segments of populations in resource-limited settings. The type of reductionism attributed to Golden Rice and the Uganda banana scenario described above is likely to occur as well when AI for Good researchers tackle SDGs without local collaborators, which leads to yet another challenge, taking Africa as an example.

### **8.3.4 *The Elephant in the Room: The Digital Divide and the Shortage of AI Talent***

AI depends on high quality broadband. This creates an obvious problem for Africa: given the continent's many connectivity challenges, people must be brought online before they can fully leverage the benefits of AI. (University of Pretoria 2018: 27)

Only an estimated 10% of Africans in rural areas have access to the internet, a figure that goes up to just 22% for urban populations (*ibid*). These figures are dramatic enough, but the ability to develop AI is another matter altogether. Analysing the potential of AI to contribute to achieving the SDGs, a United Nations Development Programme (UNDP) publication notes that the “chronic shortage of talent able to improve AI capabilities, improve models, and implement solutions” is a critical bottleneck (Chui et al. 2019).

Chronic shortage of AI talent is a worldwide challenge, even for large commercial set-ups. For instance, DeLoitte has commented that “companies [in the US] across all industries have been scrambling to secure top AI talent from a pool that’s not growing fast enough.” (Jarvis 2020) Potential new staff with AI capabilities are even lured from universities before completing their degrees to fill the shortage (Kamil 2019).

At the same time, a partnership such as 2030Vision, whose focus is the potential for AI to contribute to achieving the SDGs, is clear about what that requires.

Training significantly more people in the development and use of AI is essential ... We need to ensure we are training and supporting people to apply AI to the SDGs, as these fields are less lucrative than more commercially-oriented sectors (e.g. defense and medicine). (2030Vision 2019: 17)

Yet even universities in high-income countries are struggling to educate the next generation of AI specialists. In the context of the shortage of AI talent, one university executive speaks of “a ‘missing generation’ of academics who would normally teach students and be the creative force behind research projects”, but who are now working outside of the university sector (Kamil 2019).

To avoid the potential reductionism of helicopter research in AI for Good, local collaborators are essential, yet these need to be competent local collaborators who are trained in the technology. This is a significant challenge for AI, owing to the serious shortage of workers, never mind trainers.

### **8.3.5 *Wider Unresolved Challenges Where AI and the SDGs Are in Conflict***

Taking an even broader perspective, AI and other information and communication technologies (ICTs) might challenge rather than support the achievement of SDG 13, which focuses on climate change. Estimates of electricity needs suggest that “up to

20% of the global electricity demand by 2030” might be taken up by AI and other ICTs, a much higher figure than today’s 1% (Vinuesa et al. 2020).

An article in *Nature* argues that AI-powered technologies have a great potential to create wealth, yet argues that this wealth “may go mainly to those already well-off and educated while … others [are left] worse off” (*ibid*). The five challenges facing AI for Good that we have enumerated above must be seen in this context.

Preventing helicopter research and unintentional harm to vulnerable populations in resource-limited settings is one of the main aims of the Global Code of Conduct for Research in Resource-Poor Settings (TRUST 2018) (see also Schroeder 2019). Close collaboration with local partners and communities throughout all research phases is its key ingredient. As we went to press, the journal *Nature* followed major funders (e.g. the European Commission) and adopted the code in an effort “to improve inclusion and ethics in global research collaborations” and “to dismantle systemic legacies of exclusion” (*Nature* 2022).

## 8.4 Key Insights

Efforts by AI for Good are contributing to the achievement of Agenda 2030 against the background of a major digital divide and a shortage of AI talent, potentially leading to helicopter research that is not tailored to local needs. This digital divide is just one small phenomenon characteristic of a world that distributes its opportunities extremely unequally. According to Jeffrey Sachs, “there is enough in the world for everyone to live free of poverty and it won’t require a big effort on the part of big countries to help poor ones” (Xinhua 2018). But this help cannot be dispensed colonial-style to be effective; it has to be delivered in equitable collaborations with local partners and potential beneficiaries.

What all the challenges facing AI for Good described in this chapter have in common is the lack of equitable partnerships between those who are seeking solutions for the SDGs and those who are meant to benefit from the solutions. The small-scale farmers and seasonal workers whose livelihoods are endangered as a result of the application of seasonal climate forecasting, as well as the populations whose mobile phone data are used without proper privacy governance, are meant to be beneficiaries of AI for Good activities, yet they are not.

A saying usually attributed to Mahatma Gandhi expresses it this way: “Whatever you do for me but without me, you do against me.” To make AI for Good truly good for the SDGs, AI’s potential to “exacerbate inequality”, its potential for the “over-exploitation of resources” and its focus on “SDG issues that are mainly relevant in those nations where most AI researchers live and work” (Vinuesa et al. 2020) must be monitored and counteracted, ideally in close collaboration and engagement with potential beneficiaries in resource-limited settings.

