4. Public policy considerations

This chapter explores public policy considerations to ensure that artificial intelligence (AI) systems are trustworthy and human-centred. It covers concerns related to ethics and fairness; the respect of human democratic values, including privacy; and the dangers of transferring existing biases from the analogue world into the digital world, including those related to gender and race. The need to progress towards more robust, safe, secure and transparent AI systems with clear accountability mechanisms for their outcomes is underlined.

Policies that promote trustworthy AI systems include those that encourage investment in responsible AI research and development; enable a digital ecosystem where privacy is not compromised by a broader access to data; enable small and medium-sized enterprises to thrive; support competition, while safeguarding intellectual property; and facilitate transitions as jobs evolve and workers move from one job to the next.

Human-centred AI

Artificial intelligence (AI) plays an increasingly influential role. As the technology diffuses, the potential impacts of its predictions, recommendations or decisions on people's lives increase as well. The technical, business and policy communities are actively exploring how best to make AI human-centred and trustworthy, maximise benefits, minimise risks and promote social acceptance.

Box 4.1. "Black box" AI systems present new challenges from previous technological advancements

Neural networks have often been referred to as a "black box". Although the behaviour of such systems can indeed be monitored, the term "black box" reflects the considerable difference between the ability to monitor previous technologies compared to neural networks. Neural networks iterate on the data they are trained on. They find complex, multi-variable probabilistic correlations that become part of the model that they build. However, they do not indicate how data could interrelate (Weinberger, 2018[1]). The data are far too complex for the human mind to understand. Characteristics of AI that differ from previous technological advancements and affect transparency and accountability include:

- **Discoverability**: Rules-based algorithms can be read and audited rule-by-rule, making it comparatively straightforward to find certain types of errors. By contrast, certain types of machine-learning (ML) systems, notably neural networks, are simply abstract mathematical relationships between factors. These can be extremely complex and difficult to understand, even for those who program and train them (OECD, 2016).
- **Evolving nature**: Some ML systems iterate and evolve over time and may even change their own behaviour in unforeseen ways.
- **Not easily repeatable**: A specific prediction or decision may only appear when the ML system is presented with specific conditions and data, which are not necessarily repeatable.
- **Increased tensions in protecting personal and sensitive data**:
  - **Inferences**: Even in the absence of protected or sensitive data, AI systems may be able to infer these data and correlations from proxy variables that are not personal or sensitive, such as purchasing history or location (Kosinski, Stillwell and Graepel, 2013[2]).
  - **Improper proxy variables**: Policy and technical approaches to privacy and non-discrimination have tended to minimise data collected, prohibit use of certain data or remove data to prevent their use. But an AI system might base a prediction on proxy data that bear a close relationship to the forbidden and non-collected data. Furthermore, the only way to detect these proxies is to also collect sensitive or personal data such as race. If such data are collected, then it becomes important to ensure they are only used in appropriate ways.
  - **The data-privacy paradox**: For many AI systems, more training data can improve the accuracy of AI predictions and help reduce risk of bias from skewed samples. However, the more data collected, the greater the privacy risks to those whose data are collected.

Some types of AI – often referred to as "black boxes" – raise new challenges compared to previous technological advancements (Box 4.1). Given these challenges, the OECD – building on the work of the AI Group of Experts at the OECD (AIGO) – identified key priorities for human-centred AI. First, it should contribute to inclusive and sustainable growth and well-being. Second, it should respect human-centred values and fairness. Third, the use of AI and how its systems operate should be transparent. Fourth, AI systems should be robust and safe. Fifth, there should be accountability for the results of AI predictions and the ensuing decisions. Such measures are viewed as critical for high-stakes' predictions. They are also important in business recommendations or for less impactful uses of AI.

Inclusive and sustainable growth and well-being

AI holds significant potential to advance the agenda towards meeting the Sustainable Development Goals

AI can be leveraged for social good and to advance meeting the United Nations Sustainable Development Goals (SDGs) in areas such as education, health, transport, agriculture and sustainable cities, among others. Many public and private organisations, including the World Bank, a number of United Nations agencies and the OECD, are working to leverage AI to help advance the SDGs.

Ensuring that AI development is equitable and inclusive is a growing priority

Ensuring that AI development is equitable and inclusive is a growing priority. This is especially true in light of concerns about AI exacerbating inequality or increasing existing divides within and between developed and developing countries. These divides exist due to concentration of AI resources – AI technology, skills, datasets and computing power – in a few companies and nations. In addition, there is concern that AI could perpetuate biases (Talbot et al., 2017[3]). Some fear AI could have a disparate impact on vulnerable and under-represented populations. These include the less educated, low skilled, women and elderly, particularly in low- and middle-income countries (Smith and Neupane, 2018[4]). Canada's International Development Research Centre recently recommended the formation of a global AI for Development fund. This would establish AI Centres of Excellence in low- and middle-income countries to support the design and implementation of evidence-based inclusive policy (Smith and Neupane, 2018[4]). Its goal is to ensure that AI benefits are well distributed and lead to more egalitarian societies. Inclusive AI initiatives aim to ensure that economic gains from AI in societies are widely shared and that no one is left behind.

Inclusive and sustainable AI is an area of focus for countries such as India (NITI, 2018[5]); companies such as Microsoft;[1] and academic groups such as the Berkman Klein Center at Harvard. For example, Microsoft has launched projects such as the Seeing AI mobile application, which helps the visually impaired. It scans and recognises all the elements surrounding a person and provides an oral description. Microsoft is also investing USD 2 million in qualified initiatives to leverage AI to tackle sustainability challenges such as biodiversity and climate change (Heiner and Nguyen, 2018[6]).

Human-centred values and fairness

Human rights and ethical codes

International human rights law embodies ethical norms

International human rights law embodies ethical norms. AI can support the fulfilment of human rights, as well as create new risks that human rights might be deliberately or accidently violated. Human rights law, together with the legal and other institutional structures related to it, could also serve as one of the tools to help ensure human-centred AI (Box 4.2).

Box 4.2. Human rights and AI

International human rights refer to a body of international laws, including the International Bill of Rights,[1] as well as regional human rights systems developed over the past 70 years around the world. Human rights provide a set of universal minimum standards based on, among others, values of human dignity, autonomy and equality, in line with the rule of law. These standards and the legal mechanisms linked to them create legally enforceable obligations for countries to respect, protect and fulfil human rights. They also require that those whose rights have been denied or violated be able to obtain effective remedy.

Specific human rights include the right to equality, the right to non-discrimination, the right to freedom of association, the right to privacy and economic, social and cultural rights such as the right to education or the right to health.

Recent intergovernmental instruments such as the United Nations *Guiding Principles on Business and Human Rights* (OHCHR, 2011[7]) have also addressed private actors in the context of human rights. They provide for a "responsibility" on private actors to respect human rights. In addition, the 2011 update of government-backed recommendations to business in the *OECD Guidelines for Multinational Enterprises* (OECD, 2011[8]) contains a chapter on human rights.

Human rights overlap with wider ethical concerns and with other areas of regulation relevant to AI, such as personal data protection or product safety law. However, these other concerns and issues often have different scope.

1. Comprised of the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights.

AI promises to advance human rights

Given the potential breadth of its application and use, AI promises to advance the protection and fulfilment of human rights. Examples include using AI in the analysis of patterns in food scarcity to combat hunger, improving medical diagnosis and treatment or making health services more widely available and accessible, and shedding light on discrimination.

AI could also challenge human rights

AI may also pose a number of human rights challenges that are often reflected in discussions on AI and ethics more broadly. Specific AI systems could violate, or be used to violate, human rights accidentally or deliberately. Much focus is placed on accidental impacts. ML algorithms that predict recidivism, for example, may have undetected bias. Yet AI technologies can also be linked to intentional violations of human rights. Examples

include the use of AI technologies to find political dissidents, and restricting individuals' rights to freedom of expression or to participate in political life. In these cases, the violation in itself is usually not unique to the use of AI. However, it could be exacerbated by AI's sophistication and efficiency.

The use of AI may also pose unique challenges in situations where human rights impacts are unintentional or difficult to detect. The reason can be the use of poor-quality training data, system design or complex interactions between the AI system and its environment. One example is algorithmic exacerbation of hate speech or incitement to violence on line. Another example is the unintentional amplifying of fake news, which could impact the right to take part in political and public affairs. The likely scale and impact of harm will be linked to the scale and potential impact of decisions by any specific AI system. For example, a decision by a news recommendation AI system has a narrower potential impact than a decision by an algorithm predicting the risk of recidivism of parole inmates.

Human rights frameworks complemented by AI ethical codes

Ethical codes can address the risk that AI might not operate in a human-centred manner or align with human values. Both private companies and governments have adopted a large number of ethical codes relating to AI.

For example, Google-owned DeepMind also created a DeepMind Ethics & Society unit in October 2017.[2] The unit's goal is to help technologists understand the ethical implications of their work and help society decide how AI can be beneficial. The unit will also fund external research on algorithmic bias, the future of work, lethal autonomous weapons and more. Google itself announced a set of ethical principles to guide its research, product development and business decisions.[3] It published a white paper on AI governance, identifying issues for further clarification with governments and civil societies.[4] Microsoft's AI vision is to "amplify human ingenuity with intelligent technology" (Heiner and Nguyen, 2018[6]). The company has launched projects to ensure inclusive and sustainable development.

Human rights law, together with its institutional mechanisms and wider architecture, provides the direction and basis to ensure the ethical and human-centred development and use of AI in society.

Leveraging human rights frameworks in the AI context presents advantages

The advantages of leveraging human rights frameworks in the AI context include established institutions, jurisprudence, universal language and international acceptance:

- **Established institutions**: A wide international, regional and national human rights infrastructure has been developed over time. It is comprised of intergovernmental organisations, courts, non-governmental organisations, academia, and other institutions and communities where human rights can be invoked and remedy sought.
- **Jurisprudence**: As legal norms, the values protected by human rights are operationalised and made concrete, and legally binding, in specific situations through jurisprudence and the interpretative work by international, regional and national institutions.
- **Universal language**: Human rights provide a universal language for a global issue. This, together with the human rights infrastructure, can help enfranchise a wider variety of stakeholders. They can thus participate in the debate on the place of AI in society, alongside the AI actors directly involved in the AI lifecycle.

- **International acceptance**: Human rights have broad international acceptance and legitimacy. The mere perception that an actor may violate human rights can be significant, since the associated reputational costs can be high.

A human rights approach to AI can help identify risks, priorities, vulnerable groups and provide remedy

- **Risk identification**: Human rights frameworks can help identify risks of harm. In particular, they can carry out human rights due diligence such as human rights impact assessments (HRIAs) (Box 4.3).
- **Core requirements**: As minimum standards, human rights define inviolable core requirements. For example, in the regulation of expression on social networks, human rights jurisprudence helps demarcate hate speech as a red line.
- **Identifying high-risk contexts**: Human rights can be a useful tool to identify high-risk contexts or activities. In such situations, increased care is needed or AI could be deemed unfit for use.
- **Identifying vulnerable groups or communities**: Human rights can help identify vulnerable or at-risk groups or communities in relation to AI. Some individuals or communities may be under-represented due, for example, to limited smartphone use.
- **Remedy**: As legal norms with attendant obligations, human rights can provide remedy to those whose rights are violated. Examples of remedies include cessation of activity, development of new processes or policies, an apology or monetary compensation.

Box 4.3. Human rights impact assessments

HRIAs can help determine risks that AI lifecycle actors might not otherwise envisage. To that end, they focus on incidental human impacts rather than optimisation of the technology or its outputs. HRIAs or similar processes could ensure by-design respect for human rights throughout the lifecycle of the technology.

HRIAs assess technology against a wide range of possible human rights impacts, a broad-sweeping approach that is resource-intensive. It can be easier to start with the AI system in question and work outwards. In this way, AI focuses on a limited range of areas where rights challenges appear most likely. Industry organisations can help conduct HRIAs for small and medium-sized enterprises (SMEs) or non-tech companies that deploy AI systems but may not be literate in the technology. The Global Network Initiative exemplifies such an organisation with respect to freedom of expression and privacy. It helps companies plan ahead and incorporate human rights assessments into their plans for new products (https://globalnetworkinitiative.org/).

HRIAs have the drawback of generally being conducted company by company. Conversely, AI systems might involve many actors, which means looking at only one part may be ineffective. Microsoft was the first large technology company to conduct an HRIA on AI in 2018.

