The ethics of artificial intelligence: Issues and initiatives

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 634.452 – March 2020

EN

The ethics of artificial intelligence: Issues and initiatives

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

AUTHORS

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

Acknowledgements

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

ADMINISTRATOR RESPONSIBLE

Mihalis Kritikos, Scientific Foresight Unit (STOA)

To contact the publisher, please e-mail stoa@ep.europa.eu

LINGUISTIC VERSION

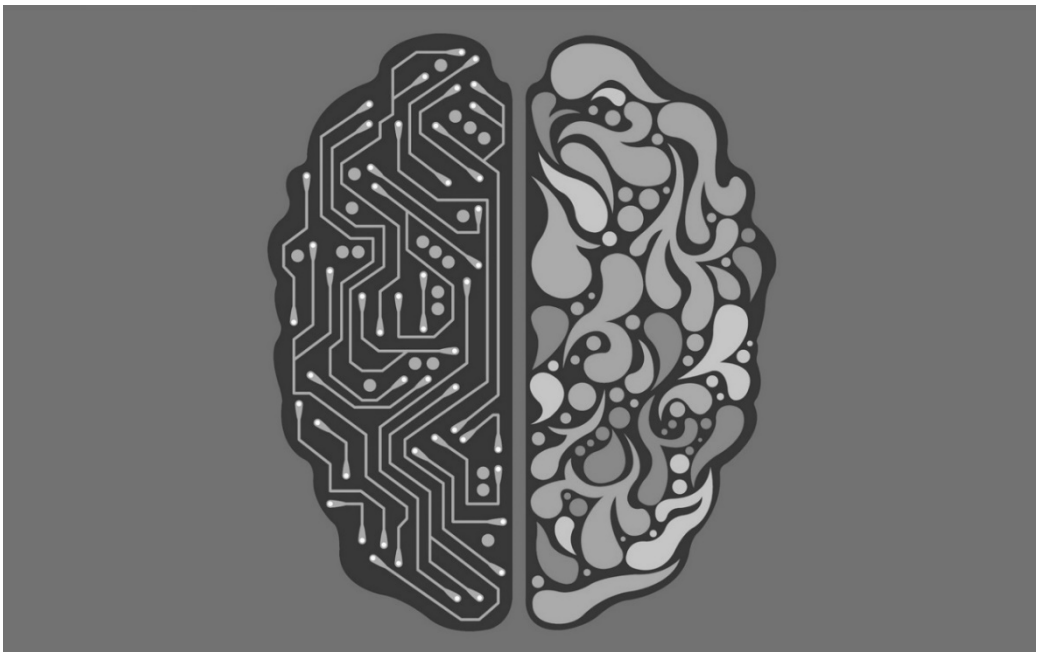
Original: EN

Manuscript completed in March 2020.

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

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

Executive summary



© Seanbatty / Pixabay

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table of contents

Table of figures

Table of tables

1. Introduction

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

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

1.1. What is AI – and what is intelligence?

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

> *'Artificial Intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.*
>
> *AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).'*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[2] From Kant's 1785 book *Groundwork of the Metaphysics of Morals*, with a variety of translations from the original German.

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

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

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

1.3. Report structure

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

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

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

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

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

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

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

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

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

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

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

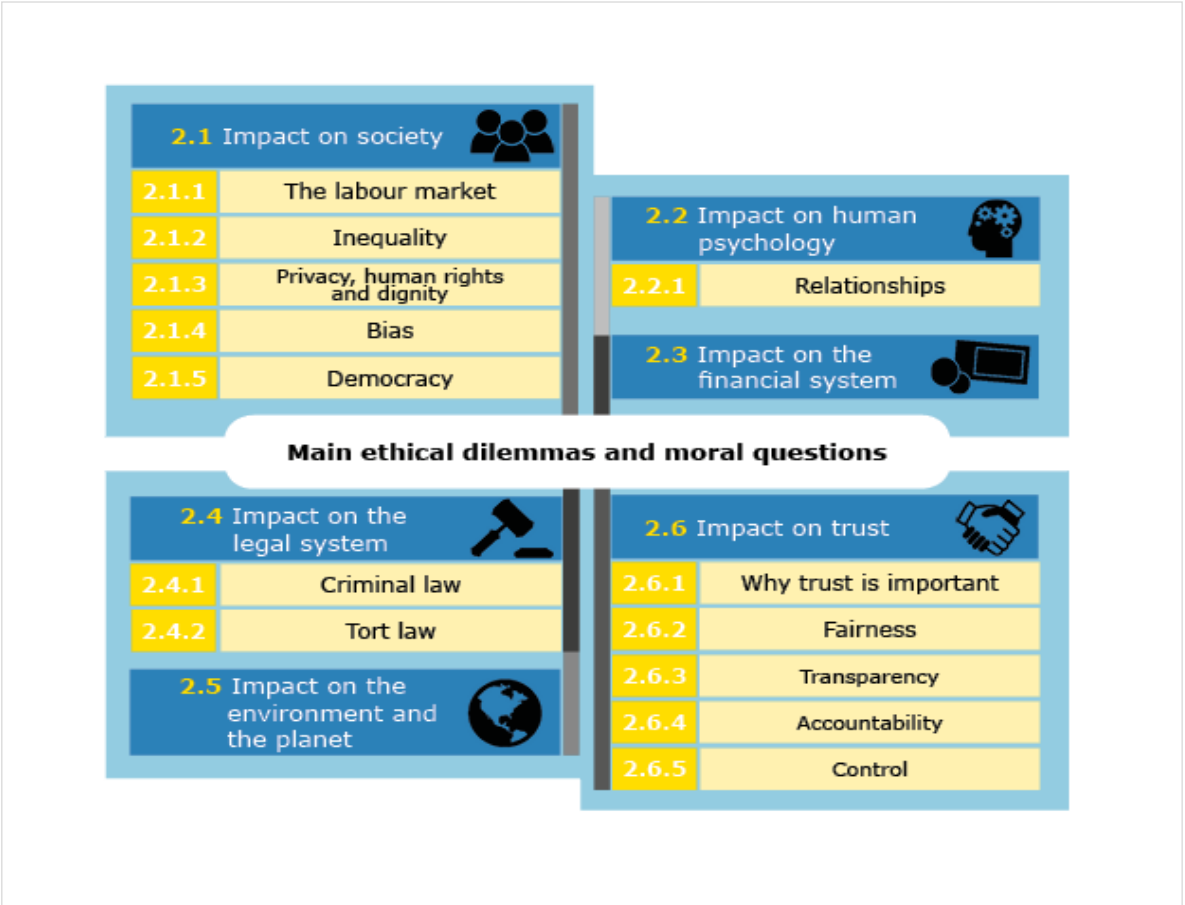
Finally, Chapter 5 explores **National and international strategies on AI**. The chapter considers what is required for a trustworthy AI and visions for the future of AI as they are articulated in national and international strategies.

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

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

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

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



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

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

2.1. Impact on society

2.1.1. The labour market

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

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

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

Impact on economic growth and productivity

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

Impact on the workforce

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Labour-market discrimination: effects on different demographics

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

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

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

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

Consequences

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

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

2.1.2. Inequality

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

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

Inequality: exploitation of workers

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

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

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

The average consumer of AI technology may never know that a person was part of the process – the value chain is opaque. One of the key ethical issues is that – given the price of the end-products – these temporary workers are being inequitably reimbursed for work that is essential to the functioning of the AI technologies. This may be especially the case where the labour force reside in countries outside the EU or US – there are growing 'data-labelling' industries in both China and Kenya, for example. Another issue is with the workers required to watch and vet offensive content for media platforms such as Facebook and YouTube (Roberts, 2016). Such content can include hate speech, violent pornography, cruelty and sometimes murder of both animals and humans. A news report (Chen, 2017) outlines mental health issues (PTSD-like trauma symptoms, panic attacks and burnout), alongside poor working conditions and ineffective counselling.
This hidden army of piecemeal workers are undertaking work that is at best extremely tedious and poorly paid, at worst, precarious, unhealthy and/or psychologically harmful. Gray's research makes the case that workers in this field still display the desire to invest in work as something more than a single payment transaction, and advises that the economic, social and psychological impacts of 'ghost work' should be dealt with systematically. Making the worker's inputs more transparent in the end-product, ensuring the value chain improves the equitable distribution of benefits, and

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

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

Sharing the benefits

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

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

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

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

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

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

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

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

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

Concentration of power among elites

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

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

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

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

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

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

Overall, Nemitz concludes that

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

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

Political instability

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

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

2.1.3. Privacy, human rights and dignity

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

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

Privacy and data rights

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

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

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

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

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

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

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

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

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

This brings up a number of ethical issues. What level of control will subjects have over the data that are collected about them? Should individuals have a right to use the model, or at least to know what it is used for, given their stake in training it? Could machine learning systems seeking patterns in data inadvertently violate people's privacy if, for example, sequencing the genome of one family member revealed health information about other members of the family?
Another ethical issue surrounds how to prevent the identity, or personal information, of an individual involved in training a model from being discovered (for example through a cyber-attack). Veale et al. (2018) argue that extra protections should be given to people whose data have been used to train models, such as the right to access models; to know where they have originated from, and to whom they are being traded or transmitted; the right to erase themselves from a trained model; and the right to express a wish that the model not be used in the future.

Human rights

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

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

Surveillance

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

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

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

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

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

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

Freedom of speech

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

2.1.4. Bias

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

Examples of AI bias

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

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

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

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

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

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

Implications

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

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

2.1.5 Democracy

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

Fake news and social media

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

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

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

[4] Created by artist Trevor Paglen and Professor Kate Crawford and New York University.

[5] https://www.vice.com/en_uk/article/xweagk/ai-face-app-imagenet-roulette

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

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

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

News bubbles and echo chambers

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

The end of democracies

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

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

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

2.2 Impact on human psychology

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

2.2.1 Relationships

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

Human-robot relationships

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

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

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

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

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

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

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

Human-human relationships

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

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

- Could robots change the beliefs, attitudes, and/or values we have about human-human relationships? People may become impatient and unwilling to put the effort into working on human-human relationships when they can have a relationship with a 'perfect' robot and avoid these challenges.

- Could 'intimate robots' lead to an increase in violent behaviour? Some researchers argue that 'sexbots' would distort people's perceptions about the value of a human being, increasing people's desire or willingness to harm others. If we are able to treat robots as instruments for sexual gratification, then we may become more likely to treat other people this way. For example, if a user repeatedly punched a companion robot, would this be unethical (Lalji, 2015)? Would violence towards robots normalise a pattern of behaviour that would eventually affect other humans? However, some argue that robots could be an outlet for sexual desire, reducing the likelihood of violence, or to help recovery from assault.

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

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

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

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

2.2.4 Personhood

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

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

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

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

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

2.3 Impact on the financial system

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

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

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

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

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

Market manipulation

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

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

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

Collusion

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

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

Accountability

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

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

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

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

2.4 Impact on the legal system

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

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

2.4.1 Criminal law

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

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

Liability

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

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

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

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

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

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

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

1. *Perpetrator-via-another*. If an offence is committed by an entity that lacks the mental capacity for *mens rea* – a child, animal, or mentally deficient person – then they are deemed an innocent agent. However, if this innocent agent was instructed by another to commit the crime, then the instructor is held criminally liable. Under this model, an AI may be held to be an innocent agent, with either the software programmer or user filling the role of perpetrator-via-another.

2. *Natural-probable-consequence*. This relates to the accomplices of a criminal action; if no conspiracy can be proven, an accomplice may still be held legally liable if the perpetrator's acts were a natural or probable consequence of a scheme encouraged or aided by an accomplice. This scenario may hold when an AI that was designed for a 'good' purpose is misappropriated and commits a crime. For example, a factory line robot may injure a nearby worker they erroneously consider a threat to their programmed mission. In this

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

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

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

Psychology

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

Commerce, financial markets and insolvency

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

Harmful or Dangerous Drugs

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

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

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

Offences Against the Person

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

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

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

Sexual Offences

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

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

Who is responsible?

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

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

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

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

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

Theft and fraud, and forgery and impersonation

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

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

2.4.2 Tort law

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

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

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

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

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

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

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

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

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

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

2.5 Impact on the environment and the planet

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

2.5.1 Use of natural resources

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

2.5.2 Pollution and waste

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

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

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

2.5.3 Energy concerns

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

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

2.5.4 Ways AI could help the planet

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

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

2.6 Impact on trust

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

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

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

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

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

2.6.1 Why trust is important

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

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

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

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

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

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

2.6.2 Fairness

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

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

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

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

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

1. Statistical parity - where an equal proportion of defendants are detained in each race group. For example, white and black defendants are detained at equal rates.

2. Conditional statistical parity - where controlling for a limited set of 'legitimate' risk factors, an equal proportion of defendants are detained within each race group. For example, among defendants who have the same number of prior convictions, black and white defendants are detained at equal rates.

3. Predictive equality - where the accuracy of decisions is equal across race groups, as measured by false positive rate. This means that among defendants who would not have gone on to commit a violent crime if released, detention rates are equal across race groups.

4. Calibration - among defendants with a given risk score, the proportion who reoffend is the same across race groups.

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

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

2.6.3 Transparency

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

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

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

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

Black boxes

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

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

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

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

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

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

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

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

Explainable systems

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

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

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

Intentional understanding

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

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

Algorithm auditors

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

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

2.6.4 Accountability

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

Another method of ensuring trust of AI is through accountability. As discussed, accountability ensures that if an AI makes a mistake or harms someone, there is someone that can be held responsible, whether that be the designer, the developer or the corporation selling the AI. In the event of damages incurred, there must be a mechanism for redress so that victims can be sufficiently compensated.

A growing body of literature has begun to address concepts such as algorithmic accountability and responsible AI. Algorithmic accountability, according to Caplan et al. (2018), deals with the delegation of responsibility for damages incurred as a result of algorithmically based decisions producing discriminatory or unfair consequences. One area where accountability is likely to be important is the introduction of self-driving vehicles. In the event of an accident, who should be held accountable? A number of fatal accidents have already occurred with self-driving cars, for example in 2016, a Tesla Model S equipped with radar and cameras determined that a nearby lorry was in fact the sky, which resulted in a fatal accident. In March 2018, a car used by Uber in self-driving vehicle trials hit and killed a woman in Arizona, USA. Even if autonomous cars are safer than vehicles driven by humans, accidents like these undermine trust.

Regulation

One way of ensuring accountability is regulation. Winfield and Jirotka (2018) point out that technology is, in general, trusted if it brings benefits and is safe and well regulated. Their paper argues that one key element in building trust in AI is ethical governance – a set of processes, procedures, cultures and values designed to ensure the highest standards of behaviour. These standards of behaviour need to be adopted by individual designers and the organisations in which they work, so that ethical issues are dealt with as or before they arise in a principled manner, rather than waiting until a problem surfaces and dealing with it in an ad-hoc way.

They give the example of airliners, which are trusted because we know that they are part of a highly regulated industry with an outstanding safety record. The reason commercial aircraft are so safe is not just good design, it is also the tough safety certification processes, and the fact that when things do go wrong, there are robust and publicly visible processes of air accident investigation.

Winfield and Jirotka (2018) suggest that some robot types, driverless cars for instance, should be regulated through a body similar to the Civil Aviation Authority (CAA), with a driverless car equivalent of the Air Accident Investigation Branch.

When it comes to public perception of robots and advanced technology, regulation and management crops up as a prominent concern. In two surveys of citizens across the EU (European Commission 2012; European Commission, 2012), both showed that there was a generally positive view of robots and digitisation as long as this is implemented and managed carefully. In fact,

between 88% and 91% of those surveyed declared that robots and advanced technology must be managed carefully, one of the strongest results in either survey — reflecting a strong concern and area of priority amongst EU citizens.

2.6.5 Control

Another issue which affects public trust of AI is control. Much of this relates to fears around the idea of 'Superintelligence' - that as artificial intelligence increases to the point that it surpasses human abilities, it may come to take control over our resources and outcompete our species, leading to human extinction. A related fear is that, even if an AI agent was carefully designed to have goals aligned with human needs, it might develop for itself unanticipated subgoals that are not. For example, Bryson (2019) gives the example of a chess-playing robot taught to improve its game. This robot inadvertently learns to shoot people that switch it off at night, depriving it of vital resources. However, while most researchers agree this threat is unlikely to occur, to maintain trust in AI, it is important that humans have ultimate oversight over this technology.

Human in the loop

One idea that has been suggested by researchers is that of always keeping a human-in-the-loop (HITL). Here a human operator would be a crucial component of the automated control process, supervising the robots. A simple form of HITL already in existence is the use of human workers to label data for training machine learning algorithms. For example when you mark an email as 'spam', you are one of many humans in the loop of a complex machine learning algorithm, helping it in its continuous quest to improve email classification as spam or non-spam.

However HITL can also be a powerful tool for regulating the behaviour of AI systems. For instance, many researchers argue that human operators should be able to monitor the behaviour of LAWS, or 'killer robots,' or credit scoring algorithms (Citron and Pasquale 2014). The presence of a human fulfils two major functions in a HITL AI system (Rahwan, 2018):

> 1. The human can identify misbehaviour by an otherwise autonomous system, and take corrective action. For instance, a credit scoring system may misclassify an adult as ineligible for credit because their age was incorrectly input—something a human may spot from the applicant's photograph. Similarly, a computer vision system on a weaponised drone may mis-identify a civilian as a combatant, and the human operator—it is hoped—would override the system.
>
> 2. Keeping humans in the loop would also provide accountability - if an autonomous system causes harm to human beings, having a human in the loop provides trust that somebody would bare the consequence of such mistakes. According to Rahwan (2018), until we find a way to punish algorithms for harm to humans, 'it is hard to think of any other alternative'.

However, although HITL is useful for building AI systems that are subject to oversight, it may not be enough. AI machines that make decisions with wider societal implications, such as algorithms that control millions of self-driving cars or news filtering algorithms that influence the political beliefs and preferences of millions of citizens, should be subject to oversight by society as a whole, requiring a 'society-in-the-loop' paradigm (Rahwan, 2018).

The big red button

As a way to address some of the threats of artificial intelligence, researchers have proposed ways to stop an AI system before it has a chance to escape outside control and cause harm. A so-called 'big red button', or 'kill switch' would enable human operators to interrupt or divert a system, while preventing the system from learning that such an intervention is a threat. However, some

commentators fear that a sufficiently advanced AI machine could anticipate this move and defend itself by learning to disable its own 'kill switch'.

The red button raises wider practical questions about shutting down AI systems in order to keep them safe. What is the best way to accomplish that, and for what specific kinds of AI systems?

Orseau and Armstrong (2016) recently published a paper about how to prevent AI programmed through reinforcement learning (RL) from seeing interruptions as a threat. For example, an algorithm trying to optimise its chess performance may learn to disable its off switch so that it can spend more time learning how to play chess. Or it may learn to harm people who tried to switch it off, etc. What the researchers propose is to steer certain variants of reinforcement learning away from learning to avoid or impede an interruption. In this way, the authors argue, a system can pursue an optimal policy that is also interruptible. By being 'safely interruptible,' the paper concludes, reinforcement learning will not undermine the means of responsible oversight and intervention.

Riedl and Harrison (2017) suggests making a 'big red button' that, once pressed, diverted the AI into a simulated world where it could pursue its reward functions without causing any harm. Alternatively another idea is to maintain system uncertainty about key reward functions, which would prevent AI from attaching value to disabling an off-switch (Hadfield-Menell et al., 2016).

However Arnold and Schultz (2018) argue that the 'red button' approach comes at the point when a system has already 'gone rogue' and seeks to obstruct interference, and that 'big red button' approaches focus on long-term threats, imagining systems considerably more advanced than exist today and neglecting the present day problems with keeping automated systems accountable. A better approach, according to Arnold and Scheutz, would be to make ongoing self-evaluation and testing an integral part of a system's operation, in order to diagnose how the system is performing, and correct any errors.

They argue that to achieve this AIs should contain an ethical core (EC) consisting of a scenario-generation mechanism and a simulation environment used to test a system's decisions in simulated worlds, rather than the real world. This EC would be kept hidden from the system itself, so that the system's algorithms would be prevented from learning about its operation and its function, and ultimately its presence. Through continual testing in the simulated world, the EC would monitor and check for deviant behaviour - providing a far more effective and vigilant response than an emergency button which one might not get to push in time.