## References

- AI for SDGsThink Tank (2019) Curbing the decline of wild and managed bees. International Research Center for AI Ethics and Governance, Institute of Automation, Chinese Academy of Sciences. <https://ai-for-sdgs.academy/case/151>. Accessed 19 May 2022
- Auerhammer H (2001) Precision farming: the environmental challenge. *Comput Electron Agric* 30(1–3):31–43. [https://doi.org/10.1016/S0168-1699\(00\)00153-8](https://doi.org/10.1016/S0168-1699(00)00153-8)
- Bonnitcha J, McCorquodale R (2017) The concept of ‘due diligence’ in the UN guiding principles on business and human rights. *Eur J Int Law* 28(3):899–919. <https://doi.org/10.1093/ejil/chx042>
- Burdett H, Wellen C (2022) Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. *Precision Agric*. <https://doi.org/10.1007/s11119-022-09897-0>
- Caine M (2020) This is how AI could feed the world’s hungry while sustaining the planet. World Economic Forum, 24 September. <https://www.weforum.org/agenda/2020/09/this-is-how-ai-could-feed-the-world-s-hungry-while-sustaining-the-planet/>. Accessed 20 May 2022
- Chui M, Chung R, Van Heteren, A (2019) Using AI to help achieve Sustainable Development Goals. United Nations Development Programme, 21 January. <https://www.undp.org/blog/using-ai-help-achieve-sustainable-development-goals>. Accessed 20 May 2022
- Dewitte S, Cornelis JP, Müller R, Munteanu A (2021) Artificial intelligence revolutionises weather forecast, climate monitoring and decadal prediction. *Remote Sens* 13(16):3209. <https://doi.org/10.3390/rs13163209>
- Fitzgerald M, Angerer C, Smith P (2021) Almost 200 dead, many still missing after floods as Germany counts devastating cost. NBC News, 19 July. <https://www.nbcnews.com/news/world/almost-200-dead-many-still-missing-after-floods-germany-counts-n1274330>. Accessed 20 May 2022
- Haelewaters D, Hofmann TA, Romero-Olivares AL (2021) Ten simple rules for Global North researchers to stop perpetuating helicopter research in the Global South. *PLoS Comput Biol* 17(8):e1009277. <https://doi.org/10.1371/journal.pcbi.1009277>
- Hammer GL, Hansen JW, Phillips JG et al (2001) Advances in application of climate prediction in agriculture. *Agric Syst* 70(2–3):515–553. [https://doi.org/10.1016/S0308-521X\(01\)00058-0](https://doi.org/10.1016/S0308-521X(01)00058-0)
- ISPA (n.d.) Precision ag definition. International Society of Precision Agriculture, Monticello IL. <https://www.ispag.org/about/definition>. Accessed 19 May 2022
- Jarvis D (2020) The AI talent shortage isn’t over yet. Deloitte Insights, 30 September. <https://www2.deloitte.com/us/en/insights/industry/technology/ai-talent-challenges-shortage.html>. Accessed 20 May 2022
- Jimenez A-F, Ortiz BV, Bondesan L et al (2021) Long short-term memory neural network for irrigation management: a case study from southern Alabama, USA. *Precision Agric* 22:475–492. <https://doi.org/10.1007/s11119-020-09753-z>
- Kamil YA (2019) Will AI’s development be hindered by a talent shortage in academia? Study International, 14 October. <https://www.studyinternational.com/news/ai-professionals-talent-shortage/>. Accessed 20 May 2022
- Kant I (1965) Grundlegung zur Metaphysik der Sitten. Felix Meiner Verlag, Hamburg
- Kant I (1990) Metaphysische Anfangsgründe der Tugendlehre. Felix Meiner Verlag, Hamburg
- Keating GC (2018) Principles of risk imposition and the priority of avoiding harm. *Rebus* 36:7–39. <https://doi.org/10.4000/rebus.4406>
- Kettenburg AJ, Hanspach J, Abson DJ, Fischer J (2018) From disagreements to dialogue: unpacking the Golden Rice debate. *Sustain Sci* 13:1469–1482. <https://doi.org/10.1007/s11625-018-0577-y>
- Klemm T, McPherson RA (2017) The development of seasonal climate forecasting for agricultural producers. *Agric for Meteorol* 232:384–399. <https://doi.org/10.1016/j.agrformet.2016.09.005>
- Leisinger K (2018) Die Kunst der verantwortungsvollen Führung. Haupt Verlag, Bern
- Lemos MC, Dilling L (2007) Equity in forecasting climate: can science save the world’s poor? *Sci Public Policy* 34(2):109–116. <https://doi.org/10.3152/030234207X190964>