There are also significant challenges to implement a human rights approach to AI. These are related to how human rights are directed towards countries, how enforcement is tied to jurisdictions, how they are better suited to remediate substantial harms to a small number of individuals and how they can be costly to business:

- **Human rights are directed towards countries, not private actors**. Yet private-sector actors play a key role in AI research, development and deployment. This challenge is not unique to AI. Several intergovernmental initiatives seek to overcome the public/private divide. Beyond such efforts, there is growing recognition that a good human rights record is good for business.[5]
- **Enforcement of human rights is tied to jurisdictions**. Generally, claimants must demonstrate legal standing in a specific jurisdiction. These approaches may not be optimal when cases involve large multinational enterprises and AI systems that span multiple jurisdictions.
- **Human rights are better suited to remediate substantial harms to a small number of individuals**, as opposed to less significant harms suffered by many. In addition, human rights and their structures can seem opaque to outsiders.
- **In some contexts, human rights have a reputation as being costly to business**. Therefore, approaches that put forward ethics, consumer protection or responsible business conduct, as well as the business case for respecting human rights, seem promising.

Some general challenges of AI such as transparency and explainability also apply with respect to human rights (Section "Transparency and explainability"). Without transparency, identifying when human rights have been violated or substantiating a claim of violation is difficult. The same is true for seeking remedy, determining causality and accountability.

Personal data protection

AI challenges notions of "personal data" and consent

AI can increasingly link different datasets and match different types of information with profound consequences. Data held separately were once considered non-personal (or were stripped of personal identifiers, i.e. "de-identified"). With AI, however, non-personal data can be correlated with other data and matched to specific individuals, becoming personal (or "re-identified"). Thus, algorithmic correlation weakens the distinction between personal data and other data. Non-personal data can increasingly be used to re-identify individuals or infer sensitive information about them, beyond what was originally and knowingly disclosed (Cellarius, 2017[9]). In 2007, for example, researchers had already used reportedly anonymous data to link Netflix's list of movie rentals with reviews posted on IMDB. In this way, they identified individual renters and accessed their complete rental history. With more data collected, and technological improvements, such links are increasingly possible. It becomes difficult to assess which data can be considered and will remain non-personal.

It is increasingly difficult to distinguish between sensitive and non-sensitive data in, for example, the European Union's General Data Protection Regulation (GDPR). Some algorithms can infer sensitive information from "non-sensitive" data, such as assessing individuals' emotional state based on their keyboard typing pattern (Privacy International and Article 19, 2018[10]). The use of AI to identify or re-identify data that originally were non-personal or de-identified also presents a legal issue. Protection frameworks, like the OECD Recommendation of the Council concerning *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* ("Privacy Guidelines"), apply to personal data (Box 4.4). Therefore, it is not clear if, or at what point, they apply to data that under some circumstances would be, or could be, identifiable (Office of the Victorian Information

Commissioner, 2018[11]). An extreme interpretation could result in vastly broadening the scope of privacy protection, which would make it difficult to apply.

Box 4.4. The OECD Privacy Guidelines

The Recommendation of the Council concerning *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* ("Privacy Guidelines") was adopted in 1980 and updated in 2013 (OECD, 2013[12]). It contains definitions of relevant terms, notably defining "personal data" as "any information relating to an identified or identifiable individual (data subject)". It also defines principles to apply when processing personal data. These principles relate to collection limitation (including, where appropriate, consent as means to ensure this principle), data quality, purpose specification, use limitation, security safeguards, openness, individual participation and accountability. They also provide that in implementing the Privacy Guidelines, members should ensure there is no unfair discrimination against data subjects. The implementation of the Privacy Guidelines was to be reviewed in 2019 to take account, among others, of recent developments, including in the area of AI.

AI also challenges personal data protection principles of collection limitation, use limitation and purpose specification

To train and optimise AI systems, ML algorithms require vast quantities of data. This creates an incentive to maximise, rather than minimise, data collection. With the growth in use of AI devices, and the Internet of Things (IoT), more data are gathered, more frequently and more easily. They are linked to other data, sometimes with little or no awareness or consent on the part of the data subjects concerned.

The patterns identified and evolution of the "learning" are difficult to anticipate. Therefore, the collection and use of data can extend beyond what was originally known, disclosed and consented to by a data subject (Privacy International and Article 19, 2018[10]). This is potentially incompatible with the Privacy Guidelines' principles of collection limitation, use limitation and purpose specification (Cellarius, 2017[9]). These first two principles rely in part on the data subject's consent (as appropriate, recognising that consent may not be feasible in some cases). This consent is either the basis for the collection of personal data, or for its use for other purposes than originally specified. AI technologies such as deep learning that are difficult to understand or monitor are also difficult to explain to the data subjects concerned. This is a challenge for companies. They report the exponential rate at which AI gains access to, analyses and uses data is difficult to reconcile with these data protection principles (OECD, 2018[13]).

The combination of AI technologies with developments in the IoT, i.e. the connection of an increasing number of devices and objects over time to the Internet, exacerbates the challenges. The increasing combination of AI and IoT technologies (e.g. IoT devices equipped with AI, or AI algorithms used to analyse IoT data) means that more data, including personal data, are constantly gathered. These can be increasingly linked and analysed. On the one hand, there is an increased presence of devices collecting information (e.g. surveillance cameras or autonomous vehicles [AVs]). On the other, there is better AI technology (e.g. facial recognition). The combination of these two trends risks leading to more invasive outcomes than either factor separately (Office of the Victorian Information Commissioner, 2018[11]).

AI can also empower individual participation and consent

AI carries potential to enhance personal data. For example, initiatives to build AI systems around principles of privacy by design and privacy by default are ongoing within a number of technical standards organisations. For the most part, they use and adapt privacy guidelines, including the OECD Privacy Guidelines. Additionally, AI is used to offer individuals tailored personalised services based on their personal privacy preferences, as learned over time (Office of the Victorian Information Commissioner, 2018[11]). These services can help individuals navigate between the different personal data processing policies of different services and ensure their preferences are considered across the board. In so doing, AI empowers meaningful consent and individual participation. A team of researchers, for example, developed Polisis, an automated framework that uses neural network classifiers to analyse privacy policies (Harkous, 2018[14]).

Fairness and ethics

ML algorithms can reflect the biases implicit in their training data

To date, AI policy initiatives feature ethics, fairness and/or justice prominently. There is significant concern that ML algorithms tend to reflect and repeat the biases implicit in their training data, such as racial biases and stereotyped associations. Because technological artefacts often embody societal values, discussions of fairness should articulate which societies technologies should serve, who should be protected and with what core values (Flanagan, Howe and Nissenbaum, 2008[15]). Disciplines such as philosophy, law and economy have grappled with different notions of fairness for decades from several angles. They illustrate the broad range of possible visions of fairness and the implications for policy.

Philosophical, legal and computational notions of fairness and ethical AI vary

Philosophy focuses on concepts of right and wrong conduct, good and evil, and morality. Three major philosophical theories are relevant in the context of ethical AI (Abrams et al., 2017[16]):

- **The fundamental human rights approach**, associated with Immanuel Kant, identifies the formal principles of ethics, which are specific rights such as privacy or freedom. It protects these principles by regulation, which AI systems should respect.
- **The utilitarian approach**, pursued by Jeremy Bentham and John Stuart Mill, focuses on public policies that maximise human welfare based on economic cost benefit analyses. For AI, the utilitarian approach raises the question of *whose* welfare to maximise (e.g. individuals, family, society or institutions/governments), which may impact algorithm design.
- **The virtue ethics approach**, based on Aristotle's work, focuses on the values and ethical norms needed for a society to support people in their everyday efforts to live a life worth living. This raises the question of which values and which ethical norms warrant protection.

The law often uses the terms "equality" and "justice" to represent concepts of fairness. The two major legal approaches to fairness are individual fairness and group fairness.

- **Individual fairness** represents the notion of equality before the law. It implies that everyone should be treated equally and not discriminated against in view of special attributes. Equality is recognised as an international human right.

- **Group fairness** focuses on the fairness of the outcome. It ensure the outcome does not differ in any systematic manner for people who, based on a protected characteristic (e.g. race or gender), belong to different groups. It reasons that differences and historical circumstances may lead different groups to react to situations differently. Different countries' approaches to group fairness differ significantly. Some, for example, use positive discrimination.

AI system designers have been considering how to represent fairness in AI systems. Different definitions of fairness embody different possible approaches (Narayanan, 2018[17]):

- The "**unaware approach**", whereby an AI system should be unaware of any identifiable factors, accompanies the individual fairness legal approach. In this case, the AI system does not consider data on sensitive or prohibited attributes, such as gender, race and sexual orientation (Yona, 2017[18]). However, many other factors may be correlated with the protected/prohibited attribute (such as gender). Removing them could limit the accuracy of an AI system.
- **Fairness through awareness** acknowledges group differences and aims to treat similar individuals in the same way. The challenge is, however, to determine who should be treated similarly to whom. Understanding who should be considered similar for a particular task requires knowledge of sensitive attributes.
- **Group fairness approaches** focus on ensuring that outcomes do not differ systematically for people who belong to different groups. There is concern about the potential for AI systems to be unfair, perpetuating or reinforcing traditional biases, since they often rely on data sets representing past activity.

Different notions of fairness translate into different results for different groups in society and different types of stakeholders. They cannot all be simultaneously achieved. Policy considerations and in some cases, choices, should inform technological design choices that could adversely impact specific groups.

AI in human resources illustrates the opportunity and challenge of AI for bias

In human resources, use of AI either perpetuates bias in hiring or helps uncover and reduce harmful bias. A Carnegie Mellon study exploring patterns of online job postings showed that an ad for higher-paid executives was displayed 1 816 times to men and just 311 times to women (Simonite, 2018[19]). Thus, one potential area for human-AI collaboration is to ensure that AI applications for hiring and evaluation are transparent. They should not codify biases, e.g. by automatically disqualifying diverse candidates for roles in historically non-diverse settings (OECD, 2017[20]).

Several approaches can help mitigate discrimination in AI systems

Approaches proposed to mitigate discrimination in AI systems include awareness building; organisational diversity policies and practices; standards; technical solutions to detect and correct algorithmic bias; and self-regulatory or regulatory approaches. For example, in predictive policing systems, some propose algorithmic impact assessments or statements. These would require police departments to evaluate the efficacy, benefits and potential discriminatory effects of all available choices for predictive policing technologies (Selbst, 2017[21]). Accountability and transparency are important to achieve fairness. However, even combined they do not guarantee it (Weinberger, 2018[22]); (Narayanan, 2018[17]).

Striving for fairness in AI systems may call for trade-offs

AI systems are expected to be "fair". This aims to result in, for example, only the riskiest defendants remaining in jail or the most suitable lending plan being proposed based on ability to pay. **False positive errors** indicate a misclassification of a person or a behaviour. For example, they could wrongly predict a defendant will reoffend when he or she will not. They could also wrongly predict a disease that is not there. **False negative errors** represent cases in which an AI system wrongly predicts, for example, that a defendant will not reoffend. As another example, a test may wrongly indicate the absence of a disease.

Group fairness approaches acknowledge different starting points for different groups. They try to account for differences mathematically by ensuring "equal accuracy" or equal error rates across all groups. For example, they would wrongly classify the same percentage of men and women as reoffenders (or equalise the false positives and false negatives).

Equalising false positives and false negatives creates a challenge. False negatives are often viewed as more undesirable and risky than false positives because they are more costly (Berk and Hyatt, 2015[23]). For example, the cost to a bank of lending to someone an AI system predicted would not default – but who does default – is greater than the gain from that loan. Someone diagnosed as not having a disease who does have that disease may suffer significantly. Equalising true positives and true negatives can also lead to undesirable outcomes. They could, for example, incarcerate women who pose no safety risk so that the same proportion of men and women are released (Berk and Hyatt, 2015[23]). Some approaches aim, for example, to equalise both false positives and false negatives at the same time. However, it is difficult to simultaneously satisfy different notions of fairness (Chouldechova, 2016[24]).

Policy makers could consider the appropriate treatment of sensitive data in the AI context

The appropriate treatment of sensitive data could be reconsidered. In some cases, organisations may need to maintain and use sensitive data to ensure their algorithms do not inadvertently reconstruct this data. Another priority for policy is to monitor unintended feedback loops. When police go to algorithmically identified "high crime" areas, for example, this could lead to distorted data collection. It would further bias the algorithm – and society – against these neighbourhoods (O'Neil, 2016[25]).

Transparency and explainability

Transparency about the use of AI and as to how AI systems operate is key

The technical and the policy meanings of the term "transparency" differ. For policy makers, transparency traditionally focuses on how a decision is made, who participates in the process and the factors used to make the decision (Kosack and Fung, 2014[26]). From this perspective, transparency measures might disclose how AI is being used in a prediction, recommendation or decision. They might also disclose when a user is interacting with an AI-powered agent.