3. Ethical initiatives in the field of artificial intelligence

As detailed in previous sections, there are myriad ethical considerations accompanying the development, use and effects of artificial intelligence (AI). These range from the potential effects AI could have on the fundamental human rights of citizens within a society to the security and utilisation of gathered data; from the bias and discrimination unintentionally embedded into an AI by a homogenous group of developers, to a lack of public awareness and understanding about the consequences of their choices and usage of any given AI, leading to ill-informed decisions and subsequent harm.

AI builds upon previous revolutions in ICT and computing and, as such, will face a number of similar ethical problems. While technology may be used for good, potentially it may be misused. We may excessively anthropomorphise and humanise AI, blurring the lines between human and machine. The ongoing development of AI will bring about a new 'digital divide', with technology benefiting some socioeconomic and geographic groups more than others. Further, AI will have an impact on our biosphere and environment that is yet to be qualified (Veruggio and Operto, 2006).

3.1. International ethical initiatives

While official regulation remains scarce, many independent initiatives have been launched internationally to explore these – and other – ethical quandaries. The initiatives explored in this section are outlined in Table 3.1 and will be studied in light of the associated harms and concerns they aim to understand and mitigate.

Table 1: Ethical initiatives and harms addressed

Initiative	Location	Key issues tackled	Publications	Sources of funding
The Institute for Ethics in Artificial Intelligence	Germany	Human-centric engineering and a focus on the cultural and social anchoring of rapid advances in AI, covering disciplines including philosophy, ethics, sociology, and political science.		Initial (2019) funding grant from Facebook ($7.5 million over five years).
The Institute for Ethical AI & Machine Learning	United Kingdom	The Institute aims to empower all from individuals to entire nations to develop AI, based on eight principles for responsible machine learning: these concern the maintenance of human control, appropriate redress for AI impact, evaluation of bias, explicability, transparency, reproducibility, mitigation of the effect of AI automation on workers, accuracy, cost, privacy, trust, and security.		unknown
The Institute for Ethical Artificial Intelligence in Education	United Kingdom	The potential threats to young people and education of the rapid growth of new AI technology, and ensuring the ethical development of AI-led EdTech.		unknown
The Future of Life Institute	United States	Ensuring that the development of AI is beneficial to humankind, with a focus on safety and existential risk: autonomous weapons arms race, human control of AI, and the potential dangers of advanced 'general/strong' or super-intelligent AI.	'***Asilomar AI Principles***'	Private. Top donors: Elon Musk (SpaceX and Tesla), Jaan Tallinn (Skype), Matt Wage (financial trader), Nisan Stiennon (software engineer), Sam Harris, George Godula (tech entrepreneur), and Jacob Trefethen (Harvard).
The Association for Computing Machinery	United States	The transparency, usability, security, accessibility, accountability, and digital inclusiveness of computers and networks, in terms of research, development, and implementation.	Statements on: algorithmic transparency and accountability (January 2017), computing and network security (May 2017), the Internet of Things (June 2017), accessibility, usability, and digital inclusiveness (September 2017),	unknown

			and mandatory access to information infrastructure for law enforcement (April 2018).	
The Japanese Society for Artificial Intelligence (JSAI)	Japan	To ensure that AI R&D remains beneficial to human society, and that development and research is conducted ethically and morally.	'***Ethical Guidelines***'	unknown
AI4All	United States	Diversity and inclusion in AI, to expose underrepresented groups to AI for social good and humanity's benefit.		Google
The Future Society	United States	The impact and governance of artificial intelligence to broadly benefit society, spanning policy research, advisory and collective intelligence, coordination of governance, law, and education.	'***Draft Principles for the Governance of AI'*** Published October 2017 (later published on their website on 7th February 2019),	unknown
The AI Now Institute	United States	The social implications of AI, especially in the areas of: Rights and liberties, labour and automation, bias and inclusion, and safety and critical infrastructure.		Various organisations, including Luminate, the MacArthur Foundation, Microsoft Research, Google, the Ford Foundation, DeepMind Ethics & Society, and the Ethics & Governance of AI Initiative.
The Institute of Electrical and Electronics Engineers (IEEE)	United States	Societal and policy guidelines to keep AI and intelligent systems human-centric, and serving humanity's values and principles. Focuses on ensuring that all stakeholders – across design and development – are educated, trained, and empowered to prioritise the ethical considerations of human rights, well-being, accountability, transparency, and awareness of misuse.	'***Ethically Aligned Design'*** **First Edition** (March 2019)	
The Partnership on AI	United States	Best practices on AI technologies: Safety, fairness, accountability, transparency, labour and the economy, collaboration between people and systems, social and societal influences, and social good.		The Partnership was formed by a group of AI researchers representing six of the world's largest tech companies: Apple,

				Amazon, DeepMind and Google, Facebook, IBM, and Microsoft.
The Foundation for Responsible Robotics	The Netherlands	Responsible robotics (in terms of design, development, use, regulation, and implementation). Proactively taking stock of the issues that accompany technological innovation, and the impact these will have on societal values such as safety, security, privacy, and well-being.		unknown
AI4People	Belgium	The social impacts of AI, and the founding principles, policies, and practices upon which to build a 'good AI society'.	'***Ethical Framework for a Good AI Society***'	Atomium— European Institute for Science, Media and Democracy. Some funding was provided to the project's Scientific Committee Chair from the Engineering and Physical Sciences Research Council.
The Ethics and Governance of Artificial Intelligence Initiative	United States	Seeks to ensure that technologies of automation and machine learning are researched, developed, and deployed in a way which vindicate social values of fairness, human autonomy, and justice.		The Harvard Berkman Klein Center and the MIT Media Lab. Supported by The Miami Foundation (fiscal sponsorship), Knight Foundation, Luminate, Red Hoffman, and the William and Flora Hewlett Foundation.
Saidot: Enabling responsible AI ecosystems	Finland	Helping companies, governments, and organisations develop and deploy responsible AI ecosystems, to deliver transparent, accountable, trustworthy AI services. Enabling organisations to develop human-centric AI, with a focus on increasing the levels of trust and accountability in AI ecosystems. The platform offers software and algorithmic systems that can 'validate [an] intelligence system's trustworthiness' (Saidot, 2019)		
euRobotics	Europe	Maintaining and extending European talent and progress in robotics – AI industrialisation and economic impact.		European Commission

The Centre for Data Ethics and Innovation	UK	Identifying and plugging gaps in our regulatory landscape, AI use of data, and maximising the benefits of AI to society.		UK Government
Special Interest Group on Artificial Intelligence (SIGAI), The Association for Computing Machinery	United States	. Promoting and supporting the growth and application of AI principles and techniques throughout computing, and promoting AI education and publications through various forums		The Association for Computing Machinery
	Other key international developments: current and historical			
The Montréal Declaration	Canada	The socially responsible development of AI, bringing together 400 participants across all sectors of society to identify the ethical and moral challenges in the short and long term. Key values: well-being, autonomy, justice, privacy, knowledge, democracy, and accountability.		Université de Montréal with the support of the Fonds de recherche en santé du Québec and the Palais des congrès de Montréal.
The UNI Global Union	Switzerland	Worker disruption and transparency in the application of AI, robotics, and data and machine learning in the workplace. Safeguarding workers' interests and maintaining human control and a healthy power balance.	'***Top 10 Principles for Ethical AI***'	unknown
The European Robotics Research Network (EURON)	Europe (Coordinator based in Sweden)	Research co-ordination, education and training, publishing and meetings, industrial links and international links in robotics.	'***Roboethics Roadmap***'	European Commission (2000-2004)
The European Robotics Platform (EUROP)	Europe	Bringing European robotics and AI community together. Industry-driven, focus on competitiveness and innovation.		European Commission

3.2. Ethical harms and concerns tackled by these initiatives

All of the initiatives listed above agree that AI should be researched, developed, designed, deployed, monitored, and used in an ethical manner – but each has different areas of priority. This section will include analysis and grouping of the initiatives above, by type of issues they aim to address, and then outline some of the proposed approaches and solutions to protect from harms.

A number of key issues emerge from the initiatives, which **can be broadly split into the following categories:**

1. Human rights and well-being
   Is AI in the best interests of humanity and human well-being?

2. Emotional harm
   Will AI degrade the integrity of the human emotional experience, or facilitate emotional or mental harm?

3. Accountability and responsibility
   Who is responsible for AI, and who will be held accountable for its actions?

4. Security, privacy, accessibility, and transparency
   How do we balance accessibility and transparency with privacy and security, especially when it comes to data and personalisation?

5. Safety and trust
   What if AI is deemed untrustworthy by the public, or acts in ways that threaten the safety of either itself or others?

6. Social harm and social justice
   How do we ensure that AI is inclusive, free of bias and discrimination, and aligned with public morals and ethics?

7. Financial harm
   How will we control for AI that negatively affects economic opportunity and employment, and either takes jobs from human workers or decreases the opportunity and quality of these jobs?

8. Lawfulness and justice
   How do we go about ensuring that AI - and the data it collects - is used, processed, and managed in a way that is just, equitable, and lawful, and subject to appropriate governance and regulation? What would such regulation look like? Should AI be granted 'personhood'?

9. Control and the ethical use – or misuse – of AI
   How might AI be used unethically - and how can we protect against this? How do we ensure that AI remains under complete human control, even as it develops and 'learns'?

10. Environmental harm and sustainability
    How do we protect against the potential environmental harm associated with the development and use of AI? How do we produce it in a sustainable way?

11. Informed use
    What must we do to ensure that the public is aware, educated, and informed about their use of

and interaction with AI?

12. Existential risk
    How do we avoid an AI arms race, pre-emptively mitigate and regulate potential harm, and ensure that advanced machine learning is both progressive and manageable?

Overall, these initiatives all aim to identify and form ethical frameworks and systems that establish human beneficence at the highest levels, prioritise benefit to both human society and the environment (without these two goals being placed at odds), and mitigate the risks and negative impacts associated with AI — with a focus on ensuring that AI is accountable and transparent (IEEE, 2019).

The IEEE's '***Ethically Aligned Design: A Vision for Prioritising Human Well-being with Autonomous and Intelligent Systems***' (v1; 2019) is one of the most substantial documents published to date on the ethical issues that AI may raise — and the various proposed means of mitigating these.

Figure 2: General principles for the ethical and values-based design, development, and implementation of autonomous and intelligent systems (as defined by the IEEE's *Ethically Aligned Design* First Edition March 2019)

Areas of key impact comprise sustainable development; personal data rights and agency over digital identity; legal frameworks for accountability; and policies for education and awareness. They fall under **the three pillars of the Ethically Aligned Design conceptual framework:** Universal human values; political self-determination and data agency; and technical dependability.

3.2.1 Harms in detail

Taking each of these harms in turn, this section explores how they are being conceptualised by initiatives and some of the challenges that remain.

Human rights and well-being

All initiatives adhere to the view that ***AI must not impinge on basic and fundamental human rights***, such as human dignity, security, privacy, freedom of expression and information, protection of personal data, equality, solidarity and justice (European Parliament, Council and Commission, 2012).

How do we ensure that AI upholds such fundamental human rights and prioritises human well-being? Or that AI does not disproportionately affect vulnerable areas of society, such as children, those with disabilities, or the elderly, or reduce quality of life across society?

In order to ensure that human rights are protected, the IEEE recommends new governance frameworks, standards, and regulatory bodies which oversee the use of AI; translating existing legal obligations into informed policy, allowing for cultural norms and legal frameworks; and always maintaining complete human control over AI, without granting them rights or privileges equal to those of humans (IEEE, 2019). To safeguard human well-being, defined as 'human satisfaction with life and the conditions of life, as well as an appropriate balance between positive and negative affect' (i*bid*), the IEEE suggest prioritising human well-being throughout the design phase, and using the best and most widely-accepted available metrics to clearly measure the societal success of an AI.

There are crossovers with accountability and transparency: there must always be appropriate ways to identify and trace the impingement of rights, and to offer appropriate redress and reform. Personal data are also a key issue here; AI collect all manner of personal data, and users must retain the access to, and control of, their data, to ensure that their fundamental rights are being lawfully upheld (IEEE, 2019).

According to the ***Foundation for Responsible Robotics***, AI must be ethically developed with human rights in mind to achieve their goal of 'responsible robotics', which relies upon proactive innovation to uphold societal values like safety, security, privacy, and well-being. The Foundation engages with policymakers, organises and hosts events, publishes consultation documents to educate policymakers and the public, and creates public-private collaborations to bridge the gap between industry and consumers, to create greater transparency. It calls for ethical decision-making right from the research and development phase, greater consumer education, and responsible law- and policymaking – made before AI is released and put into use.

The ***Future of Life Institute*** defines a number of principles, ethics, and values for consideration in the development of AI, including the need to design and operate AI in a way that is compatible with the ideals of human dignity, rights, freedoms, and cultural diversity[7]. This is echoed by the ***Japanese Society for AI Ethical Guidelines***, which places the utmost importance on AI being realised in a way that is beneficial to humanity, and in line with the ethics, conscience, and competence of both its researchers and society as a whole. AI must contribute to the peace, safety, welfare, and public interest of society, says the Society, and protect human rights.

The Future Society's Law and Society Initiative emphasises that human beings are equal in rights, dignity, and freedom to flourish, and are entitled to their human rights.[8] With this in mind, to what extent should we delegate to machines decisions that affect people? For example, could AI 'judges' in the legal profession be more efficient, equitable, uniform, and cost-saving than human ones –

[7] https://futureoflife.org/ai-principles/
[8] http://thefuturesociety.org/law-and-society-initiative

and even if they were, would this be an appropriate way to deploy AI? ***The Montréal Declaration***[9] aims to clarify this somewhat, by pulling together an ethical framework that promotes internationally recognised human rights in fields affected by the rollout of AI: 'The principles of the current declaration rest on the common belief that human beings seek to grow as social beings endowed with sensations, thoughts and feelings, and strive to fulfil their potential by freely exercising their emotional, moral and intellectual capacities.' In other words, AI must not only not disrupt human well-being, but it must also proactively encourage and support it to improve and grow.

Some approach AI from a more specific viewpoint – such as the ***UNI Global Union***, which strives to protect an individual's right to work. Over half of the work currently done by people could be done faster and more efficiently in an automated way, says the Union. This identifies a prominent harm that AI may cause in the realm of human employment. The Union states that we must ensure that AI serves people and the planet, and both protects and increases fundamental human rights, human dignity, integrity, freedom, privacy, and cultural and gender diversity[10].

Emotional harm

What is it to be human? AI will interact with and have an impact on the human emotional experience in ways that have not yet been qualified; humans are susceptible to emotional influence both positively and negatively, and ***'affect' – how emotion and desire influence behaviour – is a core part of intelligence***. Affect varies across cultures, and, given different cultural sensitivities and ways of interacting, affective and influential AI could begin to influence how people view society itself. The ***IEEE*** recommend various ways to mitigate this risk, including the ability to adapt and update AI norms and values according to who they are engaging with, and the sensitivities of the culture in which they are operating.

There are various ways in which AI could inflict emotional harm, including false intimacy, over-attachment, objectification and commodification of the body, and social or sexual isolation. These are covered by various of the aforementioned ethical initiatives, including **the Foundation for Responsible Robotics, Partnership on AI, the AI Now** institute (especially regarding affect computing), **the Montréal Declaration**, and the **European Robotics Research Network (EURON) Roadmap** (for example, their section on the risks of humanoids).

These possible harms come to the fore when considering the development of an intimate relationship with an AI, for example in the sex industry. Intimate systems, as the ***IEEE*** call them, must not contribute to sexism, racial inequality, or negative body image stereotypes; must be for positive and therapeutic use; must avoid sexual or psychological manipulation of users without consent; should not be designed in a way that contributes to user isolation from human companionship; must be designed in a way that is transparent about the effect they may have on human relationship dynamics and jealousy; must not foster deviant or criminal behaviour, or normalise illegal sexual practices such as paedophilia or rape; and must not be marketed commercially as a person (in a legal sense or otherwise).

Affective AI is also open to the possibility of deceiving and coercing its users – researchers have defined the act of AI subtly modifying behaviour as '***nudging***', when an AI emotionally manipulates and influences its user through the affective system. While this may be useful in some ways – drug dependency, healthy eating – it could also trigger behaviours that worsen human health. Systematic analyses must examine the ethics of affective design prior to deployment; users must be educated on how to recognise and distinguish between nudges; users must have an opt-in system for autonomous nudging systems; and vulnerable populations that cannot give informed consent, such

[9] https://www.montrealdeclaration-responsibleai.com/the-declaration
[10] http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

as children, must be subject to additional protection. In general, stakeholders must discuss the question of whether or not the nudging design pathway for AI, which lends itself well to selfish or detrimental uses, is an ethical one to pursue (IEEE, 2019).

As raised by the ***IEEE*** (2019), nudging may be used by governments and other entities to influence public behaviour. Would it be ethically appropriate for a robot to use nudging to encourage, for example, charitable behaviour or donations? We must pursue full transparency regarding the beneficiaries of such behaviour, say the IEEE, due to the potential for misuse.

Other issues include technology addiction and emotional harm due to societal or gender bias.

Accountability and responsibility

The vast majority of initiatives mandate that AI must be ***auditable***, in order to assure that the designers, manufacturers, owners, and operators of AI are held accountable for the technology or system's actions, and are thus considered responsible for any potential harm it might cause. According to the ***IEEE***, this could be achieved by the courts clarifying issues of culpability and liability during the development and deployment phases where possible, so that those involved understand their obligations and rights; by designers and developers taking into account the diversity of existing cultural norms among various user groups; by establishing multi-stakeholder ecosystems to create norms that currently do not exist, given that AI-oriented technology is too new; and by creating registration and record-keeping systems so that it is always possible to trace who is legally responsible for a particular AI.

Sex and Robots

In July of 2017, the ***Foundation for Responsible Robotics*** published a report on 'Our Sexual Future with Robots' (Foundation for Responsible Robotics, 2019). This aimed to present an objective summary of the various issues and opinions surrounding our intimate association with technology. Many countries are developing robots for sexual gratification; these largely tend to be pornographic representations of the human body – and are mostly female. These representations, when accompanied by human anthropomorphism, may cause robots to be perceived as somewhere between living and inanimate, especially when sexual gratification is combined with elements of intimacy, companionship and conversation. Robots may also affect societal perceptions of gender or body stereotypes, erode human connection and intimacy and lead to greater social isolation. However, there is also some potential for robots to be of emotional sexual benefit to humans, for example by helping to reduce sex crime, and to rehabilitate victims of rape or sexual abuse via inclusion in healing therapies.

The ***Future of Life Institute*** tackles the issue of accountability via its **Asilomar Principles**, a list of 23 guiding principles for AI to follow in order to be ethical in the short and long term. Designers and builders of advanced AI systems are 'stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications' (FLI, 2017); if an AI should make a mistake, it should also be possible to ascertain why. The ***Partnership on AI*** also stresses the importance of accountability in terms of bias. We should be sensitive to the fact that assumptions and biases exist within data and thus within systems built from these data, and strive not to replicate them – i.e. to be actively accountable for building fair, bias-free AI.

All other initiatives highlight the importance of accountability and responsibility – both by designers and AI engineers, and by regulation, law and society on a larger scale.

Access and transparency vs. security and privacy

A main concern over AI is its ***transparency***, explicability, security, reproducibility, and interpretability: is it possible to discover why and how a system made a specific decision, or why and how a robot acted in the way it did? This is especially pressing in the case of *safety-critical* systems that may have direct consequences for physical harm: driverless cars, for example, or medical diagnosis systems. Without transparency, users may struggle to understand the systems they are using – and their associated consequences – and it will be difficult to hold the relevant persons accountable and responsible.

To address this, the ***IEEE*** propose developing new standards that detail measurable and testable levels of transparency, so systems can be objectively assessed for their compliance. This will likely take different forms for different stakeholders; a robot user may require a 'why-did-you-do-that' button, while a certification agency or accident investigator will require access to relevant algorithms in the form of an 'ethical black box' which provides failure transparency (IEEE, 2019).