- Marcus JL, Sewell WC, Balzer LB, Krakower DS (2020) Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic. *Curr HIV/AIDS Rep* 17(3):171–179. <https://doi.org/10.1007/s11904-020-00490-6>
- Maxmen A (2019) Can tracking people through phone-call data improve lives? *Nature*, 29 May. <https://www.nature.com/articles/d41586-019-01679-5>. Accessed 20 May 2022
- Nagata JM, Magerenge RO, Young SL et al (2012) Social determinants, lived experiences, and consequences of household food insecurity among persons living with HIV/AIDS on the shore of Lake Victoria Kenya. *AIDS Care* 24(6):728–736. <https://doi.org/10.1080/09540121.2011.630358>
- Nature (2022) Nature addresses helicopter research and ethics dumping. 2 June. <https://www.nature.com/articles/d41586-022-01423-6>. Accessed 30 May 2022
- Owen R, Stilgoe J, Macnaghten P, Gorman M et al (2013) A framework for responsible innovation. In: Owen R, Bessant J, Heintz M (eds) Responsible innovation: managing the responsible emergence of science and innovation in society. John Wiley & Sons, Chichester, pp 27–50. <https://doi.org/10.1002/9781118551424.ch2>
- Rosman B, Carman M (2021) Why AI needs input from Africans. *Quartz Africa*, 25 November. <https://qz.com/africa/2094891/why-ai-needs-input-from-africans/>. Accessed 20 May 2022
- Scalable Cooperation (n.d.) Project Deep Empathy. School of Architecture and Planning, Massachusetts Institute of Technology. <https://www.media.mit.edu/projects/deep-empathy/overview/>. Accessed 18 May 2022
- Schroeder D, Chatfield K, Muthuswamy V, Kumar NK (2021) Ethics dumping: how not to do research in resource-poor settings. *Academics Stand Against Poverty* 1(1):32–55. <http://journals.asap.org/index.php/asap/article/view/4>. Accessed 20 May 2022
- Schroeder D, Chatfield K, Singh M et al (2019) Equitable research partnerships: a global code of conduct to counter ethics dumping. Springer Nature, Cham, Switzerland. <https://doi.org/10.1007/978-3-030-15745-6>
- Sen A (1983) Poverty and famines: an essay on entitlement and deprivation. Oxford University Press, New York
- Tegli JK (2017) Seeking retrospective approval for a study in resource-constrained Liberia. In: Schroeder D, Cook J, Hirsch F et al (eds) Ethics dumping. SpringerBriefs in Research and Innovation Governance. Springer, Cham, pp 115–119. [https://doi.org/10.1007/978-3-319-64731-9\\_14](https://doi.org/10.1007/978-3-319-64731-9_14)
- 2030Vision (2019) AI & the Sustainable Development Goals: the state of play. Sustainability, London. <https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/2030vision-stateofplay.pdf>. Accessed 20 May 2022
- TRUST (2018) Global Code of Conduct for Research in Resource-Poor Settings, <https://doi.org/10.48508/GCC/2018.05>
- UN (n.d.a) Goals: 9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation. United Nations Department of Economic and Social Affairs: Sustainable Development. <https://sdgs.un.org/goals/goal9>. Accessed 20 May 2022
- UN (n.d.b) The 17 goals. United Nations Department of Economic and Social Affairs: Sustainable Development. <https://sdgs.un.org/goals>. Accessed 18 May 2022
- UN (2015) Transforming our world: the 2030 Agenda for Sustainable Development. Resolution adopted by the General Assembly on 25 September 2015. Res 70/1, 21 October. <https://documents-dds-ny.un.org/doc/UNDOC/GEN/N15/291/89/PDF/N1529189.pdf?OpenElement>. Accessed 18 May 2022
- UNCTAD (2021) Inequalities threaten wider divide as digital economy data flows surge. UN Conference on Trade and Development, Geneva. <https://unctad.org/news/inequalities-threaten-wider-divide-digital-economy-data-flows-surge>. Accessed 20 May 2022
- University of Pretoria (2018) Artificial intelligence for Africa: an opportunity for growth, development, and democratisation. Access Partnership. <https://www.accesspartnership.com/cms/access-content/uploads/2018/11/WP-AI-for-Africa.pdf>. Accessed 20 May 2022
- UNSDG (n.d.) Leave no one behind. UN Sustainable Development Group. <https://unsgd.un.org/2030-agenda/universal-values/leave-no-one-behind>. Accessed 20 May 2022