For technologists, transparency of an AI system focuses largely on process issues. It means allowing people to understand how an AI system is developed, trained and deployed. It may also include insight into factors that impact a specific prediction or decision. It does not usually include sharing specific code or datasets. In many cases, the systems are too complex for these elements to provide meaningful transparency (Wachter, Mittelstadt and Russell, 2017[27]). Moreover, sharing specific code or datasets could reveal trade secrets or disclose sensitive user data.

More generally, awareness and understanding of AI reasoning processes is viewed as important for AI to become commonly accepted and useful.

Approaches to transparency in AI systems

Experts at Harvard University in the Berkman Klein Center Working Group on Explanation and the Law identify approaches to improve transparency of AI systems, and note that each entails trade-offs (Doshi-Velez et al., 2017[28]). An additional approach is that of optimisation transparency, i.e. transparency about the goals of an AI system and about their results. These approaches are: i) theoretical guarantees; ii) empirical evidence; and iii) explanation (Table 4.1).

Table 4.1. Approaches to improve the transparency and accountability of AI systems

Approach	Description	Well-suited contexts	Poorly suited contexts
Theoretical guarantees	In some situations, it is possible to give theoretical guarantees about an AI system backed by proof.	The environment is fully observable (e.g. the game of Go) and both the problem and solution can be formalised.	The situation cannot be clearly specified (most real-world settings).
Statistical evidence/ probability	Empirical evidence measures a system's overall performance, demonstrating the value or harm of the system, without explaining specific decisions.	Outcomes can be fully formalised; it is acceptable to wait to see negative outcomes to measure them; issues may only be visible in aggregate.	The objective cannot be fully formalised; blame or innocence can be assigned for a particular decision.
Explanation	Humans can interpret information about the logic by which a system took a particular set of inputs and reached a particular conclusion.	Problems are incompletely specified, objectives are not clear and inputs could be erroneous.	Other forms of accountability are possible.

Source: adapted from Doshi-Velez et al. (2017[28]), "Accountability of AI under the law: The role of explanation", https://arxiv.org/pdf/1711.01134.pdf.

Some systems offer theoretical guarantees of their operating constraints

In some cases, **theoretical guarantees** can be provided, meaning the system will demonstrably operate within narrow constraints. Theoretical guarantees apply to situations in which the environment is fully observable and both the problem and the solution can be fully formalised, such as in the game of Go. In such cases, certain kinds of outcomes cannot happen, even if an AI system processes new kinds of data. For example, a system could be designed to provably follow agreed-upon processes for voting and vote counting. In this case, explanation or evidence may not be required: the system does not need to explain how it reached an outcome because the types of outcomes that cause concern are mathematically impossible. An assessment can be made at an early stage of whether these constrains are sufficient.

Statistical evidence of overall performance can be provided in some cases

In some cases, relying on **statistical evidence** of a system's overall performance may be sufficient. Evidence that an AI system measurably increases a specific societal or individual value or harm may be sufficient to ensure accountability. For example, an autonomous aircraft landing system may have fewer safety incidents than human pilots, or a clinical diagnostic support tool may reduce mortality. Statistical evidence might be an appropriate accountability mechanism in many AI systems. This is because it both protects trade secrets and can identify widespread but low-risk harms that only become apparent in aggregate (Barocas and Selbst, 2016[29]; Crawford, 2016[30]). Questions of bias or discrimination can be ascertained statistically: for example, a loan approval system might demonstrate its bias

by approving more loans for men than women when other factors are controlled for. The permissible error rate and uncertainty tolerated varies depending on the application. For example, the error rate permissible for a translation tool may not be acceptable for autonomous driving or medical examinations.

Optimisation transparency is transparency about a system's goals and results

Another approach to the transparency of AI systems proposes to shift the governance focus from a system's means to its ends: that is, moving from requiring the explainability of a system's inner workings to measuring its outcomes – i.e. what the system is "optimised" to do. This would require a declaration of what an AI system is optimised for, with the understanding that optimisations are imperfect, entail trade-offs and should be constrained by "critical constraints", such as safety and fairness. This approach advocates for using AI systems for what they are optimised to do. It invokes existing ethical and legal frameworks, as well as social discussions and political processes where necessary to provide input on what AI systems should be optimised for (Weinberger, 2018[1]).

Explanation relates to a specific outcome from an AI system

Explanation is essential for situations in which fault needs to be determined in a specific instance – a situation that may grow more frequent as AI systems are deployed to make recommendations or decisions currently subject to human discretion (Burgess, 2016[31]). The GDPR mandates that data subjects receive meaningful information about the logic involved, the significance and the envisaged consequences of automated decision-making systems. An explanation does not generally need to provide the full decision-making process of the system. Answering one of the following questions is generally enough (Doshi-Velez et al., 2017[28]):

1. **Main factors in a decision**: For many kinds of decisions, such as custody hearings, qualifying for a loan and pre-trial release, a variety of factors must be considered (or are expressly forbidden from being considered). A list of the factors that were important for an AI prediction – ideally ordered by significance – can help ensure that the right factors were included.
2. **Determinant factors, i.e. factors that decisively affect the outcome**: Sometimes, it is important to know whether a particular factor directed the outcome. Changing a particular factor, such as race in university admissions, can show whether the factor was used correctly.
3. **Why did two similar-looking cases result in different outcomes, or vice versa**? The consistency and integrity of AI-based predictions can be assessed. For example, income should be considered when deciding whether to grant a loan, but it should not be both dispositive and irrelevant in otherwise similar cases.

Explanation is an active area of research but entails costs and possible trade-offs

Technical research is underway by individual companies, standards bodies, non-profit organisations and public institutions to create AI systems that can explain their predictions. Companies in highly regulated areas such as finance, healthcare and human resources are particularly active to address potential financial, legal and reputational risks lined to predictions made by AI systems. For example, US bank Capital One created a research team in 2016 to find ways of making AI techniques more explainable (Knight, 2017[32]). Companies such as MondoBrain have designed user interfaces to help explain meaningful

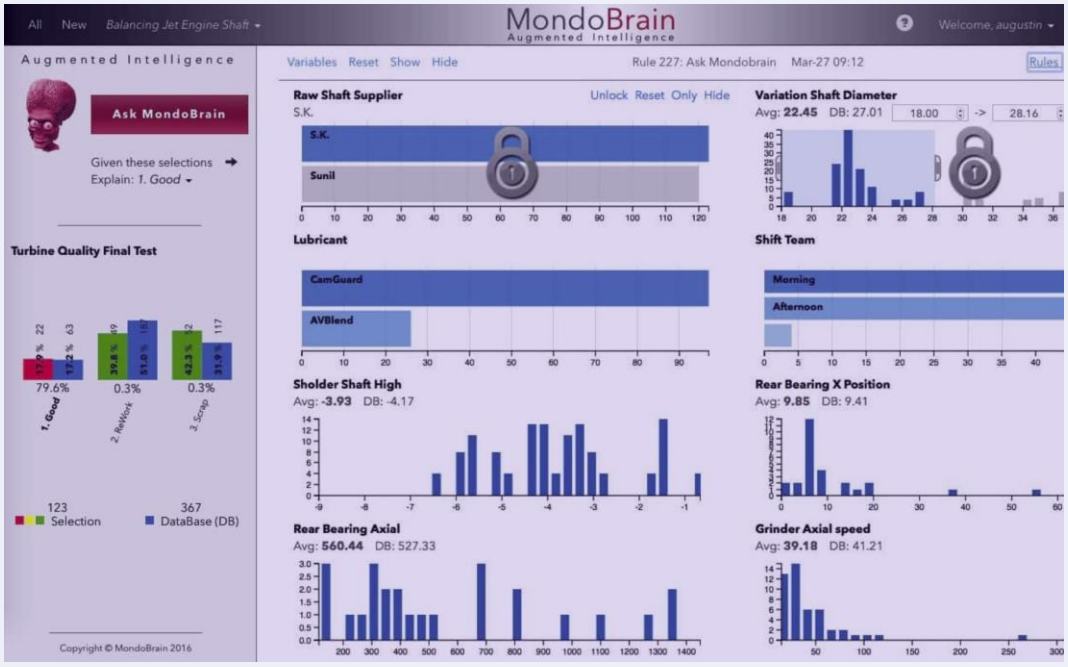
factors (Box 4.5). Non-profit organisations such as OpenAI are researching approaches to develop explainable AI and audit AI decisions. Publicly funded research is also underway. DARPA, for example, is funding 13 different research groups, working on a range of approaches to making AI more explainable.

Box 4.5. Addressing explainability issues through better-designed user interfaces

Some businesses have started to embed explainability into their solutions so that users better understand the AI processes running in the background. One example is MondoBrain. Based in France, it combines human, collective and artificial intelligence to provide an augmented reality solution for the enterprise. Through the use of interactive data-visualisation dashboards, it evaluates all existing data in a company (from Enterprise Resource Planning, Business Programme Management or Customer Relationship Management software, for instance) and provides prescriptive recommendations based on customers' queries (Figure 4.1). It uses an ML algorithm to eliminate the business variables that provide no value to the query and extract the variables with the most significant impact.

Simple traffic light colours lead users at every step of the query, facilitating their understanding of the decision process. Every single decision is automatically documented, becoming auditable and traceable. It creates a full but simple record of all the steps that led to the final business recommendation.

Figure 4.1. Illustration of data-visualisation tools to augment explainability



Source: www.mondobrain.com.

In many cases, it is possible to generate one or more of these kinds of explanations on AI systems' outcomes. However, explanations bear a cost. Designing a system to provide an explanation can be complex and expensive. Requiring explanations for all AI systems may not be appropriate in light of their purpose and may disadvantage SMEs in particular. AI systems must often be designed *ex ante* to provide a certain kind of explanation. Seeking explanations

after the fact usually requires additional work, possibly recreating the entire decision system. For example, an AI system cannot provide an explanation of all the top factors that impacted an outcome if it was designed to provide only one. Similarly, an AI system to detect heart conditions cannot be queried about the impact of gender on a diagnosis if gender data were not used to train the system. This is the case even if an AI system actually accounts for gender through proxy variables, such as other medical conditions that are more frequent in women.

In some cases, there is a trade-off between explainability and accuracy. Being explainable may require reducing the solution variables to a set small enough that humans can understand. This could be suboptimal in complex, high-dimensional problems. For example, some ML models used in medical diagnosis can accurately predict the probability of a medical condition, but are too complex for humans to understand. In such cases, the potential harm from a less accurate system that offers clear explanations should be weighed against the potential harm from a more accurate system where errors are harder to detect. For example, recidivism prediction may require simple and explainable models where errors can be detected (Dressel and Farid, 2018[33]). In areas like climate predictions, more complex models that deliver better predictions but are less explainable may be more acceptable. This is particularly the case if other mechanisms to ensure accountability exist, such as statistical data to detect possible bias or error.

Robustness, security and safety

Understanding of robustness, security and safety

Robustness can be understood as the ability to withstand or overcome adverse conditions (OECD, 2019[34]), including digital security risks. Safe AI systems can be understood as systems that do not pose unreasonable safety risks in normal or foreseeable use or misuse throughout their lifecycle (OECD, 2019[35]). Issues of robustness and safety of AI are interlinked. For example, digital security can affect product safety if connected products such as driverless cars or AI-powered home appliances are not sufficiently secure; hackers could take control of them and change settings at a distance.

Risk management in AI systems

Needed level of protections depends on risk-benefit analysis

The potential harm of an AI system should be balanced against the costs of building transparency and accountability into AI systems. Harms could include risks to human rights, privacy, fairness and robustness. But not every use of AI presents the same risks, and requiring explanation, for example, imposes its own set of costs. In managing risk, there appears to be broad-based agreement that high-stakes' contexts require higher degrees of transparency and accountability, particularly where life and liberty are at stake.

Use of risk management approaches throughout the AI lifecycle

Organisations use risk management to identify, assess, prioritise and treat potential risks that can adversely affect a system's behaviour and outcomes. Such an approach can also be used to identify risks for different stakeholders and determine how to address these risks throughout the AI system lifecycle (Section "AI system lifecycle" in Chapter 1).

AI actors – those who play an active role in the AI system lifecycle – assess and mitigate risks in the AI system as a whole, as well as in each phase of its lifecycle. Risk management in AI systems consists of the following steps, whose relevance varies depending on the phase of the AI system lifecycle:

1. **Objectives**: define objectives, functions or properties of the AI system, in context. These functions and properties may change depending on the phase of the AI lifecycle.
2. **Stakeholders and actors**: identify stakeholders and actors involved, i.e. those directly or indirectly affected by the system's functions or properties in each lifecycle phase.
3. **Risk assessment**: assess the potential effects, both benefits and risks, for stakeholders and actors. These will vary, depending on the stakeholders and actors affected, as well as the phase in the AI system lifecycle.
4. **Risk mitigation**: identify risk mitigation strategies that are appropriate to, and commensurate with, the risk. These should consider factors such as the organisation's goals and objectives, the stakeholders and actors involved, the likelihood of risks manifesting and potential benefits.
5. **Implementation**: implement risk mitigation strategies.
6. **Monitoring, evaluation and feedback**: monitor, evaluate and provide feedback on the results of the implementation.