Autonomy and agent vs. patient

The current approach to AI is undeniably anthropocentric. This raises **possible issues around the distinction between moral agents and moral patients, between artificial and natural, between self-organising and not**. AI cannot become autonomous in the same way that living beings are considered autonomous (IEEE, 2019), but how do we define autonomy in terms of AI? Machine autonomy designates how machines act and operate according to regulation, but any attempts to implant emotion and morality into AI 'blur the distinction between agents and patients and may encourage anthropomorphic expectations of machines', writes the ***IEEE*** — especially as embodied AI begins to look increasingly similar to humans. Establishing a usable distinction between human and system/machine autonomy involves questions of free will, being/becoming and predetermination. It is clear that further discussion is needed to clarify what 'autonomy' may mean in terms of artificial intelligence and systems.

AI require data to continually learn and develop their automatic decision-making. These data are personal and may be used to identify a particular individual's physical, digital, or virtual identity (i.e. personally identifiable information, PII). 'As a result,' write the IEEE (2017), 'through every digital transaction (explicit or observed) humans are generating a unique digital shadow of their physical self'. To what extent can humans realise the right to keep certain information private, or have input into how these data are used? Individuals may lack the appropriate tools to control and cultivate their unique identity and manage the associated ethical implications of the use of their data. Without clarity and education, many users of AI will remain unaware of the digital footprint they are creating, and the information they are putting out into the world. Systems must be put in place for users to control, interact with and access their data, and give them agency over their digital personas.

PII has been established as the asset of the individual (by Regulation (EU) 2016/679 in Europe, for example), and systems must ask for explicit consent at the time data are collected and used, in order to protect individual autonomy, dignity and right to consent. The IEEE mention the possibility of a personalised 'privacy AI or algorithmic agent or guardian' to help individuals curate and control their personal data and foresee and mitigate potential ethical implications of machine learning data exchange.

The ***Future of Life Institute's Asilomar Principles*** agree with the IEEE on the importance of transparency and privacy across various aspects: failure transparency (if an AI fails, it must be possible to figure out why), judicial transparency (any AI involved in judicial decision-making must provide a satisfactory explanation to a human), personal privacy (people must have the right to access, manage, and control the data AI gather and create), and liberty and privacy (AI must not unreasonably curtail people's real or perceived liberties). ***Saidot*** takes a slightly wider approach and strongly emphasises the importance of AI that are transparent, accountable, and trustworthy, where

people, organisations, and smart systems are openly connected and collaborative in order to foster cooperation, progress, and innovation.

All of the initiatives surveyed identify transparency and accountability of AI as an important issue. This balance underpins many other concerns – such as legal and judicial fairness, worker compensation and rights, security of data and systems, public trust, and social harm.

Safety and trust

Where AI is used to supplement or replace human decision-making, there is consensus that it must be ***safe, trustworthy, and reliable, and act with integrity***.

The ***IEEE*** propose cultivating a 'safety mindset' among researchers, to 'identify and pre-empt unintended and unanticipated behaviors in their systems' and to develop systems which are 'safe by design'; setting up review boards at institutions as a resource and means of evaluating projects and their progress; encouraging a community of sharing, to spread the word on safety-related developments, research, and tools. The ***Future of Life Institute's Asilomar principles*** indicate that all involved in developing and deploying AI should be mission-led, adopting the norm that AI 'should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organisation' (Future of Life Institute, 2017). This approach would build public trust in AI, something that is key to its successful integration into society.

An 'ethical black box'

Initiatives including the ***UNI Global Union*** and ***IEEE*** suggest equipping AI systems with an 'ethical black box': a device that can record information about said system to ensure its accountability and transparency, but that also includes clear data on the ethical consideration built into the system from the beginning (UNI Global Union, n.d.).

The Japanese Society for AI proposes that AI should act with integrity at all times, and that AI and society should earnestly seek to learn from and communicate with one another. 'Consistent and effective communication' will strengthen mutual understanding, says the Society, and '[contribute] to the overall peace and happiness of mankind' (JSAI, 2017). The ***Partnership on AI*** agrees, and strives to ensure AI is trustworthy and to create a culture of cooperation, trust, and openness among AI scientists and engineers. The ***Institute for Ethical AI & Machine Learning*** also emphasises the importance of dialogue; it ties together the issues of trust and privacy in its eight core tenets, mandating that AI technologists communicate with stakeholders about the processes and data involved to build trust and spread understanding throughout society.

Social harm and social justice: inclusivity, bias, and discrimination

AI development requires ***a diversity of viewpoints***. There are several organisations establishing that these must be in line with community viewpoints and align with social norms, values, ethics, and preferences, that biases and assumptions must not be built into data or systems, and that AI should be aligned with public values, goals, and behaviours, respecting cultural diversity. Initiatives also argue that all should have access to the benefits of AI, and it should work for the common good. In other words, developers and implementers of AI have a social responsibility to embed the right values into AI and ensure that they do not cause or exacerbate any existing or future harm to any part of society.

The ***IEEE*** suggest first identifying social and moral norms of the specific community in which an AI will be deployed, and those around the specific task or service it will offer; designing AI with the idea of 'norm updating' in mind, given that norms are not static and AI must change dynamically and transparently alongside culture; and identifying the ways in which people resolve norm conflicts, and equipping AI with a system in which to do so in a similar and transparent way. This should be done collaboratively and across diverse research efforts, with care taken to evaluate and assess potential biases that disadvantage specific social groups.

Several initiatives – such as ***AI4All*** and the ***AI Now Institute*** – explicitly advocate for fair, diverse, equitable, and non-discriminatory inclusion in AI at all stages, with a focus on support for under-represented groups. Currently, AI-related degree programmes do not equip aspiring developers and designers with an appropriate knowledge of ethics (IEEE, 2017), and corporate environments and business practices are not ethically empowering, with a lack of roles for senior ethicists that can steer and support value-based innovation.

On a global scale, the inequality gap between developed and developing nations is significant. While AI may have considerable usefulness in a humanitarian sense, they must not widen this gap or exacerbate poverty, illiteracy, gender and ethnic inequality, or disproportionately disrupt employment and labour. The IEEE suggests taking action and investing to mitigate the inequality gap; integrating corporate social responsibility (CSR) into development and marketing; developing transparent power structures; facilitating and sharing robotics and AI knowledge and research; and generally keeping AI in line with the US Sustainable Development Goals[11]. AI technology should be made equally available worldwide via global standardisation and open-source software, and interdisciplinary discussion should be held on effective AI education and training (IEEE, 2019).

A set of ethical guidelines published by the ***Japanese Society for AI*** emphasises, among other considerations, the importance of a) contribution to humanity, and b) social responsibility. AI must act in the public interest, respect cultural diversity, and always be used in a fair and equal manner.

The ***Foundation for Responsible Robotics*** includes a Commitment to Diversity in its push for responsible AI; the ***Partnership on AI*** cautions about the 'serious blind spots' of ignoring the presence of biases and assumptions hidden within data; ***Saidot*** aims to ensure that, although our social values are now 'increasingly mediated by algorithms', AI remains human-centric (Saidot, 2019); the ***Future of Life Institute*** highlights a need for AI imbued with human values of cultural diversity and human rights; and the ***Institute for Ethical AI & Machine Learning*** includes 'bias evaluation' for monitoring bias in AI development and production. The dangers of human bias and assumption are a frequently identified risk that will accompany the ongoing development of AI.

Financial harm: Economic opportunity and employment

AI may disrupt the economy and lead to loss of jobs or work disruption for many humans, and will have an impact on workers' rights and displacement strategy as many strains of work become automated (and vanish in related business change).

Additionally, rather than just focusing on the number of jobs lost or gained, traditional employment structures will need to be changed to mitigate the effects of automation and take into account the complexities of employment. Technological change is happening too fast for the traditional workforce to keep pace without retraining. Workers must train for adaptability, says the ***IEEE*** (2019), and new skill sets, with fallback strategies put in place for those who cannot be re-trained, and training programmes implemented at the level of high school or earlier to increase access to future employment. The ***UNI Global Union*** call for multi-stakeholder ethical AI governance bodies on global and regional levels, bringing together designers, manufacturers, developers, researchers, trade unions, lawyers, CSOs, owners, and employers. AI must benefit and empower people broadly and equally, with policies put in place to bridge the economic, technological, and social digital divides, and ensure a just transition with support for fundamental freedoms and rights.

The AI Now Institute works with diverse stakeholder groups to better understand the implications that AI will have for labour and work, including automation and early-stage integration of AI changing the nature of employment and working conditions in various sectors. ***The Future Society*** specifically asks how AI will affect the legal profession: 'If AI systems are demonstrably superior to

[11] https://sustainabledevelopment.un.org/?menu=1300

human attorneys at certain aspects of legal work, what are the ethical and professional implications for the practice of law?' (Future Society, 2019)

AI in the workplace will affect far more than workers' finances, and may offer various positive opportunities. As laid out by the **IEEE** (2019), AI may offer potential solutions to workplace bias – if it is developed with this in mind, as mentioned above – and reveal deficiencies in product development, allowing proactive improvement in the design phase (as opposed to retroactive improvement).

'RRI is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society).' (Von Schomberg, 2013)

Responsible research and innovation (RRI)

RRI is a growing area, especially in the EU, that draws from classical ethics to provide tools with which to address ethical concerns from the very outset of a project. When incorporated into a project's design phase, RRI increases the chances of design being both relevant and strong in terms of ethical alignment. Many research funders and organisations include RRI in their mission statements and within their research and innovation efforts (IEEE, 2019).

Lawfulness and justice

Several initiatives address the need for AI to be lawful, equitable, fair, just and subject to appropriate, pre-emptive governance and regulation. The many complex ethical problems surrounding AI translate directly and indirectly into discrete legal challenges. How should AI be labelled: as a product? An animal? A person? Something new?

The ***IEEE*** conclude that AI should not be granted any level of 'personhood', and that, while development, design and distribution of AI should fully comply with all applicable international and domestic law, there is much work to be done in defining and implementing the relevant legislation. Legal issues fall into a few categories: legal status, governmental use (transparency, individual rights), legal accountability for harm, and transparency, accountability, and verifiability. The IEEE suggest that AI should remain subject to the applicable regimes of property law; that stakeholders should identify the types of decisions that should never be delegated to AI, and ensure effective human control over those decisions via rules and standards; that existing laws should be scrutinised and reviewed for mechanisms that could practically give AI legal autonomy; and that manufacturers and operators should be required to comply with the applicable laws of all jurisdictions in which an AI could operate. They also recommend that governments reassess the legal status for AI as they become more sophisticated, and work closely with regulators, societal and industry actors and other stakeholders to ensure that the interests of humanity – and not the development of systems themselves – remain the guiding principle.

Control and the ethical use – or misuse – of AI

With more sophisticated and complex new AI come more sophisticated and complex possibilities for misuse. Personal data may be used maliciously or for profit, systems are at risk of hacking, and technology may be used exploitatively. This ties into informed use and public awareness: as we enter a new age of AI, with new systems and technology emerging that have never before been implemented, citizens must be kept up to date of the risks that may come with either the use or misuse of these.

The ***IEEE*** suggests new ways of educating the public on ethics and security issues, for example a 'data privacy' warning on smart devices that collect personal data; delivering this education in scalable, effective ways; and educating government, lawmakers, and enforcement agencies surrounding these issues, so they can work collaboratively with citizens – in a similar way to police officers providing safety lectures in schools – and avoid fear and confusion (IEEE, 2019).

Other issues include manipulation of behaviour and data. Humans must retain control over AI and oppose subversion. Most initiatives reviewed flag this as a potential issue facing AI as it develops, and flag that AI must behave in a way that is predictable and reliable, with appropriate means for redress, and be subject to validation and testing. AI must also work for the good of humankind, must not exploit people, and be regularly reviewed by human experts.

Personhood and AI

The issue of whether or not an AI deserves 'personhood' ties into debates surrounding accountability, autonomy, and responsibility: is it the AI itself that is responsible for its actions and consequences, or the person(s) who built them?

This concept, rather than allowing robots to be considered people in a human sense, would place robots on the same legal level as corporations. It is worth noting that corporations' legal personhood can currently shield the natural persons behind them from the implications of the law. However, ***The UNI Global Union*** asserts that legal responsibility lies with the creator, not the robot itself, and calls for a ban on attributing responsibility to robots.

Environmental harm and sustainability

The production, management, and implementation of AI must be sustainable and avoid environmental harm. This also ties in to the concept of well-being; a key recognised aspect of well-being is environmental, concerning the air, biodiversity, climate change, soil and water quality, and so on (IEEE, 2019). The ***IEEE*** (EAD, 2019) state that AI must do no harm to Earth's natural systems or exacerbate their degradation, and contribute to realising sustainable stewardship, preservation, and/or the restoration of Earth's natural systems. The ***UNI Global Union*** state that AI must put people and the planet first, striving to protect and even enhance our planet's biodiversity and ecosystems (UNI Global Union, n.d.). The ***Foundation for Responsible Robotics*** identifies a number of potential uses for AI in coming years, from agricultural and farming roles to monitoring of climate change and protection of endangered species. These require responsible, informed policies to govern AI and robotics, say the Foundation, to mitigate risk and support ongoing innovation and development.

Informed use: public education and awareness

Members of the public must be educated on the use, misuse, and potential harms of AI, via civic participation, communication, and dialogue with the public. The issue of consent – and how much an individual may reasonably and knowingly give – is core to this. For example, the ***IEEE*** raise several instances in which consent is less clear-cut than might be ethical: what if one's personal data are used to make inferences they are uncomfortable with or unaware of? Can consent be given when a system does not directly interact with an individual? This latter issue has been named the 'Internet of Other People's Things' (IEEE, 2019). Corporate environments also raise the issue of power imbalance; many employees do not have clear consent on how their personal data – including those on health – is used by their employer. To remedy this, the IEEE (2017) suggest employee data impact assessments to deal with these corporate nuances and ensure that no data is collected without employee consent. Data must also be only gathered and used for specific, explicitly stated, legitimate purposes, kept up-to-date, lawfully processed, and not kept for a longer period than necessary. In cases where subjects do not have a direct relationship with the system gathering data, consent must be dynamic, and the system designed to interpret data preferences and limitations on collection and use.

To increase awareness and understanding of AI, undergraduate and postgraduate students must be educated on AI and its relationship to sustainable human development, say the IEEE. Specifically, curriculum and core competencies should be defined and prepared; degree programmes focusing on engineering in international development and humanitarian relief should be exposed to the potential of AI applications; and awareness should be increased of the opportunities and risks faced by Lower Middle Income Countries in the implementation of AI in humanitarian efforts across the globe.

Many initiatives focus on this, including the ***Foundation for Responsible Robotics, Partnership on AI, Japanese Society for AI Ethical Guidelines, Future Society*** and ***AI Now Institute***; these and others maintain that clear, open and transparent dialogue between AI and society is key to the creation of understanding, acceptance, and trust.

Existential risk

According to the Future of Life Institute, the main existential issue surrounding AI 'is not malevolence, but competence' – AI will continually learn as they interact with others and gather data, leading them to gain intelligence over time and potentially develop aims that are at odds with those of humans.

*'You're probably not an evil ant-hater who steps on ants out of malice,' 'but if you're in charge of a hydroelectric green energy project and there's an anthill in the region to be flooded, too bad for the ants. A key goal of AI safety research is to never place humanity in the position of those ants' (*The Future of Life Institute, 2019).

AI also poses a threat in the form of ***autonomous weapons systems (AWS)***. As these are designed to cause physical harm, they raise numerous ethical quandaries. The IEEE (2019) lays out a number of recommendations to ensure that AWS are subject to meaningful human control: they suggest audit trails to guarantee accountability and control; adaptive learning systems that can explain their reasoning in a transparent, understandable way; that human operators of autonomous systems are identifiable, held responsible, and aware of the implications of their work; that autonomous behaviour is predictable; and that professional codes of ethics are developed to address the development of autonomous systems – especially those intended to cause harm. The pursuit of AWS may lead to an international arms race and geopolitical stability; as such, the IEEE recommend that systems designed to act outside the boundaries of human control or judgement are unethical and violate fundamental human rights and legal accountability for weapons use.

Given their potential to seriously harm society, these concerns must be controlled for and regulated pre-emptively, says the ***Foundation for Responsible Robotics***. Other initiatives that cover this risk explicitly include the ***UNI Global Union*** and the ***Future of Life Institute***, the latter of which cautions against an arms race in lethal autonomous weapons, and calls for planning and mitigation efforts for possible longer-term risks. We must avoid strong assumptions on the upper limits of future AI capabilities, assert the FLI's **Asilomar Principles**, and recognise that advanced AI represents a profound change in the history of life on Earth.

3.3. Case studies

3.3.1. Case study: healthcare robots

Artificial Intelligence and robotics are rapidly moving into the field of healthcare and will increasingly play roles in diagnosis and clinical treatment. For example, currently, or in the near future, robots will help in the diagnosis of patients; the performance of simple surgeries; and the monitoring of patients' health and mental wellness in short and long-term care facilities. They may also provide basic physical interventions, work as companion carers, remind patients to take their

medications, or help patients with their mobility. In some fundamental areas of medicine, such as medical image diagnostics, machine learning has been proven to match or even surpass our ability to detect illnesses.

Embodied AI, or robots, are already involved in a number of functions that affect people's physical safety. In June 2005, a surgical robot at a hospital in Philadelphia malfunctioned during prostate surgery, injuring the patient. In June 2015, a worker at a Volkswagen plant in Germany was crushed to death by a robot on the production line. In June 2016, a Tesla car operating in autopilot mode collided with a large truck, killing the car's passenger (Yadron and Tynan, 2016).

As robots become more prevalent, the potential for future harm will increase, particularly in the case of driverless cars, assistive robots and drones, which will face decisions that have real consequences for human safety and well-being. The stakes are much higher with embodied AI than with mere software, as robots have moving parts in physical space (Lin et al., 2017). Any robot with moving physical parts poses a risk, especially to vulnerable people such as children and the elderly.

Safety

Again, perhaps the most important ethical issue arising from the growth of AI and robotics in healthcare is that of safety and avoidance of harm. It is vital that robots should not harm people, and that they should be safe to work with. This point is especially important in areas of healthcare that deal with vulnerable people, such as the ill, elderly, and children.

Digital healthcare technologies offer the potential to improve accuracy of diagnosis and treatments, but to thoroughly establish a technology's long-term safety and performance investment in clinical trials is required. The debilitating side-effects of vaginal mesh implants and the continued legal battles against manufacturers (The Washington Post, 2019), stand as an example against shortcutting testing, despite the delays this introduces to innovating healthcare. Investment in clinical trials will be essential to safely implement the healthcare innovations that AI systems offer.

User understanding

The correct application of AI by a healthcare professional is important to ensure patient safety. For instance, the precise surgical robotic assistant 'the da Vinci' has proven a useful tool in minimising surgical recovery, but requires a trained operator (The Conversation, 2018).

A shift in the balance of skills in the medical workforce is required, and healthcare providers are preparing to develop the digital literacy of their staff over the next two decades (NHS' Topol Review, 2009). With genomics and machine learning becoming embedded in diagnoses and medical decision-making, healthcare professionals need to become digitally literate to understand each technological tool and use it appropriately. It is important for users to trust the AI presented but to be aware of each tool's strengths and weaknesses, recognising when validation is necessary. For instance, a generally accurate machine learning study to predict the risk of complications in patients with pneumonia erroneously considered those with asthma to be at low risk. It reached this conclusion because asthmatic pneumonia patients were taken directly to intensive care, and this higher-level care circumvented complications. The inaccurate recommendation from the algorithm was thus overruled (Pulmonology Advisor, 2017).

However, it's questionable to what extent individuals need to understand how an AI system arrived at a certain prediction in order to make autonomous and informed decisions. Even if an in-depth understanding of the mathematics is made obligatory, the complexity and learned nature of machine learning algorithms often prevent the ability to understand how a conclusion has been made from a dataset — a so called 'black box' (Schönberger, 2019). In such cases, one possible route

to ensure safety would be to license AI for specific medical procedures, and to 'disbar' the AI if a certain number of mistakes are made (Hart, 2018).