- Van Niekerk J, Wynberg R. (2018) Human food trial of a transgenic fruit. In: Schroeder D, Cook J, Hirsch F et al (eds) Ethics dumping. SpringerBriefs in Research and Innovation Governance. Springer, Cham, pp 91–98. [https://doi.org/10.1007/978-3-319-64731-9\\_11](https://doi.org/10.1007/978-3-319-64731-9_11)
- Vinuesa R, Azizpour H, Leite I et al (2020) The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11:233. <https://doi.org/10.1038/s41467-019-14108-y>
- Weckbecker K (2018) Nicht schaden – vorsichtig sein – heilen. MMW Fortschr Med 160:36. <https://doi.org/10.1007/s15006-018-0481-5>
- Wesolowski A, Buckee CO, Bengtsson L et al (2014) Commentary: containing the Ebola outbreak: the potential and challenge of mobile network data. PLoS Curr 6. <https://doi.org/10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e>
- Xinhua (2018) Greed is biggest obstacle to achieving fair societies, professor says at UN. Xinhuanet, 10 July. [http://www.xinhuanet.com/english/2018-07/10/c\\_137313107.htm](http://www.xinhuanet.com/english/2018-07/10/c_137313107.htm). Accessed 20 May 2022

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



## Chapter 9

# The Ethics of Artificial Intelligence: A Conclusion



**Abstract** The concluding chapter highlights broader lessons that can be learned from the artificial intelligence (AI) cases discussed in the book. It underlines the fact that, in many cases, it is not so much the technology itself that is the root cause of ethical concerns but the way it is applied in practice *and* its reliability. In addition, many of the cases do not differ radically from ethics cases related to other novel technologies, even though the use of AI can exacerbate existing concerns. Ethical issues can rarely be resolved to everybody's full satisfaction, not least because they often involve the balancing of competing goods. What is essential is space for human reflection and decision-making within the use of AI. Questions about what we can and should do, why we should act in particular ways and how we evaluate the ethical quality of our actions and their outcomes are part of what it means to be human. Even though Immanuel Kant believed that a good will is the only thing in the world that is ethical per se, a good will alone does not suffice where complex consequences may not be obvious. The complex nature of AI systems and their interaction with their human, social and natural environment require constant vigilance and human input.

**Keywords** AI ethics · Socio-technical systems · AI ecosystem · Solutions · Mitigation

This book of case studies on ethical issues in artificial intelligence (AI), and strategies and tools to overcome them, has provided an opportunity for learning about AI ethics. Importantly, it has also shown that AI ethics does not normally deal with clear-cut cases. While some cases provide examples of events that are obviously wrong from an ethical perspective, such cases are often about the *reliability* of the technology. For instance, it is obvious that AI-enabled robots should not present health, safety and security risks for users such as the death of a passenger in a self-driving car, or the smart-home system which allowed a man-in-the-middle attack. More difficult are cases where deliberation on the ethical pros and cons does not provide an immediate answer for the best approach—for instance, where robot use in elderly care reduces pressure on seriously overstretched staff but outsources important human contact

to machines, or where sex robots can be seen as violating the dignity of humans (especially girls and women) but at the same time helping realise sexual rights.

Looking at the cases across the different example domains in this book, one can make some general observations. The first refers to the application context of AI. Our case studies have aimed to be grounded in existing or realistic AI technologies, notably currently relevant machine learning. The ethical relevance of the cases, however, is almost always linked to the way in which the machine learning tool is applied and integrated into larger systems. The ethical concerns, then, are not focused on AI but on the way in which AI is used and the consequences this use has. For instance, the unfair dismissal case (Chap. 7) and the gender bias case (Chap. 2) are about the application of AI. Both the dismissal of staff without human input into the sequence and the training of AI devices on gender-biased CVs are about the use of AI. This is not to suggest that AI is an ethically neutral tool, but rather to highlight that the broader context of AI use, which includes existing moral preferences, social practices and formal regulation, cannot be ignored when undertaking ethical reflection and analysis.

This raises the question: how do AI ethics cases differ from other cases of technology ethics? As a first approximation it is probably fair to say that they usually do not differ radically. Many of the ethics case studies we present here are not fundamentally novel and we do not introduce issues that have never been considered before. For instance, the digital divide discussed in Chap. 8 has been debated for decades. However, the use of AI can *exacerbate* existing concerns and heighten established problems.