The use of risk management in AI systems lifecycle and the documentation of the decisions at each lifecycle phase can help improve an AI system's transparency and an organisation's accountability for the system.

Aggregate harm level should be considered alongside the immediate risk context

Viewed in isolation, some uses of AI systems are low risk. However, they may require higher degrees of robustness because of their societal effects. If a system's operation results in minor harm to a large number of people, it could still collectively cause significant harm overall. Imagine, for example, if a small set of AI tools is embedded across multiple services and sectors. These could be used to obtain loans, qualify for insurance and pass background checks. A single error or bias in one system could create numerous cascading setbacks (Citron and Pasquale, 2014[36]). On its own, any one setback might be minor. Collectively, however, they could be disruptive. This suggests that policy discussions should consider aggregate harm level, in addition to the immediate risk context.

Robustness to digital security threats related to AI

AI allows more sophisticated attacks of potentially larger magnitude

AI's malicious use is expected to increase as it becomes less expensive and more accessible, in parallel to its uses to improve digital security (Subsection "AI in digital security" in Chapter 3). Cyber attackers are increasing their AI capabilities. Faster and more sophisticated attacks pose a growing concern for digital security.[6] Against this backdrop, existing threats are expanding, new threats are being introduced and the character of threats is changing.

A number of vulnerabilities characterise today's AI systems. Malicious actors can tamper with the data on which an AI system is being trained (e.g. "data poisoning"). They can also identify the characteristics used by a digital security model to flag malware. With this information, they can design unidentifiable malicious code or intentionally cause the misclassification of information (e.g. "adversarial examples") (Box 4.6) (Brundage et al., 2018[37]). As AI technologies are increasingly available, more people can use AI to carry out more sophisticated attacks of potentially larger magnitude. The frequency and efficiency of labour-intensive digital security attacks such as targeted spear phishing could grow as they are automated based on ML algorithms.

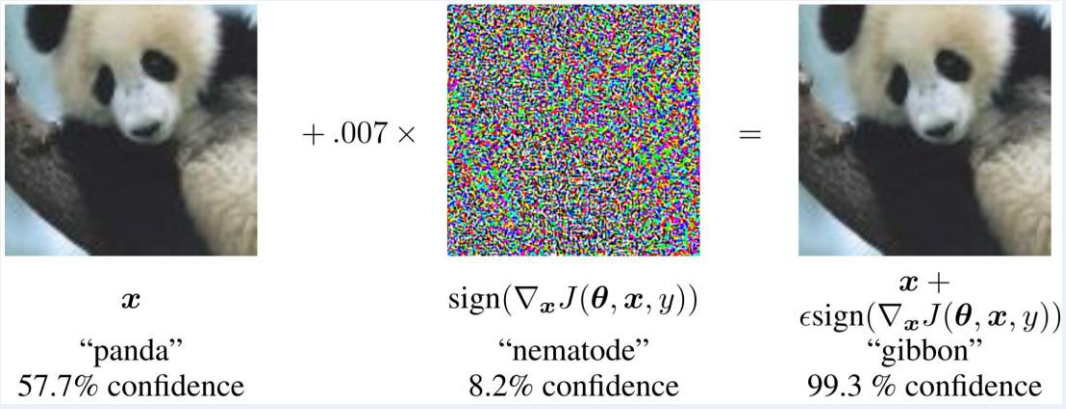
Box 4.6. The peril of adversarial examples for ML

Adversarial examples are inputs fed into ML models that an attacker intentionally designs to cause the model to make a mistake, while displaying a high degree of confidence. Adversarial examples are a real problem for AI robustness and safety because several ML models, including state-of-the-art neural networks, are vulnerable to them.

Adversarial examples can be subtle. In Figure 4.2, an imperceptibly small perturbation or "adversarial input" has been added to the image of a panda. It is specifically designed to trick the image-classification model. Ultimately, the algorithm classifies the panda as a gibbon with close to 100% confidence.

Moreover, recent research has shown that adversarial examples can be created by printing an image on normal paper and photographing it with a standard resolution smart phone. These images could be dangerous: an adversarial sticker on a stop traffic sign could trick a self-driving car into interpreting it as a "yield" or any other sign.

Figure 4.2. A small perturbation tricks an algorithm into classifying a panda as a gibbon



Sources: Goodfellow, Shlens and Szegedy (2015[38]), "Explaining and harnessing adversarial examples", https://arxiv.org/pdf/1412.6572.pdf; Kurakin, Goodfellow and Bengio (2017[39]), "Adversarial examples in the physical world", https://arxiv.org/abs/1607.02533.

Safety

Learning and autonomous AI systems impact policy frameworks for safety

The range of AI-embedded products is expanding rapidly – from robotics and driverless cars to everyday life consumer products and services such as smart appliances and smart home security systems. AI-embedded products offer significant safety benefits, while posing new practical and legal challenges to product safety frameworks (OECD, 2017[20]). Safety frameworks tend to regulate "finished" hardware products rather than software, while a number of AI software products learn and evolve throughout their lifecycle.[7] AI products can also be "autonomous" or semi-autonomous, i.e. make and execute decisions with no or little human input.

Different types of AI applications are expected to call for different policy responses (Freeman, 2017[40]). In broad terms, AI systems call for four considerations. First, they must consider how best to make sure that products are safe. In other words, products must not pose

unreasonable safety risk in normal or foreseeable use or misuse throughout their entire lifecycle. This includes cases for which there are few data to train the system on (Box 4.7). Second, they should consider who should be liable, and to what extent, for harm caused by an AI system. At the same time, they should consider which parties can contribute to the safety of autonomous machines. These parties could include users, product and sensor manufacturers, software producers, designers, infrastructure providers and data analytics companies. Third, they should consider the choice of liability principle(s). These could include strict liability, fault-based liability and the role of insurance. The opacity of some AI systems compounds the issue of liability. Fourth, they should consider how the law can be enforced, what is a "defect" in an AI product, what is the burden of proof and what remedies are available.

Box 4.7. Synthetic data for safer and more accurate AI: The case of autonomous vehicles

The use of synthetic data is gaining widespread adoption in the ML community, as it allows to simulate scenarios that are difficult to observe or replicate in real life. As Philipp Slusallek, Scientific Director at the German Research Centre for Artificial Intelligence explained, one such example is guaranteeing that a self-driving car will not hit a child running across the street.

A "digital reality" – a simulated environment replicating the relevant features of the real world – could have four effects. First, it could generate synthetic input data to train AI systems on complex situations. Second, it could validate performance and recalibrate synthetic data versus real data. Third, it could set up tests, such as for a driver's licence for AVs. Fourth, it could explore the system's decision-making process and the potential outcomes of alternative decisions. To illustrate, this approach has allowed Google to train its self-driving cars with more than 4.8 million simulated kilometres per day (equivalent to more than 500 round trips between New York City and Los Angeles).

Sources: Golson (2016[41]), "Google's self-driving cars rack up 3 million simulated miles every day", https://www.theverge.com/2016/2/1/10892020/google-self-driving-simulator-3-million-miles; Slusallek (2018[42]), *Artificial Intelligence and Digital Reality: Do We Need a CERN for AI?*, https://www.oecd-forum.org/channels/722-digitalisation/posts/28452-artificial-intelligence-and-digital-reality-do-we-need-a-cern-for-ai.

The European Union's Product Liability Directive (Directive 85/374/EEC) of 1985 establishes the principle of "liability without fault" or "strict liability". According to the principle, if a defective product causes damage to a consumer, the producer is liable even without negligence or fault. The European Commission is reviewing this directive. Preliminary conclusions find the model to be broadly appropriate (Ingels, 2017[43]). However, current and foreseeable AI technologies do impact the concepts of "product", "safety", "defect" and "damage". This makes the burden of proof more difficult.

In the AV sector, the primary concern of policy makers is ensuring safety. Policy work is needed on how to test AVs to ensure they can operate safely. This includes licensing regimes that evaluate the potential of pre-testing AV systems, or requirements that systems monitor the awareness of human drivers in fallback roles. In some cases, licensing is an issue for firms seeking to test vehicles. Governments' openness to testing also varies. There have been some calls for strict liability of manufacturers of AVs. This liability would be based on the controllability of risk. For example, it would recognise that a mere passenger of a driverless car cannot be at fault or have breached a duty of care. Legal experts suggest that even a "registered keeper" concept would not work because the keeper must be able to control the risk (Borges, 2017[44]). Some suggest that insurance could cover the risk of damage by AVs by classifying registered AVs based on risk assessments.

Working condition safety standards may require updating

Direct impacts from AI on working conditions may also include the need for new safety protocols. The imperative is growing for new or revised industry standards and technological agreements between management and workers towards reliable, safe and productive workplaces. The European Economic and Social Committee (EESC) recommended for "stakeholders to work together on complementary AI systems and their co-creation in the workplace" (EESC, 2017[45]).

Accountability

AI's growing use calls for accountability for the proper functioning of AI systems

Accountability focuses on being able to place the onus on the appropriate organisations or individuals for the proper functioning of AI systems. Criteria for accountability include respect for principles of human values and fairness, transparency, robustness and safety. Accountability is based on AI actors' individual roles, the context and the state of art. For policy makers, accountability depends on mechanisms that perform several functions. The mechanisms identify the party responsible for a specific recommendation or decision. They correct the recommendation or decision before it is acted on. They could also challenge or appeal the decision after the fact, or even challenge the system responsible for making the decision (Helgason, 1997[46]).

In practice, the accountability of AI systems often hinges on how well a system performs compared to indicators of accuracy or efficiency. Increasingly, measures also include indicators for goals of fairness, safety and robustness. However, such indicators still tend to be less used than measures of efficiency or accuracy. As with all metrics, monitoring and evaluation can be costly. Thus, the types and frequency of measurements must be commensurate with the potential risks and benefits.

The required level of accountability depends on the risk context

Policy approaches depend on context and use case. For example, accountability expectations may be higher for public sector use of AI. This is particularly true in government functions such as security and law enforcement that have the potential for substantial harms. Formal accountability mechanisms are also often required for private-sector applications in transportation, finance and healthcare, which are heavily regulated. In private-sector areas that are less heavily regulated, use of AI is less subject to formal accountability mechanisms. In these cases, technical approaches to transparency and accountability become even more important. They must ensure that systems designed and operated by private-sector actors respect societal norms and legal constraints.

Some applications or decisions may require a human to be "in the loop" to consider the social context and potential unintended consequences. When decisions significantly impact people's lives, there is broad agreement that AI-based outcomes (e.g. a score) should not be the sole decision factor. For example, the GDPR stipulates that a human must be in the loop if a decision has a significant impact on people's lives. For example, humans must be informed if AI is used to sentence criminals, make credit determinations, grant educational opportunities or conduct job screening. In high-stakes' situations, formal accountability mechanisms are often required. For example, a judge using AI for criminal sentencing is a "human-in-the-loop". However, other accountability mechanisms – including a traditional judicial appeals process – help ensure that judges consider AI recommendations as just one element in a prediction (Wachter, Mittelstadt and Floridi, 2017[47]). Low-risk contexts, such as a restaurant recommendation, could rely solely on machines. It may not require such a multi-layered approach, which may impose unnecessary costs.

AI policy environment

National policies are needed to promote trustworthy AI systems. Such policies can spur beneficial and fair outcomes for people and for the planet, especially in promising areas underserved by market-driven investments. The creation of an enabling policy environment for trustworthy AI includes, among other things, facilitating public and private investment in AI research and development and equipping people with the skills necessary to succeed as jobs evolve. The following subsections explore four policy areas that are critical to the promotion and development of trustworthy AI.

Investment in AI research and development

Long-term investment in public research can help shape AI innovation

The OECD is considering the role of innovation policies for digital transformation and AI adoption (OECD, 2018[48]). One issue being considered is the role of public research policies, knowledge transfer and co-creation policies to develop research tools and infrastructures for AI. AI calls for policy makers to reconsider the appropriate level of government involvement in AI research to address societal challenges (OECD, 2018[13]). In addition, research institutions in all areas will require capable AI systems to remain competitive, particularly in biomedical science and life science fields. New instruments such as data-sharing platforms and supercomputing facilities can help enable AI research and may call for new investments. Japan, for example, invests more than USD 120 million annually to build a high-performance computing infrastructure for universities and public research centres.