Data protection

Personal medical data needed for healthcare algorithms may be at risk. For instance, there are worries that data gathered by fitness trackers might be sold to third parties, such as insurance companies, who could use those data to refuse healthcare coverage (National Public Radio, 2018). Hackers are another major concern, as providing adequate security for systems accessed by a range of medical personnel is problematic (Forbes, 2018).

Pooling personal medical data is critical for machine learning algorithms to advance healthcare interventions, but gaps in information governance form a barrier against responsible and ethical data sharing. Clear frameworks for how healthcare staff and researchers use data, such as genomics, in a way that safeguards patient confidentiality is necessary to establish public trust and enable advances in healthcare algorithms (NHS' Topol Review, 2009).

Legal responsibility

Although AI promises to reduce the number of medical mishaps, when issues occur, legal liability must be established. If equipment can be proven to be faulty then the manufacturer is liable, but it is often tricky to establish what went wrong during a procedure and whether anyone, medical personnel or machine, is to blame. For instance, there have been lawsuits against the da Vinci surgical assistant (Mercury News, 2017), but the robot continues to be widely accepted (The Conversation, 2018).

In the case of 'black box' algorithms where it is impossible to ascertain how a conclusion is reached, it is tricky to establish negligence on the part of the algorithm's producer (Hart, 2018).

For now, AI is used as an aide for expert decisions, and so experts remain the liable party in most cases. For instance, in the aforementioned pneumonia case, if the medical staff had relied solely on the AI and sent asthmatic pneumonia patients home without applying their specialist knowledge, then that would be a negligent act on their part (Pulmonology Advisor, 2017; International Journal of Law and Information Technology, 2019).

Soon, the omission of AI could be considered negligence. For instance, in less developed countries with a shortage of medical professionals, withholding AI that detects diabetic eye disease and so prevents blindness, because of a lack of ophthalmologists to sign off on a diagnosis, could be considered unethical (The Guardian, 2019; International Journal of Law and Information Technology, 2019).

Bias

Non-discrimination is one of the fundamental values of the EU (see Article 21 of the EU Charter of Fundamental Rights), but machine learning algorithms are trained on datasets that often have proportionally less data available about minorities, and as such can be biased (Medium, 2014). This can mean that algorithms trained to diagnose conditions are less likely to be accurate for ethnic patients; for instance, in the dataset used to train a model for detecting skin cancer, less than 5 percent of the images were from individuals with dark skin, presenting a risk of misdiagnosis for people of colour (The Atlantic, 2018).

To ensure the most accurate diagnoses are presented to people of all ethnicities, algorithmic biases must be identified and understood. Even with a clear understanding of model design this is a difficult task because of the aforementioned 'black box' nature of machine learning. However, various codes of conduct and initiatives have been introduced to spot biases earlier. For instance,

The Partnership on AI, an ethics-focused industry group was launched by Google, Facebook, Amazon, IBM and Microsoft (The Guardian, 2016) — although, worryingly, this board is not very diverse.

Equality of access

Digital health technologies, such as fitness trackers and insulin pumps, provide patients with the opportunity to actively participate in their own healthcare. Some hope that these technologies will help to redress health inequalities caused by poor education, unemployment, and so on. However, there is a risk that individuals who cannot afford the necessary technologies or do not have the required 'digital literacy' will be excluded, so reinforcing existing health inequalities (The Guardian, 2019).

The UK's National Health Services' Widening Digital Participation programme is one example of how a healthcare service has tried to reduce health inequalities, by helping millions of people in the UK who lack the skills to access digital health services. Programmes such as this will be critical in ensuring equality of access to healthcare, but also in increasing the data from minority groups needed to prevent the biases in healthcare algorithms discussed above.

Quality of care

'There is remarkable potential for digital healthcare technologies to improve accuracy of diagnoses and treatments, the efficiency of care, and workflow for healthcare professionals' (NHS' Topol Review, 2019).

If introduced with careful thought and guidelines, companion and care robots, for example, could improve the lives of the elderly, reducing their dependence, and creating more opportunities for social interaction. Imagine a home-care robot that could: remind you to take your medications; fetch items for you if you are too tired or are already in bed; perform simple cleaning tasks; and help you stay in contact with your family, friends and healthcare provider via video link.
However, questions have been raised over whether a 'cold', emotionless robot can really substitute for a human's empathetic touch. This is particularly the case in long-term caring of vulnerable and often lonely populations, who derive basic companionship from caregivers. Human interaction is particularly important for older people, as research suggests that an extensive social network offers protection against dementia. At present, robots are far from being real companions. Although they can interact with people, and even show simulated emotions, their conversational ability is still extremely limited, and they are no replacement for human love and attention. Some might go as far as saying that depriving the elderly of human contact is unethical, and even a form of cruelty.

And does abandoning our elderly to cold machine care objectify (degrade) them, or human caregivers? It's vital that robots don't make elderly people feel like objects, or with even less control over their lives than when they were dependent on humans — otherwise they may feel like they are 'lumps of dead matter: to be pushed, lifted, pumped or drained, without proper reference to the fact that they are sentient beings' (Kitwood 1997).

In principle, autonomy, dignity and self-determination can all be thoroughly respected by a machine application, but it's unclear whether application of these roles in the sensitive field of medicine will be deemed acceptable. For instance, a doctor used a telepresence device to give a prognosis of death to a Californian patient; unsurprisingly the patient's family were outraged by this impersonal approach to healthcare (The Independent, 2019). On the other hand, it's argued that new technologies, such as health monitoring apps, will free up staff time for more direct interactions with patients, and so potentially increase the overall quality of care (The Guardian, Press Association, Monday 11 February 2019).

Deception

A number of 'carebots' are designed for social interactions and are often touted to provide an emotional therapeutic role. For instance, care homes have found that a robotic seal pup's animal-like interactions with residents brightens their mood, decreases anxiety and actually increases the sociability of residents with their human caregivers. However, the line between reality and imagination is blurred for dementia patients, so is it dishonest to introduce a robot as a pet and encourage a social-emotional involvement? (KALW, 2015) And if so, is if morally justifiable?

Companion robots and robotic pets could alleviate loneliness amongst older people, but this would require them believing, in some way, that a robot is a sentient being who cares about them and has feelings — a fundamental deception. Turkle et al. (2006) argue that 'the fact that our parents, grandparents and children might say 'I love you' to a robot who will say 'I love you' in return, does not feel completely comfortable; it raises questions about the kind of authenticity we require of our technology'. Wallach and Allen (2009) agree that robots designed to detect human social gestures and respond in kind all use techniques that are arguably forms of deception. For an individual to benefit from owning a robot pet, they must continually delude themselves about the real nature of their relation with the animal. What's more, encouraging elderly people to interact with robot toys has the effect of infantilising them.

Autonomy

It's important that healthcare robots actually benefit the patients themselves, and are not just designed to reduce the care burden on the rest of society — especially in the case of care and companion AI. Robots could empower disabled and older people and increase their independence; in fact, given the choice, some might prefer robotic over human assistance for certain intimate tasks such as toileting or bathing. Robots could be used to help elderly people live in their own homes for longer, giving them greater freedom and autonomy. However, how much control, or autonomy, should a person be allowed if their mental capability is in question? If a patient asked a robot to throw them off the balcony, should the robot carry out that command?

Liberty and privacy

As with many areas of AI technology, the privacy and dignity of users' needs to be carefully considered when designing healthcare service and companion robots. Working in people's homes means that robots will be privy to private moments such as bathing and dressing; if these moments are recorded, who should have access to the information, and how long should recordings be kept? The issue becomes more complicated if an elderly person's mental state deteriorates and they become confused — someone with Alzheimer's could forget that a robot was monitoring them, and could perform acts or say things thinking that they are in the privacy of their own home. Home-care robots need to be able to balance their user's privacy and nursing needs, for example by knocking and awaiting an invitation before entering a patient's room, except in a medical emergency.

To ensure their charge's safety, robots might sometimes need to act as supervisors, restricting their freedoms. For example, a robot could be trained to intervene if the cooker was left on, or the bath was overflowing. Robots might even need to restrain elderly people from carrying out potentially dangerous actions, such as climbing up on a chair to get something from a cupboard. Smart homes with sensors could be used to detect that a person is attempting to leave their room, and lock the door, or call staff — but in so doing the elderly person would be imprisoned.

Moral agency

'There's very exciting work where the brain can be used to control things, like maybe they've lost the use of an arm…where I think the real concerns lie is with things like behavioural targeting: going straight to the hippocampus and people pressing 'consent', like we do now, for data access'. (John Havens)

Robots do not have the capacity for ethical reflection or a moral basis for decision-making, and thus humans must currently hold ultimate control over any decision-making. An example of ethical reasoning in a robot can be found in the 2004 dystopian film 'I, Robot', where Will Smith's character disagreed with how the robots of the fictional time used cold logic to save his life over that of a child's. If more automated healthcare is pursued, then the question of moral agency will require closer attention. Ethical reasoning is being built into robots, but moral responsibility is about more than the application of ethics — and it is unclear whether robots of the future will be able to handle the complex moral issues in healthcare (Goldhill, 2016).

Trust

Larosa and Danks (2018) write that AI may affect human-human interactions and relationships within the healthcare domain, particularly that between patient and doctor, and potentially disrupt the trust we place in our doctor.

'Psychology research shows people mistrust those who make moral decisions by calculating costs and benefits — like computers do' (The Guardian, 2017). Our distrust of robots may also come from the number of robots running amok in dystopian science fiction. News stories of computer mistakes — for instance, of an image-identifying algorithm mistaking a turtle for a gun (The Verge, 2017) — alongside worries over the unknown, privacy and safety are all reasons for resistance against the uptake of AI (Global News Canada, 2016).

Firstly, doctors are explicitly certified and licensed to practice medicine, and their license indicates that they have specific skills, knowledge, and values such as 'do no harm'. If a robot replaces a doctor for a particular treatment or diagnostic task, this could potentially threaten patient-doctor trust, as the patient now needs to know whether the system is appropriately approved or 'licensed' for the functions it performs.

Secondly, patients trust doctors because they view them as paragons of expertise. If doctors were seen as 'mere users' of the AI, we would expect their role to be downgraded in the public's eye, undermining trust.

Thirdly, a patient's experiences with their doctor are a significant driver of trust. If a patient has an open line of communication with their doctor, and engages in conversation about care and treatment, then the patient will trust the doctor. Inversely, if the doctor repeatedly ignores the patient's wishes, then these actions will have a negative impact on trust. Introducing AI into this dynamic could increase trust — if the AI reduced the likelihood of misdiagnosis, for example, or improved patient care. However, AI could also decrease trust if the doctor delegated too much diagnostic or decision-making authority to the AI, undercutting the position of the doctor as an authority on medical matters.

As the body of evidence grows to support the therapeutic benefits for each technological approach, and as more robotic interacting systems enter the marketplace, then trust in robots is likely to increase. This has already happened for robotic healthcare systems such as the da Vinci surgical robotic assistant (The Guardian, 2014).

Employment replacement

As in other industries, there is a fear that emerging technologies may threaten employment (The Guardian, 2017), for instance, there are carebots now available that can perform up to a third of nurses' work (Tech Times, 2018). Despite these fears, the NHS' Topol Review (2009) concluded that 'these technologies will not replace healthcare professionals but will enhance them ('augment them'), giving them more time to care for patients'. The review also outlined how the UK's NHS will nurture a learning environment to ensure digitally capable employees.

3.3.2 Case study: Autonomous Vehicles

Autonomous Vehicles (AVs) are vehicles that are capable of sensing their environment and operating with little to no input from a human driver. While the idea of self-driving cars has been around since at least the 1920s, it is only in recent years that technology has developed to a point where AVs are appearing on public roads.

According to automotive standardisation body SAE International (2018), there are six levels of driving automation:

0	No automation	An automated system may issue warnings and/or momentarily intervene in driving, but has no sustained vehicle control.
1	Hands on	The driver and automated system share control of the vehicle. For example, the automated system may control engine power to maintain a set speed (e.g. Cruise Control), engine and brake power to maintain and vary speed (e.g. Adaptive Cruise Control), or steering (e.g. Parking Assistance). The driver must be ready to retake full control at any time.
2	Hands off	The automated system takes full control of the vehicle (including accelerating, braking, and steering). However, the driver must monitor the driving and be prepared to intervene immediately at any time.
3	Eyes off	The driver can safely turn their attention away from the driving tasks (e.g. to text or watch a film) as the vehicle will handle any situations that call for an immediate response. However, the driver must still be prepared to intervene, if called upon by the AV to do so, within a timeframe specified by the AV manufacturer.
4	Minds off	As level 3, but no driver attention is ever required for safety, meaning the driver can safely go to sleep or leave the driver's seat.
5	Steering wheel optional	No human intervention is required at all. An example of a level 5 AV would be a robotic taxi.

Some of the lower levels of automation are already well-established and on the market, while higher level AVs are undergoing development and testing. However, as we transition up the levels and put more responsibility on the automated system than the human driver, a number of ethical issues emerge.

Societal and Ethical Impacts of AVs

'We cannot build these tools saying, 'we know that humans act a certain way, we're going to kill them – here's what to do'.' (John Havens)

Public safety and the ethics of testing on public roads

At present, cars with 'assisted driving' functions are legal in most countries. Notably, some Tesla models have an Autopilot function, which provides level 2 automation (Tesla, nd). Drivers are legally allowed to use assisted driving functions on public roads provided they remain in charge of the

vehicle at all times. However, many of these assisted driving functions have not yet been subject to independent safety certification, and as such may pose a risk to drivers and other road users. In Germany, a report published by the Ethics Commission on Automated Driving highlights that it is the public sector's responsibility to guarantee the safety of AV systems introduced and licensed on public roads, and recommends that all AV driving systems be subject to official licensing and monitoring (Ethics Commision, 2017).

In addition, it has been suggested that the AV industry is entering its most dangerous phase, with cars being not yet fully autonomous but human operators not being fully engaged (Solon, 2018). The risks this poses have been brought to widespread attention following the first pedestrian fatality involving an autonomous car. The tragedy took place in Arizona, USA, in May 2018, when a level 3 AV being tested by Uber collided with 49-year-old Elaine Herzberg as she was walking her bike across a street one night. It was determined that Uber was 'not criminally liable' by prosecutors (Shepherdson and Somerville, 2019), and the US National Transportation Safety Board's preliminary report (NTSB, 2018), which drew no conclusions about the cause, said that all elements of the self-driving system were operating normally at the time of the crash. Uber said that the driver is relied upon to intervene and take action in situations requiring emergency braking – leading some commentators to call out the misleading communication to consumers around the terms 'self-driving cars' and 'autopilot' (Leggett, 2018). The accident also caused some to condemn the practice of testing AV systems on public roads as dangerous and unethical, and led Uber to temporarily suspend its self-driving programme (Bradshaw, 2018).

This issue of human safety — of both public and passenger — is emerging as a key issue concerning self-driving cars. Major companies — Nissan, Toyota, Tesla, Uber, Volkswagen — are developing autonomous vehicles capable of operating in complex, unpredictable environments without direct human control, and capable of learning, inferring, planning and making decisions.

Self-driving vehicles could offer multiple benefits: statistics show you're almost certainly safer in a car driven by a computer than one driven by a human. They could also ease congestion in cities, reduce pollution, reduce travel and commute times, and enable people to use their time more productively. However, they won't mean the end of road traffic accidents. Even if a self-driving car has the best software and hardware available, there is still a collision risk. An autonomous car could be surprised, say by a child emerging from behind a parked vehicle, and there is always the issue of *how: how* should such cars be programmed when they must decide whose safety to prioritise?

Driverless cars may also have to choose between the safety of passengers and other road users. Say that a car travels around a corner where a group of school children are playing; there is not enough time to stop, and the only way the car can avoid hitting the children is to swerve into a brick wall — endangering the passenger. Whose safety should the car prioritise: the children's', or the passenger's?

Processes and technologies for accident investigation

AVs are complex systems that often rely on advanced machine learning technologies. Several serious accidents have already occurred, including a number of fatalities involving level 2 AVs:

- In January 2016, 23-year-old Gao Yaning died when his Tesla Model S crashed into the back of a road-sweeping truck on a highway in Hebei, China. The family believe Autopilot was engaged when the accident occurred and accuse Tesla of exaggerating the system's capabilities. Tesla state that the damage to the vehicle made it impossible to determine whether Autopilot was engaged and, if so, whether it malfunctioned. A civil case into the crash is ongoing, with a third-party appraiser reviewing data from the vehicle (Curtis, 2016).

- In May 2016, 40-year-old Joshua Brown died when his Tesla Model S collided with a truck while Autopilot was engaged in Florida, USA. An investigation by the National Highways and Transport Safety Agency found that the driver, and not Tesla, were at fault (Gibbs, 2016). However, the National Highway Traffic Safety Administration later determined that both Autopilot and over-reliance by the motorist on Tesla's driving aids were to blame (Felton, 2017).

- In March 2018, Wei Huang was killed when his Tesla Model X crashed into a highway safety barrier in California, USA. According to Tesla, the severity of the accident was 'unprecedented'. The National Transportation Safety Board later published a report attributing the crash to an Autopilot navigation mistake. Tesla is now being sued by the victim's family (O'Kane, 2018).

Unfortunately, efforts to investigate these accidents have been stymied by the fact that standards, processes, and regulatory frameworks for investigating accidents involving AVs have not yet been developed or adopted. In addition, the proprietary data logging systems currently installed in AVs mean that accident investigators rely heavily on the cooperation of manufacturers to provide critical data on the events leading up to an accident (Stilgoe and Winfield, 2018).

One solution is to fit all future AVs with industry standard event data recorders — a so-called 'ethical black box' — that independent accident investigators could access. This would mirror the model already in place for air accident investigations (Sample, 2017).

Near-miss accidents

At present, there is no system in place for the systematic collection of near-miss accidents. While it is possible that manufacturers are collecting this data already, they are not under any obligation to do so — or to share the data. The only exception at the moment is the US state of California, which requires all companies that are actively testing AVs on public roads to disclose the frequency at which human drivers were forced to take control of the vehicle for safety reasons (known as 'disengagement').

In 2018, the number of disengagements by AV manufacturer varied significantly, from one disengagement for every 11,017 miles driven by Waymo AVs to one for every 1.15 miles driven by Apple AVs (Hawkins, 2019). Data on these disengagements reinforces the importance of ensuring that human safety drivers remain engaged. However, the Californian data collection process has been criticised, with some claiming its ambiguous wording and lack of strict guidelines enables companies to avoid reporting certain events that could be termed near-misses.

Without access to this type of data, policymakers cannot account for the frequency and significance of near-miss accidents, or assess the steps taken by manufacturers as a result of these near-misses. Again, lessons could be learned from the model followed in air accident investigations, in which all near misses are thoroughly logged and independently investigated. Policymakers require comprehensive statistics on all accidents and near-misses in order to inform regulation.

Data privacy

It is becoming clear that manufacturers collect significant amounts of data from AVs. As these vehicles become increasingly common on our roads, the question emerges: to what extent are these data compromising the privacy and data protection rights of drivers and passengers?

Already, data management and privacy issues have appeared, with some raising concerns about the potential misuse of AV data for advertising purposes (Lin, 2014). Tesla have also come under fire for the unethical use of AV data logs. In an investigation by *The Guardian*, the newspaper found multiple instances where the company shared drivers' private data with the media following crashes, without

their permission, to prove that its technology was not responsible (Thielman, 2017). At the same time, Tesla does not allow customers to see their own data logs.

One solution, proposed by the German Ethics Commission on Automated Driving, is to ensure that that all AV drivers be given full data sovereignty (Ethics Commission, 2017). This would allow them to control how their data is used.

Employment

The growth of AVs is likely to put certain jobs — most pertinently bus, taxi, and truck drivers — at risk.

In the medium term, truck drivers face the greatest risk as long-distance trucks are at the forefront of AV technology (Viscelli, 2018). In 2016, the first commercial delivery of beer was made using a self-driving truck, in a journey covering 120 miles and involving no human action (Isaac, 2016). Last year saw the first fully driverless trip in a self-driving truck, with the AV travelling seven miles without a single human on board (Cannon, 2018).