AI in its currently predominant form of machine learning has some characteristics that set it apart from other technologies, notably its apparent ability to classify phenomena, which allows it to make or suggest decisions, for example when an autonomous vehicle decides to brake because it classifies an object as an obstacle in the road, or when a law enforcement system classifies an offender as likely to commit a further crime despite a model rehabilitation record. This is often seen as an example of AI autonomy. It is important, however, to see that this autonomy is not an intrinsic part of the machine learning model but an element of the way it is integrated into the broader socio-technical system, which may or may not allow these classifications in the model to affect social reality. Autonomy is thus a function not of AI, but of the way in which AI is implemented and integrated into other systems. Ibrahim Diallo might not have been dismissed by a machine and escorted from the company building like a thief (see Chap. 7) if the AI system had been more transparent and required more human input into the dismissal process.

Indeed, another characteristic of current AI based on neural networks is their opacity. It is precisely the strength of AI that it can produce classifications without humans having to implement a model; that is, the system develops its own model. This machine learning model is frequently difficult or impossible for humans to scrutinise and understand. Opacity of this kind is often described as a problem and various approaches around explainable AI are meant to address it and give meaningful insight into what happens within an AI system. This raises questions about what constitutes explainability and explanations more broadly, including questions

about the explainability of ethical decisions: questions that may open up new avenues in moral philosophy. And while explainability is generally agreed to be an important aspect of AI ethics, one should concede that most individuals have as little understanding of how their internal combustion engine or microwave oven works as they have of the internal workings of an AI system they are exposed to. For internal combustion engines and microwave ovens, we have found ways to deal with them that address ethical concerns, which raises the question: how can similar approaches be found and implemented for AI?

A final characteristic of current AI systems is the need for large data sets in the training and validating of models. This raises questions about ownership of and access to data that relate to the existing distribution of economic resources, as shown in Chap. 4. As data sets often consist of personal data, they may create the potential for new threats and aggravate privacy and data protection harms. This may also entrench power imbalances, giving more power to those who control such information. Access to data may also be misused to poison models, which can then be used for nefarious purposes. But while AI offers new mechanisms to misuse technology, misuse itself is certainly not a new phenomenon.

What overarching conclusions can one draw from this collection of cases of ethically problematic uses of AI and the various interpretations of these issues and proposed responses and mitigation strategies?

A first point worth highlighting is that human interaction typically results in ethical questions. Adding AI to human interaction can change the specific ethical issues, but will not resolve all ethical issues or introduce completely unexpected ones. Ethical reflection on questions of what we can and should do, why we should act in particular ways and how we evaluate the ethical quality of our actions and their outcomes are part of what it means to be human. Even though Immanuel Kant believed that a good will is the only thing in the world that is ethical per se, a good will alone does not suffice where complex consequences may not be immediately obvious. For instance, as shown in Chap. 8 about AI for Good, the most vulnerable populations might be hit harder by climate change, rather than helped, as a result of the use of AI-based systems. This was the case with small-scale farmers in Brazil and Zimbabwe who were not granted credit to cope with climate change by bank managers who had access to forecasts from seasonal climate prediction. Likewise, seasonal workers in Peru were laid off earlier based on seasonal climate forecasting. In these cases, helicopter research to aid vulnerable populations in resource-limited settings ought to be avoided, as local collaborators are likely to be in a better position to predict impacts on vulnerable populations.

Ethical issues can rarely be resolved to everybody's full satisfaction, not least because they often involve the balancing of competing goods. AI raises questions such as how to balance possible crime reduction through better prediction against possible discrimination towards disadvantaged people. How do we compare access to novel AI-driven tools with the ability and motivations of the tool holders to benefit from the use of our personal data? Or what about the possibility of improving medical diagnoses amid crippling human resource shortages versus the downsides of automated misdiagnosis? Can an uncertain chance of fighting a pandemic through AI

analysis justify large-scale data collection? How could one justify the deployment of AI in resource-limited areas in the light of the intrinsic uncertainty and unpredictability of the consequences this may have on different parts of the population? What about the elderly lady whose only companion is a pet robot? All our cases can be described in terms of such competing goods, and it is rarely that a simple response can be given. The conclusion to be drawn from this is that awareness of ethical issues and the ability and willingness to reflect on them continuously need to be recognised as a necessary skill of living in technology-driven societies.

Another conclusion to be drawn from our examples is that the nature of AI as a system needs to be appreciated and included explicitly in ethical reasoning. AI is never a stand-alone artefact but is always integrated into larger technical systems that form components of broader socio-technical systems ranging from small-scale local systems in individual organisations all the way up to global systems such as air traffic control or supply chains. This systemic nature of AI means that it is typically impossible to predict the consequences of AI use accurately. That is a problem for ethical theory, which tends to work on the assumption that consequences of actions are either determined or at least statistically distributed in a way that can be accurately described. One consequence of this lack of clear causal chains in large-scale socio-technical systems is that philosophy could aim to find new ways of ethical reflection of systems.