AI is considered to be a general-purpose technology with the potential to impact a large number of industries (Agrawal, Gans and Goldfarb, 2018[49]) (Brynjolfsson, Rock and Syverson, 2017[50]). AI is also considered an "invention of a method of invention" (Cockburn, Henderson and Stern, 2018[51]) that is already widely used by scientists and inventors to facilitate innovation. Entirely new industries could be created based on the scientific breakthroughs enabled by AI. This underscores the importance of basic research and of considering long time horizons in research policy (OECD, 2018[52]).

Enabling digital ecosystem for AI

AI technologies and infrastructure

Significant advances in AI technologies have taken place over recent years. This has been due to the maturity of statistical modelling techniques such as neural networks, and more particularly, deep neural networks (known as deep learning). Many of the tools to manage and use AI exist as open-source resources in the public domain. This facilitates their adoption and allows for crowdsourcing solutions to software bugs. Tools include TensorFlow (Google), Michelangelo (Uber) and Cognitive Toolkit (Microsoft). Some companies and researchers also share curated training datasets and training tools publicly to help diffuse AI technology.

AI partly owes its recent achievements to the exponential increase in computer speeds and to Moore's Law (i.e. the number of transistors in a dense integrated circuit doubles about every two years). Together, these two developments allow AI algorithms to process enormous amounts of data rapidly. As AI projects move from concept to commercial application, specialised and expensive cloud computing and graphic-processing unit resources are often needed. Trends in AI systems continue to show extraordinary growth in the computational power required. According to one estimate, the largest recent experiment, AlphaGo Zero, required 300 000 times

the computing power needed for the largest experiment just six years before (OpenAI, 16 May 2018[53]). AlphaGo Zero's achievements in chess and Go involved computing power estimated to exceed that of the world's ten most powerful supercomputers combined (OECD, 2018[52]).

Access to and use of data

Data access and sharing can accelerate or hinder progress in AI

Current ML technologies require curated and accurate data to train and evolve. Access to high-quality datasets is thus critical to AI development. Factors related to data access and sharing that can accelerate or hinder progress in AI include (OECD, forthcoming[54]):

- **Standards**: Standards are needed to allow interoperability and data re-use across applications, to promote accessibility and to ensure that data are findable, catalogued and/or searchable and re-usable.
- **Risks**: Risks to individuals, organisations and countries of sharing data can include confidentiality and privacy breaches, risks to intellectual property rights (IPRs) and commercial interests, potential national security risks and digital security risks.
- **Costs of data**: Data collection, access, sharing and re-use require up-front and follow-up investments. In addition to data acquisition, additional investments are needed for data cleaning, data curation, metadata maintenance, data storage and processing, and secure IT infrastructure.
- **Incentives**: Market-based approaches can help provide incentives to provide access to, and share data with, data markets and platforms that commercialise data and provide added-value services such as payment and data exchange infrastructure.
- **Uncertainties about data ownership**: Legal frameworks – IPRs, (cyber-) criminal law, competition law and privacy protection law – combined with the involvement of multiple parties in the creation of data have led to uncertainties around the question of "data ownership".
- **User empowerment, including AI-powered agents**: Empowering data users and facilitating data portability – as well as enabling effective consent and choice for data subjects – can encourage individuals and businesses to share personal or business data. Some underscore how AI-powered agents that know individuals' preferences could help them negotiate complex data sharing with other AI systems (Neppel, 2017[55]).
- **Trusted third parties**: Third parties can enable trust and facilitate data sharing and re-use among all stakeholders. Data intermediaries can act as certification authorities. Trusted data-sharing platforms, such as data trusts, provide high-quality data. And institutional review boards assure respect for legitimate interests of third parties.
- **Data representativeness**: AI systems make predictions based on patterns identified in training data sets. In this context, both for accuracy and fairness, training datasets must be inclusive, diverse and representative so they do not under- or misrepresent specific groups.

Policies can enhance data access and sharing for the development of AI

Policy approaches to enhance data access and sharing include (OECD, forthcoming[54]):

- **Providing access to public sector data**, including public sector data, open government data, geo-data (e.g. maps) and transportation data.

- **Facilitating data sharing in the private sector**, usually either on a voluntary basis or, for mandatory policies, restricted data sharing with trusted users. Particular focus areas are "data of public interest", data in network industries such as transportation and energy for service interoperability, and personal data portability.
- **Developing statistical/data analytic capacities**, by establishing technology centres that provide support and guidance in the use and analysis of data.
- **Developing national data strategies**, to ensure the coherence of national data governance frameworks and their compatibility with national AI strategies.

Technical approaches are emerging to address data constraints

Some ML algorithms, such as the ones applied to image recognition, exceed average human capabilities. Yet, to get to this point, they had to be trained with large databases of millions of labelled images. The need for data has encouraged active research in machine-learning techniques that require fewer data to train AI systems. Several methods can help address such lack of data.

- **Deep reinforcement learning** is an ML technique that combines deep neural networks with reinforcement learning (Subsection "Cluster 2: ML techniques" in Chapter 1). In this way, it learns to favour a specific behaviour that leads to the desired outcome (Mousave, Schukat and Howley, 2018[56]). Artificially intelligent "agents" compete through actions in a complex environment and receive either a reward or a penalty depending on whether the action led to the desired outcome or not. The agents adjust their actions according to this "feedback".[8]
- **Transfer learning or pre-training** (Pan and Yang, 2010[57]) reuses models that have been trained to perform different tasks in the same domain. For instance, some layers of a model trained to recognise cat pictures could be reused to detect images of blue dresses. In these cases, the sample of images would be orders of magnitude smaller than traditional ML algorithms require (Jain, 2017[58]).
- **Augmented data learning**, or data "synthetisation" can artificially create data through simulations or interpolations based on existing data. This effectively augments these data and improves learning. This method is particularly beneficial in cases where privacy constraints limit data usage or to simulate scenarios seldom encountered in reality (Box 4.7).[9]
- **Hybrid learning models** can model uncertainty by combining different types of deep neural networks with probabilistic or Bayesian approaches. In this way, they can model uncertainty to improve performance and explainability and reduce the likelihood of erroneous predictions (Kendall, 23 May 2017[59]).

Privacy, confidentiality and security concerns may limit data access and sharing. This could lead to a time lag between the speed at which AI systems can learn and the availability of datasets to train them. Recent cryptographic advances, such as in secure multi-party computation (MPC) and homomorphic encryption, could help enable rights-preserving data analyses. Specifically, they could let AI systems operate without collecting or accessing sensitive data (Box 4.8). AI models, in turn, can increasingly work with encrypted data.[10] These solutions are computationally intensive and may thus be difficult to scale (Brundage et al., 2018[37]).

Box 4.8. New cryptographic tools enable privacy-preserving computation

Advances in encryption have promising application in AI. For example, an ML model could be trained using combined data from multiple organisations. The process would keep data of all participants confidential. This could help overcome barriers related to privacy or confidentiality concerns. The encryption techniques that enable this form of computation – homomorphic encryption and secure MPC – were discovered years and decades ago, respectively. However, they were too inefficient for practical use. Recent algorithmic and implementation advances mean they are increasingly becoming practical tools that can perform productive analyses on real-world datasets.

- **Homomorphic encryption**: obliviously performing computation on encrypted data without needing to view the unencrypted data.
- **Secure MPC**: computing a function of data collected from many sources without revealing information about any source's data to any other source. Secure MPC protocols allow multiple parties to jointly compute algorithms, while keeping each party's input to the algorithm private.

Sources: Brundage et al. (2018[37]), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf; Dowlin (2016[60]), *CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/CryptonetsTechReport.pdf.

Alternatively, AI models could leverage blockchain technologies that also use cryptographic tools to provide secure data storage (Box 4.9). Solutions combining AI and blockchain technologies could help increase the availability of data. At the same time, they could minimise the privacy and security risks related to unencrypted data processing.

Box 4.9. Blockchain for privacy-preserving identity verification in AI

Kairos, a face-recognition enterprise solution, has incorporated blockchain technologies into its portfolio. It combines face biometrics and blockchain technology to allow users to better protect their privacy. An algorithm compares a person's image with featured facial landmarks (or identifiers) until a unique match is constructed. This match is then converted into a unique and random string of numbers, after which the original image can be discarded. This "biometric blockchain" is built under the premise that businesses or governments do not need to know who you are to verify that it is, actually, you.

Source: https://kairos.com/.

Competition

The OECD has researched the impact and policy implications of the digital transformation on competition (OECD, 2019[61]). This subsection outlines a few possible impacts on competition particularly caused by AI. It recognises the wide recognition for the procompetitive effects of AI in facilitating new entry. It also notes that much of the attention of competition policy given to large AI players is due to their role as online platforms and holders of large amounts of data. It is not connected to their use of AI as such.

A question that relates to AI more specifically is whether there is a data-driven network effect. Under such an effect, each user's utility from using certain kinds of platforms increases

whenever others use it, too. By using one of these platforms, for example, users are helping teach its algorithms how to become better at serving users (OECD, 2019[62]). Others have put forward that data exhibit decreasing returns to scale: prediction improvements become more and more marginal as data increase beyond a certain threshold. As a result, some have questioned whether AI could generate long-term competition concerns (Bajari et al., 2018[63]; OECD, 2016[64]; Varian, 2018[65]).

There may be economies of scale in terms of the business value of additional data. If a slight lead in data quality over its competitors enables a company to get many more customers, it could generate a positive feedback loop. More customers mean more data, reinforcing the cycle and allowing for increased market dominance over time. There may also be economies of scale with respect to the expertise required to build effective AI systems.

There is also a concern that algorithms could facilitate collusion through monitoring market conditions, prices and competitors' responses to price changes. These actions could provide companies with new and improved tools for co-ordinating strategies, fixing prices and enforcing cartel agreements. A somewhat more speculative concern is that more sophisticated deep-learning algorithms would not even require actual agreements among competitors to arrive at cartel-like outcomes. Instead, these would be achieved without human intervention. That would present difficult enforcement challenges. Competition laws require evidence of agreements or a "meeting of the minds" before a cartel violation can be established and punished (OECD, 2017[66]).

Intellectual property

This subsection outlines a few possible impacts to intellectual property (IP) caused by AI. It notes this is a rapidly evolving area where evidence-based analytical work is only beginning. IP rules generally accelerate the degree and speed of discovery, invention and diffusion of new technology with regard to AI. In this way, they are similar to rules for other technologies protected by IP rights. While IP rules should reward inventors, authors, artists and brand owners, IP policy should also consider AI's potential as an input for further innovation.

The protection of AI with IPRs other than trade secrets may raise new issues on how to incentivise innovators to disclose AI innovations, including algorithms and their training. A European Parliament Office conference discussed three possible types of AI patenting (EPO, 2018[67]). The first type, Core AI, is often related to algorithms, which as mathematical methods are not patentable. In the second type – trained models/ML – claiming variations and ranges might be an issue. Finally, AI could be patented as a tool in an applied field, defined via technical effects. Other international organisations and OECD countries are also exploring the impact of AI in the IP space.[11]

Another consideration raised by the diffusion of AI is whether IP systems need adjustments in a world in which AI systems can themselves make inventions (OECD, 2017[68]). Certain AI systems can already produce patentable inventions, notably in chemistry, pharmaceuticals and biotechnology. In these fields, many inventions consist of creating original combinations of molecules to form new compounds, or in identifying new properties of existing molecules. For example, KnIT, an ML tool developed by IBM, successfully identified kinases – enzymes that act as a catalyst for the transfer of phosphate groups to specific substrates. These kinases had specific properties among a set of known kinases, which were tested experimentally. Software discovered the specific properties of those molecules, and patents were filed for the inventions. These and other matters regarding AI and IP are being considered by expert agencies of OECD countries such as the European Patent Office and the US Patent and

Trademark Office, as well as by the World Intellectual Property Organization. They could also consider issues of copyright protection of AI-processed data.

Small and medium-sized enterprises

Policies and programmes to help SMEs navigate the AI transition are an increasing priority. This is a rapidly evolving area where evidence-based analytical work is beginning. Potential tools to enable digital ecosystems for SMEs to adopt and leverage AI include:

- Upskilling, which is viewed as critical because competing for scarce AI talent is a particular concern for SMEs.
- Encouraging targeted investments in selected vertical industries. Policies to encourage investment in specific AI applications in French agriculture, for example, could benefit all players where individual SMEs could not afford to invest alone (OECD, 2018[13]).
- Helping SMEs to access data, including by creating platforms for data exchange.
- Supporting SMEs' improved access to AI technologies, including through technology transfer from public research institutes, as well as their access to computing capacities and cloud platforms (Germany, 2018[69]).
- Improving financing mechanisms to help AI SMEs scale up, e.g. through a new public investment fund and increasing the flexibility and financial limits of schemes to invest in knowledge-intensive companies (UK, 2017[70]). The European Commission is also focusing on supporting European SMEs, including through its AI4EU project, an AI-on-demand platform.