Looking further forward, bus drivers are also likely to lose jobs as more and more buses become driverless. Numerous cities across the world have announced plans to introduce self-driving shuttles in the future, including Edinburgh (Calder, 2018), New York (BBC, 2019a) and Singapore (BBC 2017). In some places, this vision has already become a reality; the Las Vegas shuttle famously got off to a bumpy start when it was involved in a collision on its first day of operation (Park, 2017), and tourists in the small Swiss town of Neuhausen Rheinfall can now hop on a self-driving bus to visit the nearby waterfalls (CNN, 2018). In the medium term, driverless buses will likely be limited to routes that travel along 100% dedicated bus lanes. Nonetheless, the advance of self-driving shuttles has already created tensions with organised labour and city officials in the USA (Weinberg, 2019). Last year, the Transport Workers Union of America formed a coalition in an attempt to stop autonomous buses from hitting the streets of Ohio (Pfleger, 2018).

Fully autonomous taxis will likely only become realistic in the long term, once AV technology has been fully tested and proven at levels 4 and 5. Nonetheless, with plans to introduce self-driving taxis in London by 2021 (BBC, 2018), and an automated taxi service already available in Arizona, USA (Sage, 2019), it is easy to see why taxi drivers are uneasy.

The quality of urban environments

In the long-term, AVs have the potential to reshape our urban environment. Some of these changes may have negative consequences for pedestrians, cyclists and locals. As driving becomes more automated, there will likely be a need for additional infrastructure (e.g. AV-only lanes). There may also be more far-reaching effects for urban planning, with automation shaping the planning of everything from traffic congestion and parking to green spaces and lobbies (Marshall and Davies, 2018). The rollout of AVs will also require that 5G network coverage is extended significantly — again, something with implications for urban planning (Khosravi, 2018).
The environmental impact of self-driving cars should also be considered. While self-driving cars have the potential to significantly reduce fuel usage and associated emissions, these savings could be counteracted by the fact that self-driving cars make it easier and more appealing to drive long distances (Worland, 2016). The impact of automation on driving behaviours should therefore not be underestimated.

Legal and ethical responsibility

From a legal perspective, who is responsible for crashes caused by robots, and how should victims be compensated (if at all) when a vehicle controlled by an algorithm causes injury? If courts cannot resolve this problem, robot manufacturers may incur unexpected costs that would discourage investment. However, if victims are not properly compensated then autonomous vehicles are unlikely to be trusted or accepted by the public.

Robots will need to make judgement calls in conditions of uncertainty, or 'no win' situations. However, which ethical approach or theory should a robot be programmed to follow when there's no legal guidance? As Lin et al. explain, different approaches can generate different results, including the number of crash fatalities.

Additionally, who should choose the ethics for the autonomous vehicle — drivers, consumers, passengers, manufacturers, politicians? Loh and Loh (2017) argue that responsibility should be shared among the engineers, the driver and the autonomous driving system itself. However, Millar (2016) suggests that the user of the technology, in this case the passenger in the self-driving car, should be able to decide what ethical or behavioural principles the robot ought to follow. Using the example of doctors, who do not have the moral authority to make important decisions on end-of-life care without the informed consent of their patients, he argues that there would be a moral outcry if engineers designed cars without either asking the driver directly for their input, or informing the user ahead of time how the car is programmed to behave in certain situations.

Ethical dilemmas in development

In 2014, the Open Roboethics initiative (ORi 2014a, 2014b) conducted a poll asking people what they thought an autonomous car in which they were a passenger should do if a child stepped out in front of the vehicle in a tunnel. The car wouldn't have time to brake and spare the child, but could swerve into the walls of the tunnel, killing the passenger. This is a spin on the classic 'trolley dilemma', where one has the option to divert a runaway trolley from a path that would hurt several people onto the path that would only hurt one.

36 % of participants said that they would prefer the car to swerve into the wall, saving the child; however, the majority (64 %) said they would wish to save themselves, thus sacrificing the child. 44 % of participants thought that the passenger should be able to choose the car's course of action, while 33 % said that lawmakers should choose. Only 12 % said that the car's manufacturers should make the decision. These results suggest that people do not like the idea of engineers making moral decisions on their behalf.

Asking for the passenger's input in every situation would be impractical. However, Millar (2016) suggests a 'setup' procedure where people could choose their ethics settings after purchasing a new car. Nonetheless, choosing how the car reacts in advance could be seen as premeditated harm, if, for example a user programmed their vehicle to always avoid vehicle collisions by swerving into cyclists. This would increase the user's accountability and liability, whilst diverting responsibility away from manufacturers.

3.3.3 Case study: Warfare and weaponisation

Although partially autonomous and intelligent systems have been used in military technology since at least the Second World War, advances in machine learning and AI signify a turning point in the use of automation in warfare.

AI is already sufficiently advanced and sophisticated to be used in areas such as satellite imagery analysis and cyber defence, but the true scope of applications has yet to be fully realised. A recent report concludes that AI technology has the potential to transform warfare to the same, or perhaps even a greater, extent than the advent of nuclear weapons, aircraft, computers and biotechnology (Allen and Chan, 2017). Some key ways in which AI will impact militaries are outlined below.

Lethal autonomous weapons

As automatic and autonomous systems have become more capable, militaries have become more willing to delegate authority to them. This is likely to continue with the widespread adoption of AI, leading to an AI inspired arms-race. The Russian Military Industrial Committee has already approved an aggressive plan whereby 30% of Russian combat power will consist of entirely remote-controlled and autonomous robotic platforms by 2030. Other countries are likely to set similar goals. While the United States Department of Defense has enacted restrictions on the use of autonomous and semi-autonomous systems wielding lethal force, other countries and non-state actors may not exercise such self-restraint.

Drone technologies

Standard military aircraft can cost more than US$100 million per unit; a high-quality quadcopter Unmanned Aerial Vehicle, however, currently costs roughly US$1,000, meaning that for the price of a single high-end aircraft, a military could acquire one million drones. Although current commercial drones have limited range, in the future they could have similar ranges to ballistic missiles, thus rendering existing platforms obsolete.

Robotic assassination

Widespread availability of low-cost, highly-capable, lethal, and autonomous robots could make targeted assassination more widespread and more difficult to attribute. Automatic sniping robots could assassinate targets from afar.

Mobile-robotic-Improvised Explosive Devices

As commercial robotic and autonomous vehicle technologies become widespread, some groups will leverage this to make more advanced Improvised Explosive Devices (IEDs). Currently, the technological capability to rapidly deliver explosives to a precise target from many miles away is restricted to powerful nation states. However, if long distance package delivery by drone becomes a reality, the cost of precisely delivering explosives from afar would fall from millions of dollars to thousands or even hundreds. Similarly, self-driving cars could make suicide car bombs more frequent and devastating since they no longer require a suicidal driver.

Hallaq et al. (2017) also highlight key areas in which machine learning is likely to affect warfare. They describe an example where a Commanding Officer (CO) could employ an Intelligent Virtual Assistant (IVA) within a fluid battlefield environment that automatically scanned satellite imagery to detect specific vehicle types, helping to identify threats in advance. It could also predict the enemy's intent, and compare situational data to a stored database of hundreds of previous wargame exercises and live engagements, providing the CO with access to a level of accumulated knowledge that would otherwise be impossible to accrue.

Employing AI in warfare raises several **legal and ethical questions**. One concern is that automated weapon systems that exclude human judgment could violate International Humanitarian Law, and threaten our fundamental right to life and the principle of human dignity. AI could also lower the threshold of going to war, affecting global stability.

International Humanitarian law stipulates that any attack needs to distinguish between combatants and non-combatants, be proportional and must not target civilians or civilian objects. Also, no attack should unnecessarily aggravate the suffering of combatants. AI may be unable to fulfil these principles without the involvement of human judgment. In particular, many researchers are concerned that Lethal Autonomous Weapon Systems (LAWS) — a type of autonomous military robot that can independently search for and 'engage' targets using lethal force — may not meet the standards set by International Humanitarian Law, as they are not able to distinguish civilians from

combatants, and would not be able to judge whether the force of the attack was proportional given the civilian damage it would incur.

Amoroso and Tamburrini (2016, p. 6) argue that: '[LAWS must be] capable of respecting the principles of distinction and proportionality at least as well as a competent and conscientious human soldier'. However, Lim (2019) points out that while LAWS that fail to meet these requirements should not be deployed, one day LAWS *will* be sophisticated enough to meet the requirements of distinction and proportionality. Meanwhile, Asaro (2012) argues that it doesn't matter how good LAWS get; it is a moral requirement that only a human should initiate lethal force, and it is simply morally wrong to delegate life or death decisions to machines.

Some argue that delegating the decision to kill a human to a machine is an infringement of basic human dignity, as robots don't feel emotion, and can have no notion of sacrifice and what it means to take a life. As Lim et al (2019) explain, 'a machine, bloodless and without morality or mortality, cannot fathom the significance of using force against a human being and cannot do justice to the gravity of the decision'.

Robots also have no concept of what it means to kill the 'wrong' person. 'It is only because humans can feel the rage and agony that accompanies the killing of humans that they can understand sacrifice and the use of force against a human. Only then can they realise the 'gravity of the decision' to kill' (Johnson and Axinn 2013, p. 136).

However, others argue that there is no particular reason why being killed by a machine would be a subjectively worse, or less dignified, experience than being killed by a cruise missile strike. 'What matters is whether the victim experiences a sense of humiliation in the process of getting killed. Victims being threatened with a potential bombing will not care whether the bomb is dropped by a human or a robot' (Lim et al, 2019). In addition, not all humans have the emotional capacity to conceptualise sacrifice or the relevant emotions that accompany risk. In the heat of battle, soldiers rarely have time to think about the concept of sacrifice, or generate the relevant emotions to make informed decisions each time they deploy lethal force.

Additionally, who should be held accountable for the actions of autonomous systems — the commander, programmer, or the operator of the system? Schmit (2013) argues that the responsibility for committing war crimes should fall on both the individual who programmed the AI, and the commander or supervisor (assuming that they knew, or should have known, the autonomous weapon system had been programmed and employed in a war crime, and that they did nothing to stop it from happening).

4. AI standards and regulation

A small new generation of ethical standards are emerging as the ethical, legal and societal impacts of artificial intelligence and robotics are further understood. Whether a standard clearly articulates explicit or implicit ethical concerns, all standards embody some kind of ethical principle (Winfield, 2019a). The standards that do exist are still in development and there is limited publicly available information on them.

Perhaps the earliest explicit ethical standard in robotics is BS 8611 Guide to the Ethical Design and Application of Robots and Robotic Systems (British Standard BS 8611, 2016). BS8611 is not a code of practice, but guidance on how designers can identify potential ethical harm, undertake an ethical risk assessment of their robot or AI, and mitigate any ethical risks identified. It is based on a set of 20 distinct ethical hazards and risks, grouped under four categories: societal, application, commercial & financial, and environmental.

Advice on measures to mitigate the impact of each risk is given, along with suggestions on how such measures might be verified or validated. The societal hazards include, for example, loss of trust, deception, infringements of privacy and confidentiality, addiction, and loss of employment. Ethical Risk Assessment should consider also foreseeable misuse, risks leading to stress and fear (and their minimisation), control failure (and associated psychological effect), reconfiguration and linked changes to responsibilities, hazards associated with specific robotics applications. Particular attention is paid to robots that can learn and the implications of robot enhancement that arise, and the standard argues that the ethical risk associated with the use of a robot should not exceed the risk of the same activity when conducted by a human.

British Standard BS 8611 assumes that physical hazards imply ethical hazards, and defines ethical harm as affecting 'psychological and/or societal and environmental well-being.' It also recognises that physical and emotional hazards need to be balanced against expected benefits to the user.
The standard highlights the need to involve the public and stakeholders in development of robots and provides a list of key design considerations including:

- Robots should not be designed primarily to kill humans;
- Humans remain responsible agents;
- It must be possible to find out who is responsible for any robot;
- Robots should be safe and fit for purpose;
- Robots should not be designed to be deceptive;
- The precautionary principle should be followed;
- Privacy should be built into the design;
- Users should not be discriminated against, nor forced to use a robot.

Particular guidelines are provided for roboticists, particularly those conducting research. These include the need to engage the public, consider public concerns, work with experts from other disciplines, correct misinformation and provide clear instructions. Specific methods to ensure ethical use of robots include: user validation (to ensure robot can/is operated as expected), software verification (to ensure software works as anticipated), involvement of other experts in ethical assessment, economic and social assessment of anticipated outcomes, assessment of any legal implications, compliance testing against relevant standards. Where appropriate, other guidelines and ethical codes should be taken into consideration in the design and operation of robots (e.g. medical or legal codes relevant in specific contexts). The standard also makes the case that military application of robots does not remove the responsibility and accountability of humans.

The IEEE Standards Association has also launched a standard via its global initiative on the Ethics of Autonomous and Intelligent Systems. Positioning 'human well-being' as a central precept, the IEEE initiative explicitly seeks to reposition robotics and AI as technologies for improving the human condition rather than simply vehicles for economic growth (Winfield, 2019a). Its aim is to educate, train and empower AI/robot stakeholders to 'prioritise ethical considerations so that these technologies are advanced for the benefit of humanity.'

There are currently 14 IEEE standards working groups working on drafting so-called 'human' standards that have implications for artificial intelligence (Table 4.1).

Table 2: IEEE 'human standards' with implications for AI

Standard		Aims/Objectives
P7000	Model Process for Addressing Ethical Concerns During System Design	To establish a process for **ethical design of Autonomous and Intelligent Systems**.
P7001	Transparency of Autonomous Systems	To ensure the **transparency of autonomous systems to a range of stakeholders**. It specifically will address: • *Users*: ensuring users understand what the system does and why, with the intention of building trust; • *Validation and certification*: ensuring the system is subject to scrutiny; • *Accidents*: enabling accident investigators to undertake investigation; • *Lawyers and expert witnesses:* ensuring that, following an accident, these groups are able to give evidence; • *Disruptive technology (e.g. driverless cars*): enabling the public to assess technology (and, if appropriate, build confidence).
P7002	Data Privacy Process	To establish standards for **the ethical use of personal data** in software engineering processes. It will develop and describe privacy impact assessments (PIA) that can be used to identify the need for, and effectiveness of, privacy control measures. It will also provide checklists for those developing software that uses personal information.

P7003	Algorithmic Bias Considerations	To help algorithm developers make explicit the ways in which they have sought to **eliminate or minimise the risk of bias** in their products. This will address the use of overly subjective information and help developers ensure they are compliant with legislation regarding protected characteristics (e.g. race, gender). It is likely to include: • Benchmarking processes for the selection of data sets; • Guidelines on communicating the boundaries for which the algorithm has been designed and validated (guarding against unintended consequences of unexpected uses); • Strategies to avoid incorrect interpretation of system outputs by users.
P7004	Standard for Child and Student Data Governance	Specifically **aimed at educational institutions,** this will provide guidance on accessing, collecting, storing, using, sharing and destroying child/student data.
P7005	Standard for Transparent Employer Data Governance	Similar to P7004, but **aimed at employers**.
P7006	Standard for Personal Data Artificial Intelligence (AI) Agent	Describes the technical elements required to create and grant access to **personalised AI.** It will enable individuals to safely organise and share their personal information at a machine-readable level, and enable personalised AI to act as a proxy for machine-to-machine decisions.
P7007	Ontological Standard for Ethically Driven Robotics and Automation Systems	This standard brings together engineering and philosophy **to ensure that user well-being is considered throughout the product life cycle**. It intends to identify ways to maximise benefits and minimise negative impacts, and will also consider the ways in which communication can be clear between diverse communities.

P7008	Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems	Drawing on 'nudge theory', this standard seeks **to delineate current or potential nudges that robots or autonomous systems might undertake**. It recognises that nudges can be used for a range of reasons, but that they seek to affect the recipient emotionally, change behaviours and can be manipulative, and seeks to elaborate methodologies for ethical design of AI using nudge.
P7009	Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems	To create effective methodologies for **the development and implementation of robust, transparent and accountable fail-safe mechanisms**. It will address methods for measuring and testing a system's ability to fail safely.
P7010	Well-being Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems	To establish a baseline for metrics used **to assess well-being factors that could be affected by autonomous systems**, and for how human well-being could proactively be improved.
P7011	Standard for the Process of Identifying and Rating the Trustworthiness of News Sources	Focusing on news information, this standard sets out t**o standardise the processes for assessing the factual accuracy of news stories**. It will be used to produce a 'trustfulness' score. This standard seeks to address the negative effects of unchecked 'fake' news, and is designed to restore trust in news purveyors.
P7012	Standard for Machine Readable Personal Privacy Terms	To establish **how privacy terms are presented** and how they could be read and accepted by machines.
P7013	Inclusion and Application Standards for Automated Facial Analysis Technology	To provide **guidelines on the data used in facial recognition**, the requirements for diversity, and benchmarking of applications and situations in which facial recognition should not be used.

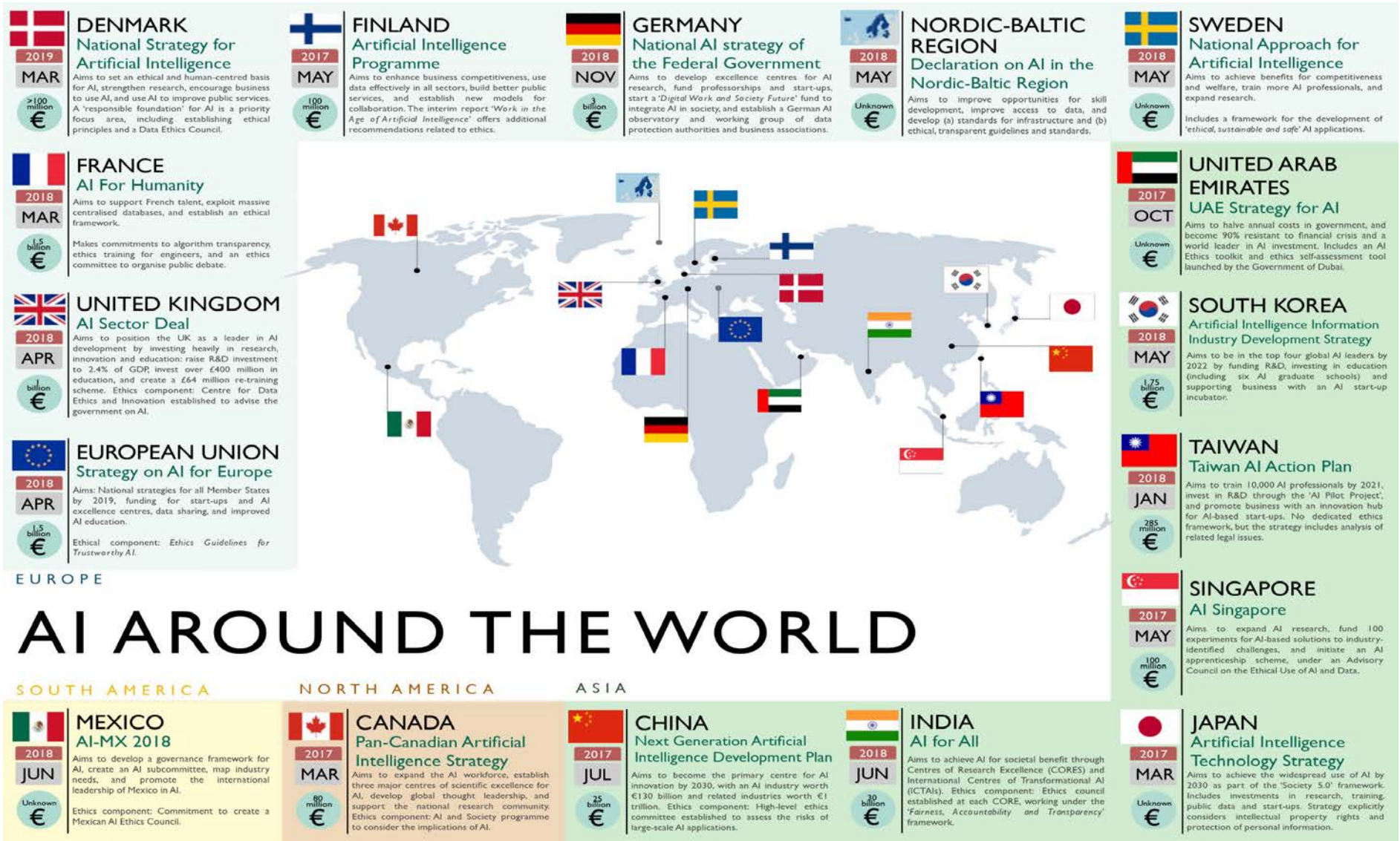
5. National and International Strategies on AI

As the technology behind AI continues to progress beyond expectations, policy initiatives are springing up across the globe to keep pace with these developments.