In practice, however, as our description of the responses to the cases has shown, there is already a significant number of responses that promise to be able to lead to a better understanding of AI ethics and to address ethical issues. These range from individual awareness, AI impact assessments, ethics-by-design approaches, the involvement of local collaborators in resource-limited settings and technical solutions such as those linked to AI explainability, all the way to legal remedies, liability rules and the setting up of new regulators. None of these is a panacea which can address the entire scope of AI ethics by itself, but collectively and taken together they offer a good chance to pre-empt the significant ethical problems or prevent them from having disastrous consequences. AI ethics as systems ethics provides a set of ethical responses. A key challenge that we face now is to orchestrate existing ethical approaches in a useful manner for societal benefit.

AI ethics as an ethics that takes systems theory seriously will need to find ways to bring together the approaches and responses to ethical challenges that we have presented. The responses and mitigation strategies put forward here do not claim to be comprehensive. There are many others, including professional bodies, standardisation, certification and the use of AI incident databases, to name but a few. Many of these already exist, and some are being developed and tailored for their application to AI. *The significant challenge will be to orchestrate them in a way that is open, transparent and subject to debate and questioning, while at the same time oriented towards action and practical outcomes.* Regulation and legislation will likely play a key role here, for example the European Union's Artificial Intelligence Act proposal, but other regulatory interventions, such as the creation of AI regulators, may prove important (Stahl et al. 2022). However, it is not just the national and international policymakers that have to play a role here. Organisations, industry

associations, professional bodies, trade unions, universities, ethics committees, the media and civil society need to contribute. All these activities are based on the effort and contributions of individuals who are willing to participate in these efforts and prepared to reflect critically on their actions.

Tackling AI ethics challenges is no simple matter, and we should not expect to be able to solve all ethical issues. Instead, we should recognise that dealing with ethics is part of what humans do and that the use of technology can add complexity to traditional or well-known ethical questions. We should furthermore recognise that AI ethics often cannot be distinguished from the ethics of technology in general, or from ethical issues related to other digital and non-digital technologies. But at the same time, it has its peculiarities that need to be duly considered.

Our aim in this book has been to encourage reflection on some interesting cases involving AI ethics. We hope that the reader has gained insights into dealing with these issues, and understands that ethical issues of technology must be reflected upon and pursued with vigilance, as long as humans use technology.

## Reference

Stahl BC, Rodrigues R, Santiago N, Macnish K (2022) A European agency for artificial intelligence: protecting fundamental rights and ethical values. *Comput Law Secur Rev* 45:105661. <https://doi.org/10.1016/j.clsr.2022.105661>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



# Index

## A

- Access, 9, 14, 20, 28, 29, 34, 39, 41, 44–46, 55, 59, 64, 66, 67, 69, 73, 81, 85, 86, 88, 89, 98–100, 102, 109  
Accountability, 19, 43, 69, 70, 73  
Acxiom, 43  
Adversarial attack, 4, 63, 65, 67, 68, 73  
Adversarial example, 68, 73  
Adversarial robustness, 63, 73  
African Charter on Human and Peoples' Rights, 64  
Agenda 2030, 95, 98, 103  
Aggregation, 40, 72  
Agriculture  
precision agriculture, 97  
AI ethicist/s, 2  
AI for Good, 5, 95–97, 99–103, 109  
AI HLEG, 5, 18, 19, 32, 33, 57, 69, 70  
AI talent, 102, 103  
Algorithmic system, 13, 70  
Algorithms, 1, 2, 12, 14, 47, 54, 55, 70, 73, 87, 95  
American Convention on Human Rights, 64  
Antitrust, 39, 44, 45, 47  
Appropriation, 39–41, 43  
Aristotle, 5, 85  
Artificial intelligence, 1–6, 9–20, 25–34, 39, 42–44, 48, 53–59, 63–65, 67–74, 79–83, 86, 87, 90, 95–98, 100–103, 107–111  
Asymmetries, 39  
Autonomous vehicles, 65, 69, 108  
Autonomy, 43, 57, 59, 67, 71, 83–85, 87–89, 108

## B

- Behavioural, 39  
Behavioural data, 39  
Behavioural prediction, 39  
Big Tech, 39, 42, 44, 45, 47, 48  
Brazil, 5, 95, 97, 98, 109