Policy environment for AI innovation

The OECD is analysing changes to innovation and other AI-relevant policies needed in the context of AI and other digital transformations (OECD, 2018[48]). Under consideration is how to improve the adaptability, reactivity and versatility of policy instruments and experiments. Governments can use experimentation to provide controlled environments for the testing of AI systems. Such environments could include regulatory sandboxes, innovation centres and policy labs. Policy experiments can operate in "start-up mode". In this case, experiments are deployed, evaluated and modified, and then scaled up or down, or abandoned quickly.

Another option to spur faster and more effective decisions is the use of digital tools to design policy, including innovation policy, and to monitor policy targets. For instance, some governments use "agent-based modelling" to anticipate the impact of policy variants on different types of businesses.

Governments can encourage AI actors to develop self-regulatory mechanisms such as codes of conduct, voluntary standards and best practices. These can help guide AI actors through the AI lifecycle, including for monitoring, reporting, assessing and addressing harmful effects or misuse of AI systems.

Governments can also establish and encourage public- and private-sector oversight mechanisms of AI systems, as appropriate. These could include compliance reviews, audits, conformity assessments and certification schemes. Such mechanisms could be used while considering the specific needs of SMEs and the constraints they face.

Preparing for job transformation and building skills

Jobs

AI is expected to complement humans in some tasks, replace them in others and generate new types of work

The OECD has researched the impact of the broader digital transformation on jobs and the policy implications in depth (OECD, 2019[61]). As a rapidly evolving area where evidence-based analytical work is beginning, AI is broadly expected to change the nature of work as it diffuses across sectors. It will complement humans in some tasks, replace them in others and also generate new types of work. This section outlines some anticipated changes to labour markets caused by AI, as well as policy considerations to accompany the transition to an AI economy.

AI is expected to improve productivity

AI is expected to improve productivity in two ways. First, some activities previously carried out by people will be automated. Second, through machine autonomy, systems will operate and adapt to circumstances with reduced or no human control (OECD, 2017[68]; Autor and Salomons, 2018[71]). Research on 12 developed economies estimated that AI could increase labour productivity by up to 40% by 2035 compared to expected baseline levels (Purdy and Daugherty, 2016[72]). Examples abound. IBM's Watson assists client advisors at Crédit Mutuel, a French bank, to field client questions 60% faster.[12] Alibaba's chatbot handled more than 95% of customer inquiries during a 2017 sale. This allowed human customer representatives to handle more complicated or personal issues (Zeng, 2018[73]). In theory, increasing worker productivity should result in higher wages, since each individual employee produces more value added.

Human-AI teams help mitigate error and could expand opportunities for human workers. Human-AI teams have been found to be more productive than either AI or workers alone (Daugherty and Wilson, 2018[74]). For example, human-AI teams in BMW factories increased manufacturing productivity by 85% compared to non-integrated teams. Beyond manufacturing, Walmart robots scan for inventory, leaving store associates to focus on helping customers. And when a human radiologist combined with AI models to screen chest X-rays for tuberculosis, net accuracy reached 100% – higher than AI or human methods alone (Lakhani and Sundaram, 2017[75]).

AI can also make previously automated tasks work better and faster. As a result, companies can produce more at lower cost. If lower costs are passed down to companies or individuals, demand for the goods can be expected to increase. This boosts labour demand both in the company – for instance, in production-related roles – as well as in downstream sectors for intermediate goods.

AI is expected to change – perhaps accelerate – the tasks that can be automated

Automation is not a new phenomenon, but AI is expected to change, and perhaps accelerate, the profile of tasks that can be automated. Unlike computers, AI technologies are not strictly pre-programmed and rules-based. Computers have tended to reduce employment in routine, middle-skill occupations. However, new AI-powered applications can increasingly perform relatively complex tasks that involve making predictions (see Chapter 3). These tasks include transcription, translation, driving vehicles, diagnosing illness and answering customer inquiries

(Graetz and Michaels, 2018[76]; Michaels, Natraj and Van Reenen, 2014[77]; Goos, Manning and Salomons, 2014[78]).[13]

Exploratory OECD measurement estimated the extent to which technologies can answer the literacy and numeracy questions of the *OECD Survey of Adult Skills* (*PIAAC*) (Elliott, 2017[79]). This research suggested that, in 2017, AI systems could answer the literacy questions at a level comparable to that of 89% of adults in OECD countries. In other words, only 11% of adults were above the level that AI was close to reproducing in terms of literacy skills. The report predicted more economic pressure to apply computer capabilities for certain literacy and numeracy skills. This would likely decrease demand for human workers to perform tasks using low- and mid-level literacy skills, reversing recent patterns. The report underscored the difficulty of designing education policies to adults above the current computer level. It suggested new tools and incentives for promoting adult skills or combining skills policies with other interventions, including social protection and social dialogue (OECD, 2018[13]).

AI's impact on jobs will depend on its speed of diffusion across different sectors

AI's impact on jobs will also depend on the speed of the development and diffusion of AI technologies in different sectors over the coming decades. AVs are widely expected to disrupt driving and delivery service jobs. Established truck companies such as Volvo and Daimler, for example, are competing with start-ups like Kodiak and Einride to develop and test driverless trucks (Stewart, 2018[80]). According to the International Transport Forum, driverless trucks may be a regular presence on many roads within the next ten years. Some 50% to 70% of the 6.4 million professional trucking jobs in the United States and Europe could be eliminated by 2030 (ITF, 2017[81]). However, new jobs will be created in parallel to provide support services for the increased number of driverless trucks. Driverless trucks could reduce operating costs for road freight in the order of 30%, notably due to savings in labour costs. This could drive traditional trucking companies out of business, resulting in an even faster decline in trucking jobs.

AI technologies are likely to impact traditionally higher-skilled tasks

AI technologies are also performing prediction tasks traditionally performed by higher-skilled workers, from lawyers to medical personnel. A robolawyer has successfully appealed over USD 12 million worth of traffic tickets (Dormehl, 2018[82]). In 2016, IBM's Watson and DeepMind Health outperformed human doctors in diagnosing rare cancers (Frey and Osborne, 2017[83]). AI has proven to be better at predicting stock exchange variations than finance professionals (Mims, 2010[84]).

AI can complement people and create new types of work

AI complements people and is also likely to create job opportunities for human workers. Notable areas include those that complement prediction and leverage human skills such as critical thinking, creativity and empathy (EOP, 2016[85]; OECD, 2017[20]).

- **Data scientists and ML experts**: Specialists are needed to create and clean data and to program and develop AI applications. However, although data and ML lead to some new tasks, they are unlikely to generate large numbers of new tasks for workers.
- **Actions**: Some actions are inherently more valuable when done by a human than a machine, as professional athletes, child carers or salespeople illustrate. Many think

it is likely that humans will increasingly focus on work to improve each other's lives, such as childcare, physical coaching and care for the terminally ill.

- **Judgment to determine what to predict**: Perhaps most important is the concept of judgment – the process of determining the reward to a particular action in a particular environment. When AI is used for predictions, a human must decide what to predict and what to do with the predictions. Posing dilemmas, interpreting situations or extracting meaning from text requires people with qualities such as judgment and fairness (OECD, 2018[13]). In science, for example, AI can complement humans in charge of the conceptual thinking necessary to build research frameworks and to set the context for specific experiments.
- **Judgment to decide what to do with a prediction**: A decision cannot be made with a prediction alone. For example, the trivial decision of whether to take an umbrella when going outside for a walk will consider a prediction about the likelihood of rain. However, the decision will depend largely on preferences such as the degree to which one dislikes being wet and carrying an umbrella. This example can be broadened to many important decisions. In cybersecurity, a prediction about whether a new inquiry is hostile will need to be measured against the risk of turning away a friendly inquiry and letting a hostile inquiry obtain unauthorised information.

Predictions regarding AI's net impact on the quantity of work vary widely

Over the past five years, widely varying estimates have been made of the overall impacts of automation on job loss (Winick, 2018[86]; MGI, 2017[87]; Frey and Osborne, 2017[83]). For example, a Frey and Osborne predicted that 47% of US jobs are at risk of displacement in the next 10 to 15 years. Using a task-oriented approach, the McKinsey Global Institute found in 2017 that about one-third of activities in 60% of jobs are automatable. However, identified jobs affected by automation are not due to the development and deployment of AI alone, but also to other technological developments.

In addition, anticipating future job creation in new areas is challenging. One study estimated that AI would lead to a net job creation of 2 million by 2025 (Gartner, 2017[88]). Job creation is likely both as a result of new occupations arising and through more indirect channels. For example, AI is likely to reduce the cost of producing goods and services, as well as to increase their quality. This will lead to increased demand and, as a result, higher employment.

The most recent OECD estimates allow for heterogeneity of tasks within narrowly defined occupations, using data of the Programme for the International Assessment of Adult Competencies (PIAAC). Based on existing technologies, 14% of jobs in member countries are at high risk of automation; another 32% of workers are likely to see substantial change in how their jobs are carried out (Nedelkoska and Quintini, 2018[89]). The risk of automation is highest among teenagers and senior workers. Recent OECD analysis finds employment decline in occupations classified as "highly automatable" in 82% of regions across 16 European countries. At the same time, it identifies a greater increase in "low automation" jobs in 60% of regions that offsets job loss. This research supports the idea that automation may be shifting the mix of jobs, without driving down overall employment (OECD, 2018[90]).

AI will change the nature of work

AI adoption is broadly expected to change the nature of work. AI may help make work more interesting by automating routine tasks, allowing more flexible work and possibly a better work-life balance. Human creativity and ingenuity can leverage increasingly powerful

computation, data and algorithm resources to create new tasks and directions that require human creativity (Kasparov, 2018[91]).

More broadly, AI may accelerate changes to how the labour market operates by increasing efficiency. Today, AI techniques coupled with big data hold potential to help companies to identify roles for workers – as well as participate in matching people to jobs. IBM, for example, uses AI to optimise employee training, recommending training modules to employees based on their past performance, career goals and IBM skills needs. Companies such as KeenCorp and Vibe have developed text analytics techniques to help companies parse employee communications to help assess metrics such as morale, worker productivity and network effects (Deloitte, 2017[92]). As a result of this information, AI may help companies optimise worker productivity.

Parameters for organisational change will need to be set

The imperative is growing for new or revised industry standards and technological agreements between management and workers towards reliable, safe and productive workplaces. The EESC recommended for "stakeholders to work together on complementary AI systems and their co-creation in the workplace" (EESC, 2017[45]). Workplaces also need flexibility, while safeguarding workers' autonomy and job quality, including the sharing of profits. The recent collective agreement between the German sector union *IG Metall* and employers (*Gesamtmetall)* gives an economic case for variable working times. It shows that, depending on organisational and personal (care) needs in the new world of work, employers and unions can reach agreements without revising legal employment protections (Byhovskaya, 2018[93]).

Using AI to support labour market functions – with safeguards – is also promising

AI has already begun to make job matching and training more efficient. It can help better connect job seekers, including displaced workers, with the workforce development programmes they need to qualify for emerging and expanding occupations. In many OECD countries, employers and public employment services already use online platforms to fill jobs (OECD, 2018[90]). Looking ahead, AI and other digital technologies can improve innovative and personalised approaches to job-search and hiring processes and enhance the efficiency of labour supply and demand matching. The LinkedIn platform uses AI to help recruiters find the right candidates and to connect candidates to the right jobs. It draws on data about the profile and activity of the platform's 470 million registered users (Wong, 2017[94]).

AI technologies leveraging big data can also help inform governments, employers and workers about local labour market conditions. This information can help identify and forecast skills demands, direct training resources and connect individuals with jobs. Projects to develop labour market information are already underway in countries such as Finland, the Czech Republic and Latvia (OECD, 2018[90]).

Governing the use of workers' data

While AI requires large datasets to be productive, there are some potential risks when these data represent individual workers, especially if the AI systems that analyse the data are opaque. Human resources and productivity planning will increasingly leverage employee data and algorithms. As they do, public policy makers and stakeholders could investigate how data collection and processing affect employment prospects and terms. Data may be collected from applications, fingerprints, wearables and sensors in real time, indicating the location and workplace of an employee. In customer service, AI software analyses the

friendliness of employees' tone. According to workers' accounts, however, it did not consider speech patterns and challenging the scoring was difficult (UNI, 2018[95]).

In contrast, agreements on workers' data and the right to disconnect are emerging in some countries. The French telecommunications company Orange France Telecom and five trade union centres were among the first to settle on commitments to protect employee data. Specific protections include transparency over use, training and the introduction of new equipment. To close the regulatory gap on workers' data, provisions could include establishing data governance bodies in companies, accountability on behalf of (personal) data use, data portability, explanation and deletion rights (UNI, 2018[95]).