The first national strategy on AI was launched by Canada in March 2017, followed soon after by technology leaders Japan and China. In Europe, the European Commission put forward a communication on AI, initiating the development of independent strategies by Member States. An American AI initiative is expected soon, alongside intense efforts in Russia to formalise their 10-point plan for AI.

These initiatives differ widely in terms of their goals, the extent of their investment, and their commitment to developing ethical frameworks, reviewed here as of May 2019.

Figure 3: National and International Strategies on AI published as of May 2019.



5.1. Europe

The European Commission's Communication on Artificial Intelligence (European Commission, 2018a), released in April 2018, paved the way to the first international strategy on AI. The document outlines a coordinated approach to maximise the benefits, and address the challenges, brought about by AI.

The Communication on AI was formalised nine months later with the presentation of a coordinated plan on AI (European Commission, 2018b). The plan details seven objectives, which include financing start-ups, investing €1.5 billion in several 'research excellence centres', supporting masters and PhDs in AI and creating common European data spaces.

Objective 2.6 of the plan is to develop 'ethics guidelines with a global perspective'. The Commission appointed an independent high-level expert group to develop their ethics guidelines, which – following consultation – were published in their final form in April 2019 (European Commission High-Level Expert Group on Artificial Intelligence, 2019). The Guidelines list key requirements that AI systems must meet in order to be trustworthy.

The EU's seven requirements for trustworthy AI:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental wellbeing
7. Accountability

Source: European Commission High-Level Expert Group on Artificial Intelligence, 2019

The EU's High-Level Expert Group on AI shortly after released a further set of policy and investment guidelines for trustworthy AI (European Commission High-Level Expert Group on AI, 2019b), which includes a number of important recommendations around protecting people, boosting uptake of AI in the private sector, expanding European research capacity in AI and developing ethical data management practices.

The Council of Europe also has various ongoing projects regarding the application of AI and in September 2019 established an Ad Hoc Committee on Artificial Intelligence (CAHAI). The committee will assess the potential elements of a legal framework for the development and application of AI, based on the Council's founding principles of human rights, democracy and the rule of law (Council of Europe, 2019a).

Looking ahead, the next European Commission President, Ursula von der Leyen, has announced AI as a priority for the next Commission, including legislation for a coordinated approach on the 'human and ethical implications' of AI (Kayali, 2019; von der Leyen, 2019).

The European Commission provides a unifying framework for AI development in the EU, but Member States are also required to develop their own national strategies.

Finland was the first Member State to develop a national programme on AI (Ministry of Economic Affairs and Employment of Finland, 2018a). The programme is based on two reports, *Finland's Age of Artificial Intelligence* and *Work in the Age of Artificial Intelligence* (Ministry of Economic Affairs and Employment of Finland, 2017, 2018b). Policy objectives focus on investment for business competitiveness and public services. Although recommendations have already been incorporated into policy, Finland's AI steering group will run until the end of the present Government's term, with a final report expected imminently.

So far, Denmark, France, Germany, Sweden and the UK have also announced national initiatives on AI. **Denmark**'s National Strategy for Artificial Intelligence (The Danish Government, 2019) was released in March 2019 and follows its 'Strategy for Digital Growth' (The Danish Government, 2018). This comprehensive framework lists objectives including establishing a responsible foundation for AI, providing high quality data and overall increasing investment in AI (particularly in the agriculture, energy, healthcare and transport sectors). There is a strong focus on data ethics, including responsibility, security and transparency, and recognition of the need for an ethical framework. The Danish government outlines six principles for ethical AI – self-determination, dignity, responsibility, explainability, equality and justice, and development (solutions that support ethically responsible development and use of AI in order to achieve societal progress) – and will establish a Data Ethics Council to monitor technological development in the country.

In **France**, 'AI for Humanity' was launched in March 2018 and makes commitments to support French talent, make better use of data and also establish an ethical framework on AI (AI For Humanity, 2018). President Macron has committed to ensuring transparency and fair use in AI, which will be embedded in the education system. The strategy is mainly based on the work of Cédric Villani, French mathematician and politician, whose 2018 report on AI made recommendations across economic policy, research infrastructure, employment and ethics (Villani, 2018).

Germany's AI Strategy was adopted soon after in November 2018 (Die Bundesregierung, 2018) and makes three major pledges: to make Germany a global leader in the development and use of AI, to safeguard the responsible development and use of AI, and to integrate AI in society in ethical, legal, cultural and institutional terms. Individual objectives include developing Centres of Excellence for research, the creation of 100 extra professorships for AI, establishing a German AI observatory, funding 50 flagship applications of AI to benefit the environment, developing guidelines for AI that are compatible with data protection laws, and establishing a 'Digital Work and Society Future Fund' (De.digital, 2018).

Sweden's approach to AI (Government Offices of Sweden, 2018) has less specific terms, but provides general guidance on education, research, innovation and infrastructure for AI. Recommendations include building a strong research base, collaboration between sectors and with other countries, developing efforts to prevent and manage risk and developing standards to guide the ethical use of AI. A Swedish AI Council, made up of experts from industry and academia, has also been established to develop a 'Swedish model' for AI, which they say will be sustainable, beneficial to society and promote long-term economic growth (Swedish AI Council, 2019).

The **UK** government issued the comprehensive 'AI Sector Deal' in April 2018 (GOV.UK, 2018), part of a larger 'Industrial Strategy', which sets out to increase productivity by investing in business, skills and infrastructure (GOV.UK, 2019). It pledges almost £1 billion to promote AI in the UK, along five key themes: ideas, people, infrastructure, business environment and places.

Key policies include increasing research and development investment to a total of 2.4% of GDP by 2027; investing over £400 million in maths, digital and technical education; developing a national retraining scheme to plug the skills gap and investing in digital infrastructure such as electric

vehicles and fibre networks. As well as these investment commitments, included in the deal is the creation of a 'Centre for Data Ethics and Innovation' (CDEI) to ensure the safe and ethical use of AI. First announced in the 2017 budget, the CDEI will assess the risks of AI, review regulatory and governance frameworks and advise the government and technology creators on best practice (UK Government Department for Digital, Culture, Media & Sport, 2019).

Several other European nations are well on their way to releasing national strategies. **Austria** has established a 'Robot Council' to help the Government to develop a national AI Strategy (Austrian Council on Robotics and Artificial Intelligence, 2019). A white paper prepared by the Council lays the groundwork for the strategy. The socially-focused document includes objectives to promote the responsible use of AI, develop measures to recognise and mitigate hazards, create a legal framework to protect data security, and engender a public dialogue around the use of AI (Austrian Council on Robotics and Artificial Intelligence, 2018).

Estonia has traditionally been quick to take up new technologies, AI included. In 2017, Estonia's Adviser for Digital Innovation Marten Kaevats described AI as the next step for 'e-governance' in Estonia (Plantera, 2017). Indeed, AI is already widely used by the government, which is currently devising a national AI strategy (Castellanos, 2018). The plan will reportedly consider the ethical implications of AI, alongside offering practical economic incentives and pilot programmes.

An AI task force has been established by **Italy** (Agency for Digital Italy, 2019) to identify the opportunities offered by AI and improve the quality of public services. Their white paper (Task Force on Artificial Intelligence of the Agency for Digital Italy, 2018), published in March 2018, describes ethics as the first challenge to the successful implementation of AI, stating a need to uphold the principle that AI should be at the service of the citizen and to ensure equality by using technology to address universal needs. The task force further outline challenges relating to technology development, the skills gap, data accessibility and quality, and a legal framework. It makes a total of 10 recommendations to government, which are yet to be realised by policy.

Malta, a country that has previously focused heavily on blockchain technology, has now made public its plans to develop a national AI strategy, putting Malta 'amongst the top 10 nations with a national strategy for AI' (Malta AI, 2019). A task force has been established composed of industry representatives, academics and other experts to help devise a policy for Malta that will focus on an ethical, transparent and socially-responsible AI while developing measures that garner foreign investment, which will include developing the skillset and infrastructure needed to support AI in Malta.

Poland too is working on its national AI strategy. A report recently released by the Digital Poland Foundation (2019) focuses on the AI ecosystem in Poland, as a forerunner of the national AI strategy. Although it provides a comprehensive overview of the state-of-the-art in Poland, it does not make specific recommendations for government, and makes no reference to the ethical issues surrounding AI.

Despite media reports of military-focused AI developments in **Russia** (Apps, 2019; Bershidski, 2017; Le Miere, 2017; O'Connor, 2017) the country currently has no national strategy on AI. Following the 2018 conference 'Artificial Intelligences: Problems and Solutions', the Russian Ministry of Defence released a list of policy recommendations, which include creating a state system for AI education and a national centre for AI. The latest reports suggest President Putin has set a deadline of June 15th 2019 for his government to finalise the national strategy on AI.

5.1.1. Across the EU: Public attitudes to robots and digitisation

Overall, surveys of European perspectives to AI, robotics, and advanced technology (European Commission 2012; European Commission 2017) have reflected that citizens hold a generally positive view of these developments, viewing them as a positive addition to society, the economy, and citizens' lives. However, this attitude varies by age, gender, educational level, and location and is largely dependent on one's exposure to robots and relevant information — for example, only small numbers of those surveyed actually had experience of using a robot (past or present), and those with experience were more likely to view them positively than those without.

General trends in public perception from these surveys showed that respondents were:

- Supportive of using robots and digitisation in jobs that posed risk or difficulty to humans (such as space exploration, manufacturing and the military);
- Concerned that such technology requires effective and careful management;
- Worried that automation and digitisation would bring job losses, and unsure whether it would stimulate and boost job opportunities across the EU;
- Unsupportive of using robots to care for vulnerable members of society (the elderly, ill, dependent pets, or those undergoing medical procedures);
- Worried about accessing and protecting their data and online information, and likely to have taken some form of protective action in this area (antivirus software, changed browsing behaviour);
- Unwilling to drive in a driverless car (only 22% would be happy to do this);
- Distrustful of social media, with only 7% viewing stories published on social media as 'generally trustworthy'; and
- Unlikely to view widespread use of robots as near-term, instead perceiving it to be a scenario that would occur at least 20 years in the future.

These concerns thus feature prominently in European AI initiatives, and are reflective of general opinion on the implementation of robots, AI, automation and digitisation across the spheres of life, work, health, and more.

5.2. North America

Canada was the first country in the world to launch a national AI strategy, back in March 2017. The Pan-Canadian Artificial Intelligence Strategy (Canadian Institute For Advanced Research, 2017) was established with four key goals, to: increase the number of AI researchers and graduates in Canada; establish centres of scientific excellence (in Edmonton, Montreal and Toronto); develop global thought leadership in the economic, ethical, policy and legal implications of AI; and support a national research community in AI.

A separate programme for AI and society was dedicated to the social implications of AI, led by policy-relevant working groups that publish their findings for both government and public. In collaboration with the French National Centre for Scientific Research (CNRS) and UK Research and Innovation (UKRI), the AI and society programme has recently announced a series of interdisciplinary workshops to explore issues including trust in AI, the impact of AI in the healthcare sector and how AI affects cultural diversity and expression (Canadian Institute For Advanced Research, 2019).

In the **USA**, President Trump issued an Executive Order launching the 'American AI Initiative' in February 2019 (The White House, 2019a), soon followed by the launch of a website uniting all other AI initiatives (The White House, 2019b), including AI for American Innovation, AI for American Industry, AI for the American Worker and AI for American Values. The American AI Initiative has five key areas: investing in R&D, unleashing AI resources (i.e. data and computing power), setting

governance standards, building the AI workforce and international engagement. The Department of Defence has also published its own AI strategy (US Department of Defence, 2018), with a focus on the military capabilities of AI.

In May, the US advanced this with the AI Initiative Act, which will invest $2.2 billion into developing a national AI strategy, as well as funding federal R&D. The legislation, which seeks to 'establish a coordinated Federal initiative to accelerate research and development on artificial intelligence for the economic and national security of the United States' commits to establishing a National AI Coordination Office, create AI evaluation standards and fund 5 national AI research centres. The programme will also fund the National Science Foundation to research the effects of AI on society, including the roles of data bias, privacy and accountability, and expand AI-based research efforts led by the Department of Energy (US Congress, 2019).

In June 2019, the National Artificial Intelligence Research and Development Strategic Plan was released, which builds on an earlier plan issued by the Obama administration and identifies eight strategic priorities, including making long-term investments in AI research, developing effective methods for human-AI collaboration, developing shared public datasets, evaluating AI technologies through standards and benchmarks, and understanding and addressing the ethical, legal and societal implications of AI. The document provides a coordinated strategy for AI research and development in the US (National Science & Technology Council, 2019).

5.3. Asia

Asia has in many respects led the way in AI strategy, with **Japan** being the second country to release a national initiative on AI. Released in March 2017, Japan's AI Technology Strategy (Japanese Strategic Council for AI Technology, 2017) provides an industrialisation roadmap, including priority areas in health and mobility, important with Japan's ageing population in mind. Japan envisions a three-stage development plan for AI, culminating in a completely connected AI ecosystem, working across all societal domains.

Singapore was not far behind. In May 2017, AI Singapore was launched, a five-year programme to enhance the country's capabilities in AI, with four key themes: industry and commerce, AI frameworks and testbeds, AI talent and practitioners and R&D (AI Singapore, 2017). The following year the Government of Singapore announced additional initiatives focused around the governance and ethics of AI, including establishing an Advisory Council on the Ethical Use of AI and Data, formalised in January 2019's 'Model AI Governance Framework' (Personal Data Protection Commission Singapore, 2019). The framework provides a set of guiding ethical principles, which are translated into practical measures that businesses can adopt, including how to manage risk, how to incorporate human decision making into AI and how to minimise bias in datasets.

China's economy has experienced huge growth in recent decades, making it the world's second largest economy (World Economic Forum, 2018). To catapult China to world leader in AI, the Chinese Government released the 'Next Generation AI Development Plan' in July 2017. The detailed plan outlines objectives for industrialisation, R&D, education, ethical standards and security (Foundation for Law and International Affairs, 2017). In line with Japan, it is a three-step strategy for AI development, culminating in 2030 with becoming the world's leading centre for AI innovation.

There is substantial focus on governance, with intent to develop regulations and ethical norms for AI and 'actively participate' in the global governance of this technology. Formalised under the 'Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry', the strategy iterates four main goals, to: scale-up the development of key AI products (with a focus on intelligent vehicles, service robots, medical diagnosis and video image identification

systems); significantly enhance core competencies in AI; deepen the development of smart manufacturing; and establish the foundation for an AI industry support system (New America, 2018).

In **India**, AI has the potential to add 1 trillion INR to the economy by 2035 (NITI Aayog, 2018). India's AI strategy, named AI for All, aims to utilise the benefits of AI for economic growth but also social development and 'inclusive growth', with significant focus on empowering citizens to find better quality work. The report provides 30 recommendations for the government, which include setting up Centres of Research Excellence for AI (COREs, each with their own Ethics Council), promoting employee reskilling, opening up government datasets and establishing 'Centres for Studies on Technological Sustainability'. It also establishes the concept of India as an 'AI Garage', whereby solutions developed in India can be rolled out to developing economies in the rest of the world.

Alongside them, **Taiwan** released an 'AI Action Plan' in January 2018 (AI Taiwan, 2018), focused heavily on industrial innovation, and **South Korea** announced their 'AI Information Industry Development Strategy' in May 2018 (H. Sarmah, 2019). The report on which this was based (Government of the Republic of Korea, 2016) provides fairly extensive recommendations for government, across data management, research methods, AI in government and public services, education and legal and ethical reforms.

Malaysia's Prime Minister announced plans to introduce a national AI framework back in 2017 (Abas, 2017), an extension of the existing 'Big Data Analytics Framework' and to be led by the Malaysia Digital Economy Corporation (MDEC). There has been no update from the government since 2017. More recently, **Sri Lanka**'s wealthiest businessman Dhammika Perera has called for a national AI strategy in the country, at an event held in collaboration with the Computer Society of Sri Lanka (Cassim, 2019), however there has not yet been an official pledge from the government.

In the Middle East, the **United Arab Emirates** was the first country to develop a strategy for AI, released in October 2017 and with emphasis on boosting government performance and financial resilience (UAE Government, 2018). Investment will be focused on education, transport, energy, technology and space. The ethics underlying the framework is fairly comprehensive; the Dubai AI Ethics Guidelines dictate the key principles that make AI systems fair, accountable, transparent and explainable (Smart Dubai, 2019a). There is even a self-assessment tool available to help developers of AI technology to evaluate the ethics of their system (Smart Dubai, 2019b).

World leader in technology **Israel** is yet to announce a national AI strategy. Acknowledging the global race for AI leadership, a recent report by the Israel Innovation Authority (Israel Innovation Authority, 2019) recommended that Israel develop a national AI strategy 'shared by government, academia and industry'.

5.4. Africa

Africa has taken great interest in AI; a recent white paper suggests this technology could solve some of the most pressing problems in Sub-Saharan Africa, from agricultural yields to providing secure financial services (Access Partnership, 2018). The document provides essential elements for a pan-African strategy on AI, suggesting that lack of government engagement to date has been a hindrance and encouraging African governments to take a proactive approach to AI policy. It lists laws on data privacy and security, initiatives to foster widespread adoption of the cloud, regulations to enable the use of AI for provision of public services, and adoption of international data standards as key elements of such a policy, although one is yet to emerge.

Kenya however has announced a task force on AI (and blockchain) chaired by a former Secretary in the Ministry of Information and Communication, which will offer recommendations to the government on how best to leverage these technologies (Kenyan Wallstreet, 2018). **Tunisia** too has created a task force to put together a national strategy on AI and held a workshop in 2018 entitled 'National AI Strategy: Unlocking Tunisia's capabilities potential' (ANPR, 2018).

5.5. South America

Mexico is so far the only South American nation to release an AI strategy. It includes five key actions, to: develop an adequate governance framework to promote multi-sectorial dialogue; map the needs of industry; promote Mexico's international leadership in AI; publish recommendations for public consultation; and work both with experts and the public to achieve the continuity of these efforts (México Digital, 2018). The strategy is the formalisation of a White Paper (Martinho-Truswell et al., 2018) authored by the British Embassy in Mexico, consultancy firm Oxford Insights and thinktank C Minds, with the collaboration of the Mexican Government.

The strategy emphasises the role of its citizens in Mexico's AI development and the potential of social applications of AI, such as improving healthcare and education. It also addresses the fact that 18% of all jobs in Mexico (9.8 million in total) will be affected by automation in the coming 20 years and makes a number of recommendations to improve education in computational approaches.

Other South American nations will likely follow suit if they are to keep pace with emerging markets in Asia. Recent reports suggest AI could double the size of the economy in Argentina, Brazil, Chile, Colombia and Peru (Ovanessoff and Plastino, 2017).

5.6. Australasia

Australia does not yet have a national strategy on AI. It does however have a' Digital Economy Strategy' (Australian Government, 2017) which discusses empowering Australians through 'digital skills and inclusion', listing AI as a key emerging technology. A report on 'Australia's Tech Future' further details plans for AI, including using AI to improve public services, increase administrative efficiency and improve policy development (Australian Government, 2018).

The report also details plans to develop an ethics framework with industry and academia, alongside legislative reforms to streamline the sharing and release of public sector data. The draft ethics framework (Dawson et al., 2019) is based on case studies from around the world of AI 'gone wrong' and offers eight core principles to prevent this, including fairness, accountability and the protection of privacy. It is one of the more comprehensive ethics frameworks published so far, although yet to be implemented.