## C

- Canadian Supreme Court, 79  
Care robots, 5, 79, 85–90  
Case study, case studies, 1–3, 5, 15, 20, 30, 42, 55, 57, 67–69, 72, 96, 101, 107, 108  
Catholic, 81  
Child, children, 4, 84, 88–90  
ChoicePoint, 43  
Civil society, 48, 59, 111  
Clearview AI, 40, 41  
Commercialisation, 39  
Commission Nationale de l'Informatique et des Libertés, CNIL, 17, 40, 41  
Compensation, 41, 72  
Competition, 41, 42, 45  
Consumer, 26, 39, 41, 44, 46, 53, 56, 58, 65, 71  
Convention on the Rights of Persons with Disabilities, 64  
Convention on the Rights of the Child, 64  
Corporate veil, 48  
COVID-19, 11, 40, 68  
Crime, 12, 13, 30, 63, 67, 80, 108, 109  
Criminal, 12, 15, 30, 63, 66, 67, 82, 90  
Crown Prosecution Service, 89

**D**

- Data Act, 46
- Data broker, 43
- Data desert, 100
- Data management, 72
- Data ownership, 39, 46
- Data poisoning, 73
- Data sharing, 39, 44–46
- Deception, 40, 87
- Deep learning, 1, 2, 58, 68
- Democracy, 39, 43, 45, 47, 54–56
- Developers, 2, 10, 18, 63, 73
- Diagnosis, 1, 4, 28, 63, 67, 68
- Digital divide, 1, 100, 102, 103, 108
- Digital Markets Act, 46
- Dignity
  - aspirational dignity, 80, 82
  - dignity as slogan, 80, 90
  - intrinsic dignity, 80, 83, 84
- Disabilities, 9, 14–16, 90
- Disclosure, 44
- Discrimination, 3, 9, 10, 13–20, 84, 87, 109
- Domination, 40
- Do no harm, 43, 96
- Double standards, 101

**E**

- Ebola, 5, 98, 99
- Elderly, 5, 85–90, 107, 110
- European Commission, 17, 18, 42, 46, 64, 69–72, 103
- European Convention on Human Rights, 64
- European Union's High Level Expert Group on AI, AI HLEG, 5, 18, 33, 57, 69
- Experian, 43
- Explainability, 39, 43, 58, 59, 108–110
- Exploitation, 34, 41, 43, 69, 103

**F**

- Facebook, 4, 40, 42, 54, 56, 57
- Famine, 101
- Fitbit, 41, 42
- Fitness, 41
- France, 40
- Freedom of association, 67
- Friend, 44, 54

**G**

- General Data Protection Regulation (GDPR), 26, 31–33, 41, 46
- Genetic manipulation (GM), 99, 100

Golden Rice, 100, 101

Google, 40–42, 55

Guinea, 5, 98

**H**

- Harm, 25, 30, 33, 34, 41, 66, 68–72, 74, 79, 83–85, 89, 90, 96, 98, 99, 103, 109
- Healthcare, 27, 28, 33, 40, 63, 65, 68, 87
- Health data, 4, 39, 41, 43
- Helicopter research, 95, 98–103, 109
- High-risk, 12, 13, 64, 70, 72, 73, 84
- HIV, 101
- Home intrusions, 67
- Human rights, 1, 3, 10, 11, 17, 28, 32, 34, 43, 56, 63, 64, 66, 69, 71, 73, 79, 83, 84, 90

**I**

- Identity theft, 67
- Illegal, 2, 3, 9, 14–20, 40, 41, 45, 47, 48, 58
- Impact assessment, 3, 9, 17–20, 25, 32, 33, 59, 110
- Indignation, 44
- Individual, 9, 10, 12, 13, 16, 18, 25, 26, 29, 30, 33, 34, 39, 41, 43, 44, 46, 48, 53, 54, 56–59, 63, 69, 70, 80, 83, 96, 97, 109–111
- Inequality, 39, 42, 55, 90, 95, 98, 103
- Infantilisation, 87
- Information capitalism, 39
- Innovation, 2, 17, 19, 26, 45, 47, 64, 96–98
- Inter-American Convention on Protecting the Human Rights of Older Persons, 64
- Interference, 2, 43, 84
- International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families, 64
- International Covenant on Civil and Political Rights, 64
- Intrinsic worth, 81
- Intrusion, 67, 100

**K**

- Kant, Immanuel, 80, 81, 83, 96, 107, 109

**L**

- Law enforcement, 9, 12, 13, 15, 17, 29, 40, 64, 65, 108
- Leak, 4, 29, 41
- Legal basis, 33, 41

Legal framework, 47  
Liability, 63, 69–72, 88, 110  
Liberia, 5, 98  
Liberty, 4, 12, 57, 63–65, 67, 69, 71, 74, 84  
Life, 3, 12, 16, 17, 43, 44, 47, 48, 63–65,  
    67–69, 71, 73, 74, 82, 84, 85, 88, 89,  
    96  
Literacy, 44  
Location tracking, 67