Managing the AI transition

Policies for managing the AI transition, including social protection, are key

There is a possibility of disruption and turbulence in labour markets as technology outpaces organisational adaptation (OECD, 2018[13]). Long-term optimism does not imply a smooth transition to an economy with more and more AI: some sectors are likely to grow, while others decline. Existing jobs may disappear, while new ones are created. Thus, key policy questions with respect to AI and jobs relate to managing the transition. Policies for managing the transition include social safety nets, health insurance, progressive taxation of labour and capital, and education. Moreover, OECD analysis also points to the need for attention to competition policies and other policies that might affect concentration, market power and income distribution (OECD, 2019[61]).

Skills to use AI

As jobs change, so will the skills required of workers

As jobs change, so will the skills required of workers (OECD, 2017[96]; Acemoglu and Restrepo, 2018[97]; Brynjolfsson and Mitchell, 2017[98]). The present subsection outlines a few possible repercussions of AI on skills, noting this is a rapidly evolving area where evidence-based analytical work is only beginning. Education policy is expected to require adjustments to expand lifelong learning, training and skills development. As with other areas of technology, AI is expected to generate demand in three skills areas. First, **specialist skills** will be needed to program and develop AI applications. These could include skills for AI-related fundamental research, engineering and applications, as well as data science and computational thinking. Second, **generic skills** will be needed to leverage AI, including through AI-human teams on the factory floor and quality control. Third, AI will need **complementarity skills**. These could include leveraging human skills such as critical thinking; creativity, innovation and entrepreneurship; and empathy (EOP, 2016[85]; OECD, 2017[20]).

Initiatives to build and develop AI skills are required to address AI skills shortage

The AI skills shortage is expected to grow, and may become more evident as demand for specialists in areas such as ML accelerates. SMEs, public universities and research centres already compete with dominant firms for talent. Initiatives to build and develop AI skills are starting to emerge in the public, private and academic sectors. For instance, the Singaporean government has set up a five-year research programme on governance of AI and data use in Singapore Management University. Its Centre for AI & Data Governance focuses on industry-relevant research, covering AI and industry, society and commercialisation. On the academic side, the Massachusetts Institute of Technology (MIT) has committed USD 1 billion

to create the Schwarzman College of Computing. It aims to equip students and researchers in all disciplines to use computing and AI to advance their disciplines and vice versa.

The AI skills shortage has also led some countries to streamline immigration processes for high-skilled experts. For example, the United Kingdom doubled the number of its Tier 1 (Exceptional Talent) visas to 2 000 a year and streamlined the process for top students and researchers to work there (UK, 2017[99]). Similarly, Canada introduced two-week processing times for visa applications from high-skilled workers and visa exemptions for short-term research assignments. This was part of its 2017 Global Skills Strategy to attract high-skilled workers and researchers from abroad (Canada, 2017[100]).

Generic skills to be able to leverage AI

All OECD countries assess skills and anticipate need for skills in the current, medium or long term. Finland proposed the Artificial Intelligence Programme, which includes a skills account or voucher-based lifelong learning programme to create demand for education and training (Finland, 2017[101]). The United Kingdom is promoting a diverse AI workforce and investing about GBP 406 million (USD 530 million) in skills. It focuses on science, technology, engineering and mathematics, and computer science teachers (UK, 2017[99]).

Practitioners must now be what some call "bilinguals". These are people who may be specialised in one area such as economics, biology or law, but who are also skilled at AI techniques such as ML. In this vein, the MIT announced in October 2018 the most significant change to its structure in 50 years. It plans a new school of computing that will sit outside the engineering discipline and intertwine with all other academic departments. It will train these "bilingual" students who apply AI and ML to the challenges of their own disciplines. This represents a complete shift in the way the MIT teaches computer science. The MIT is allocating USD 1 billion for the creation of this new college within the Institute (MIT, 2018[102]).

Complementary skills

There is a strong focus on emerging, "softer" skills. Based on existing research, these skills may include human judgment, analysis and interpersonal communication (Agrawal, Gans and Goldfarb, 2018[103]; Deming, 2017[104]; Trajtenberg, 2018[105]). In 2021, the OECD will include a module on the Programme for International Student Assessment (PISA) to test creative and critical thinking skills. The results will help provide a benchmark creativity assessment across countries to inform policy and social partner actions.

Measurement

The implementation of human-centred and trustworthy AI depends on context. However, a key part of policy makers' commitment to ensuring human-centred AI will be to identify objectives and metrics to assess performance of AI systems. These include areas such as accuracy, efficiency, advancement of societal goals, fairness and robustness.

References

Abrams, M. et al. (2017), *Artificial Intelligence, Ethics and Enhanced Data Stewardship*, The Information Accountability Foundation, Plano, Texas. [16]

Acemoglu, D. and P. Restrepo (2018), *Artificial Intelligence, Automation and Work*, National Bureau of Economic Research , Cambridge, MA, http://dx.doi.org/10.3386/w24196. [97]

Agrawal, A., J. Gans and A. Goldfarb (2018), "Economic policy for artificial intelligence", *NBER Working Paper*, No. 24690, http://dx.doi.org/10.3386/w24690. [49]

Agrawal, A., J. Gans and A. Goldfarb (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business School Press, Brighton, MA. [103]

Autor, D. and A. Salomons (2018), "Is automation labor-displacing? Productivity growth, employment, and the labor share", *NBER Working Paper*, No. 24871, http://dx.doi.org/10.3386/w24871. [71]

Bajari, P. et al. (2018), "The impact of big data on firm performance: An empirical investigation", *NBER Working Paper*, No. 24334, http://dx.doi.org/10.3386/w24334. [63]

Barocas, S. and A. Selbst (2016), "Big data's disparate impact", *California Law Review*, Vol. 104, pp. 671-729, http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf. [29]

Berk, R. and J. Hyatt (2015), "Machine learning forecasts of risk to inform sentencing decisions", *Federal Sentencing Reporter*, Vol. 27/4, pp. 222-228, http://dx.doi.org/10.1525/fsr.2015.27.4.222. [23]

Borges, G. (2017), *Liability for Machine-Made Decisions: Gaps and Potential Solutions*, presentation at the "AI: Intelligent Machines, Smart Policies" conference, Paris, 26-27 October, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-borges.pdf. [44]

Brundage, M. et al. (2018), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Centre for a New American Security, Electronic Frontier Foundation and Open AI, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf. [37]

Brynjolfsson, E. and T. Mitchell (2017), "What can machine learning do? Workforce implications", *Science*, Vol. 358/6370, pp. 1530-1534, http://dx.doi.org/10.1126/science.aap8062. [98]

Brynjolfsson, E., D. Rock and C. Syverson (2017), "Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics", *NBER Working Paper*, No. 24001, http://dx.doi.org/10.3386/w24001. [50]

Burgess, M. (2016), "Holding AI to account: Will algorithms ever be free of bias if they are created by humans?", *WIRED,* 11 January, https://www.wired.co.uk/article/creating-transparent-ai-algorithms-machine-learning. [31]

Byhovskaya, A. (2018), *Overview of the National Strategies on Work 4.0: A Coherent Analysis of the Role of the Social Partners*, European Economic and Social Committee, Brussels, https://www.eesc.europa.eu/sites/default/files/files/qe-02-18-923-en-n.pdf. [93]

Canada (2017), "Government of Canada launches the Global Skills Strategy", News Release, Immigration, Refugees and Citizenship Canada, 12 June, https://www.canada.ca/en/immigration-refugees-citizenship/news/2017/06/government_of_canadalaunchestheglobalskillsstrategy.html. [100]

Cellarius, M. (2017), *Artificial Intelligence and the Right to Informational Self-determination*, The OECD Forum, OECD, Paris, https://www.oecd-forum.org/users/75927-mathias-cellarius/posts/28608-artificial-intelligence-and-the-right-to-informational-self-determination. [9]

Chouldechova, A. (2016), "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", *arXiv*, Vol. 07524, https://arxiv.org/abs/1610.07524. [24]

Citron, D. and F. Pasquale (2014), "The scored society: Due process for automated predictions", *Washington Law Review*, Vol. 89, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376209. [36]

Cockburn, I., R. Henderson and S. Stern (2018), "The impact of artificial intelligence on innovation", *NBER Working Paper*, No. 24449, http://dx.doi.org/10.3386/w24449. [51]

Crawford, K. (2016), "Artificial intelligence's white guy problem", *The New York Times,* 26 June, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=0. [30]

Daugherty, P. and H. Wilson (2018), *Human Machine: Reimagining Work in the Age of AI*, Harvard Business Review Press, Cambridge, MA. [74]

Deloitte (2017), *HR Technology Disruptions for 2018: Productivity, Design and Intelligence Reign*, Deloitte, http://marketing.bersin.com/rs/976-LMP-699/images/HRTechDisruptions2018-Report-100517.pdf. [92]

Deming, D. (2017), "The growing importance of social skills in the labor market", *The Quarterly Journal of Economics*, Vol. 132/4, pp. 1593-1640, http://dx.doi.org/10.1093/qje/qjx022. [104]

Dormehl, L. (2018), "Meet the British whiz kid who fights for justice with robo-lawyer sidekick", *Digital Trends,* 3 March, https://www.digitaltrends.com/cool-tech/robot-lawyer-free-acess-justice/. [82]

Doshi-Velez, F. et al. (2017), "Accountability of AI under the law: The role of explanation", *arXiv* 21 November, https://arxiv.org/pdf/1711.01134.pdf. [28]

Dowlin, N. (2016), *CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*, Microsoft Research, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/CryptonetsTechReport.pdf. [60]

Dressel, J. and H. Farid (2018), "The accuracy, fairness and limits of predicting recidivism", *Science Advances*, Vol. 4/1, http://advances.sciencemag.org/content/4/1/eaao5580. [33]

EESC (2017), *Artificial Intelligence – The Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society*, European Economic and Social Committee, Brussels, https://www.eesc.europa.eu/en/our-work/opinions-inf. [45]

Elliott, S. (2017), *Computers and the Future of Skill Demand*, Educational Research and Innovation, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264284395-en. [79]

EOP (2016), *Artificial Intelligence, Automation and the Economy*, Executive Office of the President, Government of the United States, https://www.whitehouse.gov/sites/whitehouse.gov/files/images/EMBARGOED AI Economy Report.pdf. [85]

EPO (2018), *Patenting Artificial Intelligence - Conference Summary*, European Patent Office, Munich, 30 May, http://documents.epo.org/projects/babylon/acad.nsf/0/D9F20464038C0753C125829E0031B814/$FILE/summary_conference_artificial_intelligence_en.pdf. [67]

Finland (2017), *Finland's Age of Artificial Intelligence - Turning Finland into a Leader in the Application of AI*, webpage, Finnish Ministry of Economic Affairs and Employment, https://tem.fi/en/artificial-intelligence-programme. [101]

Flanagan, M., D. Howe and H. Nissenbaum (2008), "Embodying values in technology: Theory and practice", in van den Hoven, J. and J. Weckert (eds.), *Information Technology and Moral Philosophy*, Cambridge University Press, Cambridge, http://dx.doi.org/10.1017/cbo9780511498725.017. [15]

Freeman, R. (2017), *Evolution or Revolution? The Future of Regulation and Liability for AI*, presentation at the "AI: Intelligent Machines, Smart Policies" conference, Paris, 26-27 October, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-freeman.pdf. [40]

Frey, C. and M. Osborne (2017), "The future of employment: How susceptible are Jobs to computerisation?", *Technological Forecasting and Social Change*, Vol. 114, pp. 254-280, http://dx.doi.org/10.1016/j.techfore.2016.08.019. [83]

Gartner (2017), "Gartner says by 2020, artificial intelligence will create more jobs than it eliminates", Gartner, Press Release, 13 December, https://www.gartner.com/en/newsroom/press-releases/2017-12-13-gartner-says-by-2020-artificial-intelligence-will-create-more-jobs-than-it-eliminates. [88]

Germany (2018), "Key points for a federal government strategy on artificial intelligence", Press Release, 18 July, BMWI, https://www.bmwi.de/Redaktion/EN/Pressemitteilungen/2018/20180718-key-points-for-federal-government-strategy-on-artificial-intelligence.html. [69]

Golson, J. (2016), "Google's self-driving cars rack up 3 million simulated miles every day", *The Verge,* 1 February, https://www.theverge.com/2016/2/1/10892020/google-self-driving-simulator-3-million-miles. [41]

Goodfellow, I., J. Shlens and C. Szegedy (2015), "Explaining and harnessing adversarial examples", *arXiv*, Vol. 1412.6572, https://arxiv.org/pdf/1412.6572.pdf. [38]