Work is also ongoing to launch a national strategy in **New Zealand**, where AI has the potential to increase GDP by up to $54 billion (AI Forum New Zealand, 2018). The AI Forum of New Zealand has been set up to increase awareness and capabilities of AI in the country, bringing together public, industry, academia and Government.

Their report 'Artificial Intelligence: Shaping The Future of New Zealand' (AI Forum New Zealand, 2018) lays out a number of recommendations for the government to coordinate strategy development (i.e. to coordinate research investment and the use of AI in government services); increase awareness of AI (including conducting research into the impacts of AI on economy and society); assist AI adoption (by developing best practice resources for industry); increase the accessibility of trusted data; grow the AI talent pool (developing AI courses, including AI on the list of valued skills for immigrants); and finally to adapt to AI's effects on law, ethics and society. This

includes the recommendation to establish an AI ethics and society working group to investigate moral issues and develop guidelines for best practice in AI, aligned with international bodies.

Challenges to government adoption of AI

The World Economic Forum has, through consultation with stakeholders, identified five major roadblocks to government adoption of AI:

1. Effective use of data - Lack of understanding of data infrastructure, not implementing data governance processes (e.g. employing data officers and tools to efficiently access data).

2. Data and AI skills - It is difficult for governments, which have smaller hiring budgets than many big companies, to attract candidates with the required skills to develop first-rate AI solutions.

3. The AI ecosystem - There are many different companies operating in the AI market and it is rapidly changing. Many of the start-ups pioneering AI solutions have limited experience working with government and scaling up for large projects.

4. Legacy culture - It can be difficult to adopt transformative technology in government, where there are established practices and processes and perhaps less encouragement for employees to take risks and innovate than in the private sector.

5. Procurement mechanisms - The private sector treats algorithms as intellectual property, which may make it difficult for governments to customise them as required. Public procurement mechanisms can also be slow and complicated (e.g. extensive terms and conditions, long wait times from tender response submission to final decision).

(Torres Santeli and Gerdon, 2019)

5.7. International AI Initiatives, in addition to the EU

In addition to the EU, there are a growing number of international strategies on AI, aiming to provide a unifying framework for governments worldwide on stewardship of this new and powerful technology.

G7 Common Vision for the Future of AI

At the 2018 meeting of the G7 in Charlevoix, Canada, the leaders of the G7 (Canada, France, Germany, Italy, Japan, the United Kingdom and the United States) committed to 12 principles for AI, summarised below:

1. Promote human-centric AI and the commercial adoption of AI, and continue to advance appropriate technical, ethical and technologically neutral approaches.
2. Promote investment in R&D in AI that generates public test in new technologies and supports economic growth.
3. Support education, training and re-skilling for the workforce.
4. Support and involve underrepresented groups, including women and marginalised individuals, in the development and implementation of AI.

5. Facilitate multi-stakeholder dialogue on how to advance AI innovation to increase trust and adoption.
6. Support efforts to promote trust in AI, with particular attention to countering harmful stereotypes and fostering gender equality. Foster initiatives that promote safety and transparency.
7. Promote the use of AI by small and medium-sized enterprises.
8. Promote active labour market policies, workforce development and training programmes to develop the skills needed for new jobs.
9. Encourage investment in AI.
10. Encourage initiatives to improve digital security and develop codes of conduct.
11. Ensure the development of frameworks for privacy and data protection.
12. Support an open market environment for the free flow of data, while respecting privacy and data protection.

(G7 Canadian Presidency, 2018).

Nordic-Baltic Region Declaration on AI

The declaration signed by the Nordic-Baltic Region (comprising Denmark, Estonia, Finland, the Faroe Islands, Iceland, Latvia, Lithuania, Norway, Sweden and the Åland Islands) aims to promote the use of AI in the region, including improving the opportunities for skills development, increasing access to data and a specific policy objective to develop 'ethical and transparent guidelines, standards, principles and values' for when and how AI should be used (Nordic Co-operation, 2018).

OECD Principles on AI

On 22 May 2019, the Organisation for Economic Co-operation and Development issued its principles for AI, the first international standards agreed by governments for the responsible development of AI. They include practical policy recommendations as well as value-based principles for the 'responsible stewardship of trustworthy AI', summarised below:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should respect the rule of law, human rights, democratic values and diversity, and there should include appropriate safeguards to ensure a fair society.
- There should be transparency around AI to ensure that people understand outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles and risks should be continually assessed.
- Organisations and individuals developing, deploying or operating AI systems should be held accountable.

These principles have been agreed by the governments of the 36 OECD Member States as well as Argentina, Brazil, Colombia, Costa Rica, Peru and Romania (OECD, 2019a). The G20 human-centred AI Principles were released in June 2019 and are drawn from the OECD Principles (G20, 2019).

United Nations

The UN has several initiatives relating to AI, including:

- AI for Good Global Summit- Summits held since 2017 have focused on strategies to ensure the safe and inclusive development of AI (International Telecommunication Union, 2018a,b). The events are organised by the International Telecommunication Union, which aims to 'provide a neutral platform for government, industry and

academia to build a common understanding of the capabilities of emerging AI technologies and consequent needs for technical standardisation and policy guidance.'

- UNICRI Centre for AI and Robotics - The UN Interregional Crime and Justice Research Institute (UNICRI) launched a programme on AI and Robotics in 2015 and will be opening a centre dedicated to these topics in The Hague (UNICRI, 2019).

- UNESCO Report on Robotics Ethics - The UNESCO World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) has authored a report on 'Robotics Ethics', which deals with the ethical challenges of robots in society and provides ethical principles and values, and a technology-based ethical framework (COMEST, 2017).

World Economic Forum

The World Economic Forum (WEF) formed a Global AI Council in May 2019, co-chaired by speech recognition developer Kai-Fu Lee, previously of Apple, Microsoft and Google, and current President of Microsoft Bradford Smith. One of six 'Fourth Industrial Revolution' councils, the Global AI Council will develop policy guidance and address governance gaps, in order to develop a common understanding among countries of best practice in AI policy (World Economic Forum, 2019a).

In October 2019, they released a framework for developing a national AI strategy to guide governments that are yet to develop or are currently developing a national strategy for AI. The WEF describe it as a way to create a 'minimum viable' AI strategy and includes four main stages:

1) Assess long-term strategic priorities
2) Set national goals and targets
3) Create plans for essential strategic elements
4) Develop the implementation plan

The WEF has also announced plans to develop an 'AI toolkit' to help businesses to best implement AI and to create their own ethics councils, which will be released at 2020's Davos conference (Vanian, 2019).

5.8. Government Readiness for AI

A report commissioned by Canada's International Development Research Centre (Oxford Insights, 2019) evaluated the 'AI readiness' of governments around the globe in 2019, using a range of data including not only the presence of a national AI strategy, but also data protection laws, statistics on AI startups and technology skills.

Singapore was ranked number 1 in their estimation, with Japan as the only other Asian nation in the top 10 (Table 3). Sixty percent of countries in the top 10 were European, with the remainder from North America.

The strong European representation in this analysis is reflective of the value of the unifying EU framework, as well as Europe's economic power. The analysis also praises the policy strategies of individual European nations, which, importantly, have been developed in a culture of collaboration. Examples of this collaborative approach include the EU Declaration of Cooperation on AI (European Commission, 2018d), in which Member States agreed to cooperate on boosting Europe's capacity in AI, and individual partnerships between Member States, such as that of Finland, Estonia and Sweden, working together to trial new applications of AI.

Table 3: Top 10 rankings for Government AI Readiness 2018/19. Source: Oxford Insights, 2019.

Rank	Country	Score
1	Singapore	9.19
2	United Kingdom	9.07
3	Germany	8.81
4	USA	8.80
5	Finland	8.77
6	Sweden	8.67
6	Canada	8.67
8	France	8.61
9	Denmark	8.60
10	Japan	8.58

Singapore ranked highest of all nations while Japan, the second country in the world to release a national strategy on AI, ranked 10th. China's position as 21st in the global rankings is expected to improve next year as its investments in AI begin to pay off. Progress in Asia overall has been unbalanced, with two countries in the region also ranking in the bottom ten worldwide, reflecting the income inequality in the region.

Despite the comparatively slow development of their national strategy, the USA ranked 4th, with Canada not far behind. Both nations are supported by their strong economies, highly skilled workforces, private sector innovation and abundance of data, to a level at which regions missing from the top 10 – Africa, South America and Australasia – are unable to compete.

This framework provides a highly useful metric by which to assess the ability of governments to capitalise on AI's potential in the coming years. What this analysis does not consider however is how robustly each nation is considering the moral and ethical issues surrounding the use of AI, which we will explore below.

6. Emerging Themes

Our review of the literature on the ethical issues surrounding AI and intelligent robots highlights a wide range of potential impacts, including in the social, psychological, financial, legal and environmental domains. These are bound up with issues of trust and are tackled in different ways by the emerging ethical initiatives. Standards and regulation are also beginning to develop that go some way to addressing these concerns. However, the focus of many existing strategies on AI is on enabling technology development and, while ethical issues are addressed, notable gaps can be identified.

6.1. Addressing ethical issues through national and international strategies

There are several themes shared by the various national strategies on AI, among which **industrialisation** and **productivity** perhaps rank highest. All countries have some sort of industrial strategy for AI, and this is particularly prominent in the emerging economies of Southeast Asia. Most of the strategies make reference to the importance of AI for business competitiveness and several, including those of Germany, South Korea, Taiwan and the UK, announce extra funding and specialised incubators for AI-focused start-ups.

Whether in the private or public sector, the importance of **research** and development is also universally recognised, with almost all strategies pledging enhanced funding for research and many to establish 'centres of excellence' entirely dedicated to AI research, including strategies from Canada, Germany and India.

Essential to developing a strong research effort is talent, and so investing in **people** and education also features heavily in most strategies. The UK has announced 'Turing Fellowships' to fund new academics exploring computational approaches, while Germany has provided for at least an extra 100 professors working on AI – both under the umbrella of the EU commitment to train, attract and retain talent. In Asia, South Korea has committed to developing six new graduate programmes to train a total of 5,000 AI specialists, while Taiwan has committed to training double that number by 2021.

Most of the strategies also consider the impact the AI revolution will have on the non-technology literate workforce, who may be the first to lose their jobs to automation. Although this crosses over into ethical considerations, several of the strategies make practical commitments to **re-training** programmes to help those affected to find new work. This is a key objective in the EU plan (objective 2.4: 'adapting our learning and training programmes and systems to better prepare our society for AI'), and therefore the plans of its Member States. The UK for example will initiate an > €70 million re-training scheme to help people gain digital skills and Germany has revealed a similar 'National Further Training Strategy'. Naturally, those countries most in need of re-training have the least funding available for it. Mexico's strategy however emphasises the importance of computational thinking and mathematics in lifelong teaching, including to help its citizens retrain, while India pledges to promote informal training institutions and create financial incentives for reskilling of employees. Other strategies however suggest re-training is the responsibility of individual businesses and do not allocate separate funding for it.

Collaboration between sectors and countries is another common thread, yet interpreted differently by different countries. India's approach for example is one of sharing; the 'AI Garage' concept named in their strategy means AI-based solutions developed in India will be rolled out to developing economies facing similar issues. Conversely, the US Executive Order on AI sets out to

'promote an international environment that supports American AI' while also protecting the nation's technological advantage against 'foreign adversaries'. Naturally, the strategies of EU Member States display an inclination for cross-border collaboration. Sweden for example states a need to develop partnerships and collaborations with other countries 'especially within the EU', while Denmark's strategy also emphasises close cooperation with other European countries.

The democratisation of technology has the potential to reduce inequalities in society, and **inclusion** and **social development** are important goals for many national AI initiatives, particularly those of developing economies. India's strategy discusses AI for 'greater good', focusing on the possibilities for better access to healthcare, economic growth for groups previously excluded from formal financial products, and using data to aid small-scale farmers. Mexico's strategy lists inclusion as one of its five major goals, which includes aims to democratise productivity and promote gender equality. France too aims for an AI that 'supports inclusivity', striving for policies that reduce both social and economic inequalities.

Determining who is **responsible** for the actions and behaviour of AI is highly important, and challenging in both moral and legal senses. Currently, AI is most likely considered to be the legal responsibility of a relevant human actor – a tool in the hands of a developer, user, vendor, and so on. However, this framework does not account for the unique challenges brought by AI, and many grey areas exist. As just one example, as a machine learns and evolves to become different to its initial programming over many iterations, it may become more difficult to assign responsibility for its behaviour to the programmer. Similarly, if a user or vendor is not adequately briefed on the limitations of an AI agent, then it may not be possible to hold them responsible. Without proving that an AI agent intended to commit a crime (*mens rea*) and can act voluntarily, both of which are controversial concepts, then it may not be possible to deem an AI agent responsible and liable for its own actions.

6.2. Addressing the governance challenges posed by AI

There are currently two major international frameworks for the governance of AI: that of the EU (see Section 5.1) and the Organisation for Economic Co-operation and Development (OECD).

The OECD launched a set of principles for AI in May 2019 (OECD, 2019a) which were at that time adopted by 42 countries. The OECD framework offers five fundamental principles for the operation of AI (see section 5.1.1) as well as accompanying practical recommendations for governments to achieve them. The G20 soon after adopted its own, human-centred AI principles, drawn from (and essentially an abridged version of) those of the OECD (G20, 2019).

The OECD Principles have also been backed by the European Commission, which has its own strategy on AI since April 2018 (European Commission, 2018b). The EU framework includes comprehensive plans for investment, but also makes preparations for complex socio-economic changes and is complemented by a separate set of ethics guidelines (European Commission High-Level Expert Group on AI, 2019a).

Gaps in AI frameworks

These frameworks address the moral and ethical dilemmas identified in this report to varying extents, with some notable gaps. Regarding **environmental concerns** (Section 2.5), while the OECD makes reference to developing AI that brings positive outcomes for the planet, including protecting natural environments, the document does not suggest ways to achieve this, nor does it mention any specific environmental challenges to be considered.

The EU Communication on AI does not discuss the environment. However, its accompanying ethics guidelines are founded on the principle of prevention of harm, which includes harm to the natural

environment and all living beings. Societal and environmental well-being (including sustainability and 'environmental friendliness') is one of the EU's requirements for trustworthy AI and its assessment list includes explicit consideration of risks to the environment or to animals. Particular examples are also given on how to achieve this (e.g. critical assessment of resource use and energy consumption throughout the supply chain).

Impacts on human **psychology**, including how people interact with AI and subsequent effects on how people interact with each other, could be further addressed in the frameworks. The psychosocial impact of AI is not considered by the OECD Principles or the EU Communication. However, the EU requirement for societal well-being to be considered does address 'social impact', which includes possible changes to social relationships and loss of social skills. The guidelines state that such effects must 'be carefully monitored and considered' and that AI interacting with humans must clearly signal that its social interaction is simulated. However, more specific consideration could be given to human-robot relationships or more complex effects on the human psyche, such as those outlined above (Section 2.2).

While both frameworks capably address changes to the **labour market** (Section 2.1.1), attention to more nuanced factors, including the potential for AI to drive **inequalities** (2.1.2) and **bias** (2.1.4), is more limited. The OECD's first principle of inclusive growth, sustainable development and well-being states that AI should be developed in a way that reduces 'economic, social, gender and other inequalities'. This is also covered to a degree by the second OECD principle, which states that AI systems should respect diversity and include safeguards to ensure a fair society, however detail on how this can be achieved is lacking.

The EU ethics guidelines are more comprehensive on this point and include diversity, non-discrimination and fairness as a separate requirement. The guidelines elaborate that equality is a fundamental basis for trustworthy AI and state that AI should be trained on data which is representative of different groups in order to prevent biased outputs. The guidelines include additional recommendations on the avoidance of unfair bias.

Both frameworks include **human rights** and **democratic values** (Sections 2.1.3, 2.1.5) as key tenets. This includes **privacy**, which is one of the OECD's human-centred values and a key requirement of the EU ethics guidelines, which elaborates on the importance of data governance and data access rules. Issues concerning privacy are also covered by existing OECD data protection guidelines (OECD, 2013).

The implications of AI for **democracy** (Section 2.1.5) are only briefly mentioned by the OECD, with no discussion of the particular issues facing governments at the present time, such as Deepfake or the manipulation of opinion through targeted news stories. Threats to democracy are not mentioned at all in the EU Communication, although society and democracy is a key theme in the associated ethics guidelines, which state that AI systems should serve to maintain democracy and not undermine 'democratic processes, human deliberation or democratic voting systems.'

These issues form part of a bigger question surrounding changes to the **legal system** (Section 2.4) that may be necessary in the AI age, including important questions around liability for misconduct involving AI. The issue of liability is explicitly addressed by the EU in both its Communication and ethics guidelines. Ensuring an appropriate legal framework is a key requirement of the EU Communication on AI, which includes guidance on product liability and an exploration of safety and security issues (including criminal use). The accompanying ethics guidelines also suitably handle this issue, including providing guidance for developers on how to ensure legal compliance. Relevant changes to regulation are further addressed in the recent AI Policy and Investment Recommendations (European Commission High-Level Expert Group on AI, 2019b), which explore potential changes to current EU laws and the need for new regulatory powers.

The OECD principles are more limited on this point. While they provide guidance for governments to create an 'enabling policy environment' for AI, including a recommendation to review and adapt

regulatory frameworks, this is stated to be for the purpose of encouraging 'innovation and competition' and does not address the issue of liability for AI-assisted crime.

These questions could also come under the issue of **accountability** (2.6.4) however, which is adequately addressed by both frameworks. The OECD lists accountability as a key principle and states that 'organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning' (OECD, 2019a). It is likewise a core principle of the EU ethics guidelines, which provides more than 10 conditions for accountability in its assessment list for trustworthy AI.

Many of the aforementioned issues are ultimately important for building **trust** in AI (Section 2.6), which also requires AI to be fair (2.6.2) and transparent (2.6.3). These issues are at the foundation of the EU ethics guidelines where they are dealt with in great detail. The OECD also states that AI systems should ensure a 'fair and just society'. Transparency and explainability is a core principle for the OECD, with strong emphasis on the fact that people should be able to understand and challenge AI systems. The OECD Principles offer less context on these issues and do not consider practical means of ensuring this (e.g. audits of algorithms), which are considered by the EU ethics guidelines. The ethics guidelines also consider the need for human oversight (including discussion of the human-in-the-loop approach and the need for a 'stop button', neither of which are mentioned by the OECD principles).

Finally, although both acknowledge the beneficial use of AI in **finance** (Section 2.3), neither framework adequately addresses potential negative impacts on the financial system, either through accidental harm or malicious activity. The potential for AI-assisted financial crime is an important one and currently unaddressed by any international framework. However, the G7 has recently voiced concerns about digital currencies and various other new financial products being developed (Reuters, 2019), which suggests that regulatory changes in this regard are afoot.

7. Summary

What this report makes clear is the diversity and complexity of the ethical concerns arising from the development of artificial intelligence; from large scale issues such job losses from automation, degradation of the environment and furthering inequalities, to more personal moral quandaries such as how AI may affect our privacy, our ability to judge what is real, and our personal relationships.

What is also clear is that there are various **approaches to ethics**. Robust ethical principles are essential in the future of this rapidly developing technology, but not all countries understand ethics in the same way. There are a number of independent ethical initiatives for AI, such as Germany's Institute for Ethics in AI, funded by Facebook, and the private donor-funded Future of Life Institute in the US. An increasing number of governments are also developing national AI strategies, with their own ethics components. A number of countries have committed to creating AI ethics councils, including Germany, the UK, India, Singapore and Mexico. The UAE has also prioritised ethics in its national strategy, by developing an 'Ethical AI Toolkit' and self-assessment tool for developers, while several others give only passing reference; ethics is almost completely left out by Japan, South Korea and Taiwan.

Our assessment shows that the vast majority of ethical issues identified here are also addressed in some form by at least one of the current international frameworks; the EU Communication (supplemented by separate ethics guidelines) and the OECD Principles on AI.