## M

Machine learning, 1, 2, 10, 11, 16, 17, 25,  
    26, 30, 33, 58, 96, 97, 100, 108  
Mandatory, 44  
Mandela, Nelson, 80  
Man-in-the-middle attacks, 4, 67, 107  
Manipulation, 4, 34, 43, 53–59, 68, 79, 89,  
    90  
Market control, 39  
Marketplace, 40  
Medical, 1, 4, 14, 28, 29, 63, 64, 67, 68, 89,  
    109  
Merger, 40, 42, 45  
Migrant workers, 64  
Modification, 30, 39, 67  
Monetisation, 39, 41, 43  
Movement, 5, 27, 67, 83, 84, 98–100

## N

Naming, 44  
New York, 12, 40, 41, 48

## P

Pandemic, 11, 40, 109  
Personal data, 20, 25–27, 30, 31, 33, 34,  
    39–41, 43, 44, 58, 88, 109  
Peru, 5, 95, 97, 98, 109  
Power, 27, 39, 41–45, 47, 56, 58, 59, 84,  
    88, 90, 109  
Privacy, 3, 4, 17–20, 25–28, 30–32, 34, 39,  
    43, 45, 53, 56–59, 67, 69, 73, 84, 87,  
    88, 98, 101, 103, 109  
Private sector, 40  
Profit, 39, 41, 44, 56

## Q

Quality management, 63, 72

## R

Regulation, 1, 25–27, 31, 39, 42–45, 47, 66,  
    82, 90, 108, 110  
Resource-limited settings, 5, 95, 97,  
    99–101, 103, 109, 110  
Responsibility, 33, 46, 66, 67, 69, 70, 72,  
    73, 87  
Retail, 40, 65  
Revenue, 39  
Rights, 17, 25, 26, 32, 41, 46, 48, 56, 63–65,  
    67, 69, 73, 79, 81, 83, 84, 87–89, 98  
Right to life, liberty and security of person,  
    4, 63, 65  
Risk, 3, 12–14, 18, 28, 32, 33, 41, 43, 45,  
    63, 64, 66, 68–73, 85, 88, 90, 97, 99,  
    107  
Risk assessment tools, 63  
Robot  
    assistive robot, 85  
    companion robot, 85, 86  
    monitoring robot, 85, 86  
    pet robot, 85, 110  
R. v. Kapp, 80

## S

Safeguards, 25, 45, 68, 74  
Safety, 1, 48, 64–68, 71, 74, 88, 99, 107  
Sanctions, 30, 41, 42  
Sartre, Jean-Paul, 86  
Seasonal climate forecasting, 5, 95, 97, 98,  
    101, 103, 109  
Security, 4, 18, 19, 28, 30, 41, 63–69, 71,  
    73, 74, 81, 101, 107  
Self-driving cars, 4, 63, 65, 66, 71, 107  
Sen, Amartya, 101  
Sex robots, 3, 4, 79, 83–87, 89, 90, 108  
Sexual rights, 83, 84, 90, 108  
Sexual violence, 83  
Sierra Leone, 5, 98  
Slavery, 79, 81, 89, 90  
Smart home, 4, 65–67, 69  
Social media, 27, 28, 34, 40, 42–44, 54, 56  
Socio-economic, 39, 44, 59, 90  
Surveillance capitalism, 4, 34, 39, 40,  
    42–48, 59, 100  
Sustainable Development Goals (SDGs), 5,  
    95, 96, 98–103

## T

Telehealth, 40  
Tesla, 4, 39, 65, 66

Threat, 26, 30–32, 41, 43, 56, 67, 73, 88, 109  
Touch, 34, 67, 83  
Transparency, 19, 39, 43, 45, 48, 53, 58, 59, 82  
Transport, 63–65

**U**

UN Educational, Scientific and Cultural Organization (UNESCO), 2, 17, 43, 68  
Unethical, 44, 57  
Unfair discrimination, 2, 3, 9, 14–20, 34  
Unfair dismissal, 81, 82, 87, 108  
Universal Declaration of Human Rights, 10, 14, 64, 79, 81, 83  
Unlawful, 41, 45, 69, 82

Usability, 67  
Users, 2, 4, 5, 15, 18, 40–44, 46, 55–57, 63, 66, 67, 70, 73, 87, 89, 98, 107

**V**

Virtue, 80, 81, 96  
Vulnerability, 4, 41, 54, 55, 57, 65–69, 73

**W**

Whistleblower, 48  
World Health Organization, 83, 84, 89

**Z**

Zimbabwe, 5, 95, 97, 98, 109