Goos, M., A. Manning and A. Salomons (2014), "Explaining job polarization: Routine-biased technological change and offshoring", *American Economic Review*, Vol. 104/8, pp. 2509-2526, http://dx.doi.org/10.1257/aer.104.8.2509. [78]

Graetz, G. and G. Michaels (2018), "Robots at work", *Review of Economics and Statistics*, Vol. 100/5, pp. 753-768, http://dx.doi.org/10.1162/rest_a_00754. [76]

Harkous, H. (2018), "Polisis: Automated analysis and presentation of privacy policies using deep learning", *arXiv* 29 June, https://arxiv.org/pdf/1802.02561.pdf. [14]

Heiner, D. and C. Nguyen (2018), "Amplify Human Ingenuity with Intelligent Technology", *Shaping Human-Centered Artificial Intelligence, A.Ideas Series*, The Forum Network, OECD, Paris, https://www.oecd-forum.org/users/86008-david-heiner-and-carolyn-nguyen/posts/30653-shaping-human-centered-artificial-intelligence. [6]

Helgason, S. (1997), *Towards Performance-Based Accountability: Issues for Discussion*, Public Management Service, OECD Publishing, Paris, http://www.oecd.org/governance/budgeting/1902720.pdf. [46]

Ingels, H. (2017), *Artificial Intelligence and EU Product Liability Law*, presentation at the "AI: Intelligent Machines, Smart Policies" conference, Paris, 26-27 October, http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-ingels.pdf. [43]

ITF (2017), "Driverless trucks: New report maps out global action on driver jobs and legal issues" , International Transport Forum, https://www.itf-oecd.org/driverless-trucks-new-report-maps-out-global-action-driver-jobs-and-legal-issues. [81]

Jain, S. (2017), "NanoNets : How to use deep learning when you have limited data, Part 2 : Building object detection models with almost no hardware", *Medium,* 30 January, https://medium.com/nanonets/nanonets-how-to-use-deep-learning-when-you-have-limited-data-f68c0b512cab. [58]

Kasparov, G. (2018), *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*, Public Affairs, New York. [91]

Kendall, A. (23 May 2017), "Deep learning is not good enough, we need Bayesian deep learning for safe AI", Alex Kendall blog, https://alexgkendall.com/computer_vision/bayesian_deep_learning_for_safe_ai/. [59]

Knight, W. (2017), "The financial world wants to open AI's black boxes", *MIT Technology Review,* 13 April, https://www.technologyreview.com/s/604122/the-financial-world-wants-to-open-ais-black-boxes/. [32]

Kosack, S. and A. Fung (2014), "Does transparency improve governance?", *Annual Review of Political Science*, Vol. 17, pp. 65-87, https://www.annualreviews.org/doi/pdf/10.1146/annurev-polisci-032210-144356. [26]

Kosinski, M., D. Stillwell and T. Graepel (2013), "Private traits and attributes are predictable from digital records of human behavior", *PNAS,* 11 March, http://www.pnas.org/content/pnas/early/2013/03/06/1218772110.full.pdf. [2]

Kurakin, A., I. Goodfellow and S. Bengio (2017), "Adversarial examples in the physical world", *arXiv* 02533, https://arxiv.org/abs/1607.02533. [39]

Lakhani, P. and B. Sundaram (2017), "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks", *Radiology*, Vol. 284/2, pp. 574-582, http://dx.doi.org/10.1148/radiol.2017162326. [75]

Matheson, R. (2018), *Artificial intelligence model "learns" from patient data to make cancer treatment less toxic*, 9 August, http://news.mit.edu/2018/artificial-intelligence-model-learns-patient-data-cancer-treatment-less-toxic-0810. [107]

MGI (2017), *Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation*, McKinsey Global Institute, New York. [87]

Michaels, G., A. Natraj and J. Van Reenen (2014), "Has ICT polarized skill demand? Evidence from eleven countries over twenty-five years", *Review of Economics and Statistics*, Vol. 96/1, pp. 60-77, http://dx.doi.org/10.1162/rest_a_00366. [77]

Mims, C. (2010), "AI that picks stocks better than the pros", *MIT Technology Review,* 10 June, https://www.technologyreview.com/s/419341/ai-that-picks-stocks-better-than-the-pros/. [84]

MIT (2018), "Cybersecurity's insidious new threat: Workforce stress", *MIT Technology Review,* 7 August, https://www.technologyreview.com/s/611727/cybersecuritys-insidious-new-threat-workforce-stress/. [102]

Mousave, S., M. Schukat and E. Howley (2018), "Deep reinforcement learning: An overview", *arXiv* 1806.08894, https://arxiv.org/abs/1806.08894. [56]

Narayanan, A. (2018), "Tutorial: 21 fairness definitions and their politics", https://www.youtube.com/watch?v=jIXIuYdnyyk. [17]

Nedelkoska, L. and G. Quintini (2018), "Automation, skills use and training", *OECD Social, Employment and Migration Working Papers*, No. 202, OECD Publishing, Paris, https://dx.doi.org/10.1787/2e2f4eea-en. [89]

Neppel, C. (2017), *AI: Intelligent Machines, Smart Policies*, presentation at the "AI: Intelligent Machines, Smart Policies" conference, Paris, 26-27 October, http://oe.cd/ai2017. [55]

NITI (2018), *National Strategy for Artificial Intelligence #AIforall*, NITI Aayog, June, http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf. [5]

OECD (2019), *An Introduction to Online Platforms and Their Role in the Digital Transformation*, OECD Publishing, Paris, https://dx.doi.org/10.1787/53e5f593-en. [62]

OECD (2019), *Going Digital: Shaping Policies, Improving Lives*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264312012-en. [61]

OECD (2019), *Recommendation of the Council on Artificial Intelligence*, OECD, Paris. [35]

OECD (2019), *Scoping Principles to Foster Trust in and Adoption of AI – Proposal by the Expert Group on Artificial Intelligence at the OECD (AIGO)*, OECD, Paris, http://oe.cd/ai. [34]

OECD (2018), "AI: Intelligent machines, smart policies: Conference summary", *OECD Digital Economy Papers*, No. 270, OECD Publishing, Paris, http://dx.doi.org/10.1787/f1a650d9-en. [13]

OECD (2018), *Job Creation and Local Economic Development 2018: Preparing for the Future of Work*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264305342-en. [90]

OECD (2018), *OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption*, OECD Publishing, Paris, https://dx.doi.org/10.1787/sti_in_outlook-2018-en. [52]

OECD (2018), "Perspectives on innovation policies in the digital age", in *OECD Science, Technology and Innovation Outlook 2018: Adapting to Technological and Societal Disruption*, OECD Publishing, Paris, https://dx.doi.org/10.1787/sti_in_outlook-2018-8-en. [48]

OECD (2017), *Algorithms and Collusion: Competition Policy in the Digital Age*, OECD Publishing, Paris, http://www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.html. [66]

OECD (2017), *Getting Skills Right: Skills for Jobs Indicators*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264277878-en. [96]

OECD (2017), *OECD Digital Economy Outlook 2017*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264276284-en. [20]

OECD (2017), *The Next Production Revolution: Implications for Governments and Business*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264271036-en. [68]

OECD (2016), *Big Data: Bringing Competition Policy to the Digital Era (Executive Summary)*, OECD DAF Competition Committee, https://one.oecd.org/document/DAF/COMP/M(2016)2/ANN4/FINAL/en/pdf. [64]

OECD (2013), *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, OECD, Paris, http://www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf. [12]

OECD (2011), *OECD Guidelines for Multinational Enterprises, 2011 Edition*, OECD Publishing, Paris, https://dx.doi.org/10.1787/9789264115415-en. [8]

OECD (forthcoming), *Enhanced Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-Use across Societies*, OECD Publishing, Paris. [54]

OHCHR (2011), *Guiding Principles on Business and Human Rights*, United Nations Human Rights Office of the High Commissioner, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf. [7]

O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Broadway Books, New York. [25]

OpenAI (16 May 2018), "AI and compute", OpenAI blog, San Francisco, https://blog.openai.com/ai-and-compute/. [53]

Pan, S. and Q. Yang (2010), "A survey on transfer learning", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22/10, pp. 1345-1359. [57]

Paper, I. (ed.) (2018), "Artificial intelligence and privacy", June, Office of the Victorian Information Commissioner, https://ovic.vic.gov.au/wp-content/uploads/2018/08/AI-Issues-Paper-V1.1.pdf. [11]

Patki, N., R. Wedge and K. Veeramachaneni (2016), "The Synthetic Data Vault", *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, http://dx.doi.org/10.1109/dsaa.2016.49. [106]

Privacy International and Article 19 (2018), *Privacy and Freedom of Expression in the Age of Artificial Intelligence*, https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf. [10]

Purdy, M. and P. Daugherty (2016), "Artificial intelligence poised to double annual economic growth rate in 12 developed economies and boost labor productivity by up to 40 percent by 2035, according to new research by Accenture", Accenture, Press Release, 28 September, http://www.accenture.com/futureofAI. [72]

Selbst, A. (2017), "Disparate impact in big data policing", *Georgia Law Review*, Vol. 52/109, http://dx.doi.org/10.2139/ssrn.2819182. [21]

Simonite, T. (2018), "Probing the dark side of Google's ad-targeting system", *MIT Technology Review,* 6 July, https://www.technologyreview.com/s/539021/probing-the-dark-side-of-googles-ad-targeting-system/. [19]

Slusallek, P. (2018), *Artificial Intelligence and Digital Reality: Do We Need a CERN for AI?*, The Forum Network, OECD, Paris, https://www.oecd-forum.org/channels/722-digitalisation/posts/28452-artificial-intelligence-and-digital-reality-do-we-need-a-cern-for-ai. [42]

Smith, M. and S. Neupane (2018), *Artificial Intelligence and Human Development: Toward a Research Agenda*, International Development Research Centre, Ottawa, https://idl-bnc-idrc.dspacedirect.org/handle/10625/56949. [4]

Stewart, J. (2018), "As Uber gives up on self-driving trucks, another startup jumps in", *WIRED,* 8 July, https://www.wired.com/story/kodiak-self-driving-semi-trucks/. [80]

Talbot, D. et al. (2017), "Charting a roadmap to ensure AI benefits all", *Medium,* 30 November, https://medium.com/berkman-klein-center/charting-a-roadmap-to-ensure-artificial-intelligence-ai-benefits-all-e322f23f8b59. [3]

Trajtenberg, M. (2018), "AI as the next GPT: A political-economy perspective"*, NBER Working Paper*, No. 24245, http://dx.doi.org/10.3386/w24245. [105]

UK (2017), *UK Digital Strategy*, Government of the United Kingdom, https://www.gov.uk/government/publications/uk-digital-strategy/uk-digital-strategy. [70]

UK (2017), *UK Industrial Strategy: A Leading Destination to Invest and Grow*, Great Britain and Northern Ireland, http://https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/668161/uk-industrial-strategy-international-brochure.pdf. [99]

UNI (2018), *10 Principles for Workers' Data Rights and Privacy*, UNI Global Union, http://www.thefutureworldofwork.org/docs/10-principles-for-workers-data-rights-and-privacy/. [95]

Varian, H. (2018), "Artificial intelligence, economics and industrial organization"*, NBER Working Paper*, No. 24839, http://dx.doi.org/10.3386/w24839. [65]

Wachter, S., B. Mittelstadt and L. Floridi (2017), "Transparent, explainable and accountable AI for robotics", *Science Robotics,* 31 May, http://robotics.sciencemag.org/content/2/6/eaan6080. [47]

Wachter, S., B. Mittelstadt and C. Russell (2017), "Counterfactual explanations without opening the black box: Automated decisions and the GDPR", *arXiv* 00399, https://arxiv.org/pdf/1711.00399.pdf. [27]

Weinberger, D. (2018), "Optimization over explanation - Maximizing the benefits of machine learning without sacrificing its intelligence", *Medium,* 28 January, https://medium.com/@dweinberger/optimization-over-explanation-maximizing-the-benefits-we-want-from-machine-learning-without-347ccd9f3a66. [1]

Weinberger, D. (2018), *Playing with AI Fairness*, Google PAIR, 17 September, https://pair-code.github.io/what-if-tool/ai-fairness.html. [22]

Winick, E. (2018), "Every study we could find on what automation will do to jobs, in one chart", *MIT Technology Review,* 25 January, https://www.technologyreview.com/s/610005/every-study-we-could-find-on-what-automation-will-do-to-jobs-in-one-chart/. [86]

Wong, Q. (2017), "At LinkedIn, artificial intelligence is like 'oxygen'", *Mercury News,* 1 June, http://www.mercurynews.com/2017/01/06/at-linkedin-artificial-intelligence-is-like-oxygen. [94]