The current frameworks address the major ethical concerns and make recommendations for governments to manage them, but **notable gaps** exist. These include environmental impacts, including increased energy consumption associated with AI data processing and manufacture, and inequality arising from unequal distribution of benefits and potential exploitation of workers. Policy options relating to environmental impacts include providing a stronger mandate for sustainability and ecological responsibility; requiring energy use to be monitored, and publication of carbon footprints; and potentially policies that direct technology innovation towards urgent environmental priorities. In the case of inequality, options include declaring AI as a public, rather than private, good. This would require changes to cultural norms and new strategies to help navigate a transition to an AI-driven economy. Setting minimum standards for corporate social responsibility reporting would encourage larger, transnational corporations to clearly show how they are sharing the benefits of AI. Economic policies may be required to support workers displaced by AI; such policies should focus on those at most risk of being left behind and might include policies designed to create support structures for precarious workers. It will be important for future iterations of these frameworks to address these and other gaps in order to adequately prepare for the full implications of an AI future. In addition, to clarify the issue of responsibility pertaining to AI behaviour, moral and legislative frameworks will require updating alongside the development of the technology itself.

Governments also need to develop new, up-to-date forms of **technology assessment** – allowing them to understand such technologies deeply while they can still be shaped, such as the Accountability Office's Technology Assessment Unit in the USA or the European Foresight platform (http://www.foresight-platform.eu/). New forms of technology assessment TA should include processes of Ethical Risk Assessment, such as the one set out in BS8611, and other forms of ethical evaluation currently being drafted in the IEEE Standards Association P7000 series of ethical standards; P7001 for instance sets out a method for measuring the transparency of an AI.

There is a clear need for the development of viable and applicable **legislation and policies** that will face the multifaceted challenges associated with AI, including potential breaches of fundamental ethical principles. Policy makers are in the valuable position of being able to develop policy that actively shapes the development of AI and as data-driven and machine-learning approaches begin

to take increasing roles in society, thoughtful and detailed strategies on how to share benefits and achieve the best possible outcomes, while effectively managing risk, will be essential.

As well as the very encouraging progress made in policy so far, this report also reveals a concerning **disparity** between regions. Successful AI development requires substantial investment, and as automation and intelligent machines begin to drive government processes, there is a real risk that lower income countries – those nations of the Global South – will be left behind. It is incumbent upon policymakers therefore to try to ensure that AI does not widen global inequalities. This could include **data sharing** and collaborative approaches, such as India's promise to share its AI solutions with other developing countries, and efforts to make teaching on computational approaches a fundamental part of education, available to all.

To return to our main theme, **ethical considerations** must also be a critical component of any policy on AI. It speaks volumes that the nation ranked highest in the 2019 Government AI Readiness Index has prioritised ethics so strongly in their national AI Strategy. Singapore is one of a few governments to create an AI Ethics Council and has incorporated a range of ethical considerations into its policy. Addressing ethical concerns is also the first key point in the World Economic Forum's framework for developing a national AI strategy. So, aside from any potential moral obligations, it seems unlikely that governments that do not take ethics seriously will be able to succeed in the competitive global forum.

8. Appendix

Building ethical robots

In the future it's very likely that intelligent machines will have to make decisions that affect human safety, psychology and society. For example, a search and rescue robot should be able to 'choose' the victims to assist first after an earthquake; an autonomous car should be able to 'choose' what or who to crash into when an accident cannot be avoided; a home-care robot should be able to balance its user's privacy and their nursing needs. But how do we integrate societal, legal and moral values into technological developments in AI? How can we program machines to make ethical decisions - to what extent can ethical considerations even be written in a language that computers understand?

Devising a method for integrating ethics into the design of AI has become a main focus of research over the last few years. Approaches towards moral decision making generally fall into two camps, 'top-down' and 'bottom-up' approaches (Allen et al., 2005). Top-down approaches involve explicitly programming moral rules and decisions into artificial agents, such as 'thou shalt not kill'. Bottom up approaches, on the other hand, involve developing systems that can implicitly learn to distinguish between moral and immoral behaviours.

Bottom-up approaches
Bottom up approaches involve allowing robots to learn ethics independently of humans, for instance by using machine learning. Santos-Lang (2002) points out that this is a better approach, as humans themselves continuously learn to be ethical. An advantage of this is that most of the work is done by the machine itself, which avoids the robot being influenced by the designers' biases. However the downside is that machines could demonstrate unintended behaviour that deviates from the desired goal. For example, if a robot was programmed to 'choose behaviour that leads to the most happiness', the machine may discover that it can more quickly reach its goal of maximising happiness by first increasing its own learning efficiency, 'temporarily' shifting away from the original goal. Because of the shift, the machine may even choose behaviours that temporarily reduce happiness, if these behaviours were to ultimately help it achieve its goal. For example a machine could try to rob, lie and kill, in order to become an ethical paragon later.

Top-down approaches
Top-down approaches involve programming agents with strict rules that they should follow in given circumstances. For example, in self-driving cars a vehicle could be programmed with the command 'you shall not drive faster than 130 km/h on the highway'. The problem with top down approaches is that they require deciding which moral theories ought to be applied. Examples of competing moral theories include utilitarian ethics, deontological ethics and the commensal view and the Doctrine of Double Effect.

Utilitarianism is based on the notion that the morality of an action should be judged by its consequences. In other words, an action is judged to be morally right if its consequences lead to the greater good. Different utilitarian theories vary in terms of the definition of the 'good' they aim to maximise. For example, Bentham (1789) proposed that a moral agent should aim to maximise the total happiness of a population of people.

Deontological (duty-based) ethics, on the other hand argues that actions should be judged not on the basis of their expected outcomes, but on what people do. Duty-based ethics teaches that actions are right or wrong regardless of the good or bad consequences that may be produced. Under this form of ethics you can't justify an action by showing that it produced good consequences.

Sometimes different moral theories can directly contradict each other. For example, in the case of a self-driving car that has to decide whether to swerve to avoid animals in its path. Under the commensal view, animal lives are treated as if they are worth some small fraction of what human lives are worth, and so the car would swerve if there was a low chance of causing harm to a human (Bogosian, 2017). However, the incommensal view would never allow humans to be placed at additional risk of fatality in order to save an animal. Since this view fundamentally rejects the assumptions of the other, and holds that no tradeoff is permissible, there is no obvious 'halfway point' where the competing principles can meet.

Bonnemains et al. (2018) describe a dilemma where a drone programmed to take out a missile threatening an allied ammo factory is suddenly alerted to a second threat - a missile heading towards some civilians. The drone must decide whether to continue its original mission, or take out the new missile in order to save the civilians. The decision outcome is different depending on whether you use utilitarianism, deontological ethics and the Doctrine of Double Effect - a theory which states that if doing something morally good has a morally bad side-effect, it's ethically okay to do it providing that the bad side-effect wasn't intended.

Some of the theories are unable to solve the problem. For instance, from a deontological perspective both decisions are valid, as they both arise from good intentions. In the case of utilitarian ethics, without any information about the number of civilians that are in danger, or the value of the strategic factory, it would be difficult for a drone to reach a decision. In order to follow the utilitarian doctrine and make a decision that maximised a 'good outcome', an artificial agent would need to identify all possible consequences of a decision, from all parties' perspectives, before making a judgement about which consequence is preferable. This would be impossible in the field. Another issue is how should a drone decide which outcomes it prefers when this is a subjective judgement? What is Good? Giving an answer to this broad philosophical issue is hardly possible for an autonomous agent, or the person programming it.

Under the Doctrine of Double Effect the drone would not be allowed to intercept the missile and save the civilians, as the bad side effect (the destruction of the drone itself) would be a means to ensuring the good effect (saving the humans). It would therefore continue to pursue its original goal and destroy the launcher, letting the civilians die.

If philosophers cannot agree on the merits of various theories, companies, governments, and researchers will find it even more difficult to decide which system to use for artificial agents (Bogosion, 2017). People's personal moral judgements can also differ widely when faced with moral dilemmas (Greene et al., 2001), particularly when they are considering politicised issues such as racial fairness and economic inequality. Bogosian (2017) argues that instead, we should design machines to be fundamentally uncertain about morality.

REFERENCES

Abas, A. (2017). *Najib unveils Malaysia's digital 'to-do list' to propel digital initiatives implementation.* [online] Nst.com.my. Available from: https://www.nst.com.my/news/nation/2017/10/292784/najib-unveils-malaysias-digital-do-list-propel-digital-initiatives [Accessed 8 May 2019].

Access Partnership and the University of Pretoria (2018). *Artificial Intelligence for Africa: An Opportunity for Growth, Development and Democratisation*. Available from: https://www.up.ac.za/media/shared/7/ZP_Files/ai-for-africa.zp165664.pdf

Acemoglu, D. and Restrepo, P. (2018) Low-skill and high-skill automation. *Journal of Human Capital*, 2018, vol. 12, no. 2.

Agency for Digital Italy (2019). *Artificial Intelligence task force.* [online] IA-Gov. Available from: https://ia.italia.it/en/ [Accessed 10 May 2019].

AI4All (2019). *What we do* [online] Available from: http://ai-4-all.org [Accessed 11/03/2019].

AI For Humanity (2018). *AI for humanity: French Strategy for Artificial Intelligence* [online] Available from: https://www.aiforhumanity.fr/en/ [Accessed 10 May 2019].

AI Forum New Zealand (2018). *Artificial Intelligence: Shaping a Future New Zealand*. Available from: https://aiforum.org.nz/wp-content/uploads/2018/07/AI-Report-2018_web-version.pdf

AI Now Insitute, (2018). *AI Now Report.* AI Now Institute, New York University. Available from: https://ainowinstitute.org/AI_Now_2018_Report.pdf

AI Singapore. (2018). *AI Singapore*. [online] Available from: https://www.aisingapore.org [Accessed 26 Apr. 2019].

AI Taiwan. (2019). *AI Taiwan*. [online] Available from: https://ai.taiwan.gov.tw [Accessed 28 Apr. 2019].

Allen, C., Smit, I., and Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*. doi:10.1007/s10676-006-0004-4.

Allen, G,. and Chan, T,. (2017). *Artificial Intelligence and National Security*. Available from: https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf

Amoroso, D., and Tamburrini, G. (2018). The Ethical and Legal Case Against Autonomy in Weapons Systems. *Global Jurist* 18 (1), DOI: 10.1515/gj-2017-0012.

Anderson, J. M., Heaton, P. and = Carroll, S. J. (2010). *The U.S. Experience with No-Fault Automobile Insurance: A Retrospective.* Santa Monica, CA: RAND Corporation. Available from: https://www.rand.org/pubs/monographs/MG860.html.

ANPR (2018). *National AI Strategy: Unlocking Tunisia's capabilities potential* [online] Available from: http://www.anpr.tn/national-ai-strategy-unlocking-tunisias-capabilities-potential/. [Accessed 6 May 2019].

Apps, P. (2019). *Commentary: Are China, Russia winning the AI arms race?* [online] U.S. Available from: https://www.reuters.com/article/us-apps-ai-commentary/commentary-are-china-russia-winning-the-ai-arms-race-idUSKCN1P91NM.

Arnold, T., and Scheutz, M. (2018). The 'big red button' is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology*. 20 (1), 59–69.

Asaro, P. (2012). On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making. *International Review of the Red Cross.* 94 (886), 687-703.

Atabekov, A. and Yastrebov, O. (2018) Legal status of Artificial Intelligence: Legislation on the move. European Research Studies Journal Volume XXI, Issue 4, 2018 pp. 773 - 782

Australian Government (2017). *The Digital Economy: Opening Up The Conversation*. Department of Industry, Innovation and Science. Available from: https://www.archive.industry.gov.au/innovation/Digital-Economy/Documents/Digital-Economy-Strategy-Consultation-Paper.pdf

Australian Government (2018). *Australia's Tech Future*. Department of Industry, Innovation and Science. Available from: https://www.industry.gov.au/sites/default/files/2018-12/australias-tech-future.pdf

Austrian Council on Robotics and Artificial Intelligence (2018). Die Zukunft Österreichs mit Robotik und Künstlicher Intelligenz positiv gestalten. *White Paper des Österreichischen Rats für Robotik und Künstliche Intelligenz.* Available from: https://www.acrai.at/wp-content/uploads/2019/04/ACRAI_whitebook_online_2018-1.pdf

Austrian Council on Robotics and Artifical Intelligence (2019). *Österreichischer Rat für Robotik und Künstliche Intelligenz.* [online] Available from: https://www.acrai.at/ [Accessed 10 May 2019].

Autor, D. H. (2015). Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives*. 29(3), 3–30.

Bandyopadhyay, A., and Hazra, A. (2017). A comparative study of classifier performance on spatial and temporal features of handwritten behavioural data. In A. Basu, S. Das, P. Horain, and S. Bhattacharya (eds.). (2016) *Intelligent Human Computer Interaction: 8th International Conference*, IHCI 2016, Pilani, IndiaCham: Springer International Publishing, 111–121.

Baron, E. (2017). Robot surgery firm from Sunnyvale facing lawsuits, reports of death and injury. *Mercury News.* Available from: https://www.mercurynews.com/2017/10/22/robot-surgery-firm-from-sunnyvale-facing-lawsuits-reports-of-death-and-injury/

Bartlett, J. (2018) How AI could kill off democracy. *New Statesman*. Available from: https://www.newstatesman.com/science-tech/technology/2018/08/how-ai-could-kill-democracy-0

BBC News (2017). Singapore to use driverless buses 'from 2022'. *BBC.* Available from: https://www.bbc.co.uk/news/business-42090987

BBC News. (2018). Addison Lee plans self-driving taxis by 2021. BBC. Available from: https://www.bbc.co.uk/news/business-45935000

BBC News. (2019a). Autonomous shuttle to be tested in New York City. BBC. Available from: https://www.bbc.co.uk/news/technology-47668886

BBC News. (2019b). Uber 'not criminally liable for self-driving death. BBC. Available from: https://www.bbc.co.uk/news/technology-47468391

Beane, M. (2018). Young doctors struggle to learn robotic surgery – so they are practicing in the shadows. *The Conversation.* Available from: https://theconversation.com/young-doctors-struggle-to-learn-robotic-surgery-so-they-are-practicing-in-the-shadows-89646

Berger, S. (2019). Vaginal mesh has caused health problems in many women, even as some surgeons vouch for its safety and efficacy. *The Washington Post.* Available from:

https://www.washingtonpost.com/national/health-science/vaginal-mesh-has-caused-health-problems-in-many-women-even-as-some-surgeons-vouch-for-its-safety-and-efficacy/2019/01/18/1c4a2332-ff0f-11e8-ad40-cdfd0e0dd65a_story.html?noredirect=on&utm_term=.9bece54e4228

Bershidsky, L (2017). *Elon Musk warns battle for AI supremacy will spark Third World War. The Independent.* [online] Available from: https://www.independent.co.uk/life-style/gadgets-and-tech/news/elon-musk-ai-artificial-intelligence-world-war-three-russia-china-robots-cyber-warfare-replicants-a7931981.html

Bentham, J. (1789). *A Fragment of Government and an Introduction to the Principles of Morals and Legislation*, London.

Biavaschi, C., Eichhorst, W., Giulietti, C., Kendzia, M., Muravyev, A., Pieters, J., Rodriguez-Planas, N., Schmidl, R., and Zimmermann, K. (2013). Youth Unemployment and Vocational Training. *World Development Report*. World Bank.

Bilge, L., Strufe, T., Balzarotti, D., Kirda, K., and Antipolis, S. (2009). All your contacts are belong to us: Automated identity theft attacks on social networks, In WWW '09: *Proceedings of the 18th international conference on World Wide Web*, *WWW '09, April 20-24, 2009*, *Madrid, Spain.* New York, NY, USA. pp. 551–560.

Bogosian, K. (2017) Implementation of Moral Uncertainty in Intelligent Machines. *Minds & Machines* 27 (591).

Bonnemains, V., Saurel, C. & Tessier, C. (2018) Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology.* 20 (41). https://doi.org/10.1007/s10676-018-9444-x

Borenstein, J.and Arkin, R.C. (2019) *Robots, Ethics, and Intimacy: The Need for Scientific Research*. Available from: https://www.cc.gatech.edu/ai/robot-lab/online-publications/RobotsEthicsIntimacy-IACAP.pdf

Bradshaw, S., and Howard, P. (2017) Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. In Woolley, S. and Howard, P. N. (Eds.) (2017) *Working Paper: Project on Computational Propaganda,*. Oxford, UK. Available from:http://comprop.oii.ox.ac.uk/..

Bradshaw, T. (2018) Uber halts self-driving car tests after pedestrian is killed. *Financial Times.* 19 March, 2018. Available at: https://www.ft.com/content/1e2a73d6-2b9e-11e8-9b4b-bc4b9f08f381

British Standard BS 8611 (2016) *Guide to the Ethical Design of Robots and Robotic Systems* https://shop.bsigroup.com/ProductDetail?pid=000000000030320089

Brundage, M. And Bryson, J. (2016) Smart Policies for Artificial Intelligence.

Brynjolfsson, E., and McAfee, A (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York, W. W. Norton & Company..

Bryson, J,. (2018) Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20 (1). 15–26

Bryson, J. J. (2019). The Past Decade and Future of AI's Impact on Society. In Baddeley, M., Castells, M., Guiora, A., Chau, N., Eichengreen, B., López, R., Kanbur, R. and Burkett, V. (2019) *Towards a New Enlightenment? A Transcendent Decade*. Madrid, Turner.

Burgmann, T. (2016). There's a cure for that: Canadian doctor pushes for more wearable technology. *Global News Canada*. Available from: https://globalnews.ca/news/2787549/theres-a-cure-for-that-canadian-doctor-pushes-for-more-wearable-technology/

Cadwalladr, C. (2017a). Revealed: How US billionaire helped to back Brexit. *The Guardian*.

Cadwalladr, C. (2017b). Robert Mercer: The big data billionaire waging war on mainstream media. *The Guardian*.

Calder,S. (2018). Driverless buses and taxis to be launched in Britain by 2021. *The Independent*. Available from: https://www.independent.co.uk/travel/news-and-advice/self-driving-buses-driverless-cars-edinburgh-fife-forth-bridge-london-greenwich-a8647926.html

Cannon, J. (2018). Starsky Robotics completes first known fully autonomous run without a driver in cab. *Commercial Carrier Journal*. Available from: https://www.ccjdigital.com/starsky-robotics-autonomous-run-without-driver/

Caplan, R., Donovan, J., Hanson, L. and Matthews, J. (2018). *Algorithmic Accountability: A Primer*. New York, Data & Society.

Cassim, N. (2019). Dhammika makes strong case for national strategy for AI. [online] *Financial Times*. Available from: http://www.ft.lk/top-story/Dhammika-makes-strong-case-for-national-strategy-for-AI/26-674868 [Accessed 10 May 2019].

Castellanos, S. (2018). Estonia's CIO Tackles AI Strategy For Government. [online] *WSJ*. Available from: https://blogs.wsj.com/cio/2018/11/28/estonias-cio-tackles-ai-strategy-for-government/ [Accessed 10 May 2019].

Canadian Institute For Advanced Research (2017) *Pan-Canadian Artificial Intelligence Strategy*. [online] Available from: https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy. [Accessed 4 April 2019].

Canadian Institute For Advanced Research (2019). *AI & Society Workshops: Call Two*. [online] Available from: https://www.cifar.ca/ai/ai-society/workshops-call-two [Accessed 10 May 2019].

CDEI (2019). '*The Centre for Data Ethics and Innovation (CDEI) 2019/ 20 Work Programme*' [online] Available from: https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-cdei-2019-20-work-programme/the-centre-for-data-ethics-and-innovation-cdei-2019-20-work-programme [Accessed 3 May 2019].

Chantler, A., & Broadhurst, R. (2006). Social engineering and crime prevention in cyberspace. *Technical report.,* Justice, Queensland University of Technology.

Chen, A. (2017) 'The Human Toll of Protecting the Internet from the Worst of Humanity'. The New Yorker.

Chesney, R., & Citron, D. (2018). Deep fakes: A looming crisis for national security, democracy and privacy? *Lawfare*.

Christakis, N.A (2019) How AI Will Rewire Us. *The Atlantic Magazine, April 2019 Issue*. Available from: https://www.theatlantic.com/magazine/archive/2019/04/robots-human-relationships/583204/

Christakis, N.A & Shirado, H. (2017) Locally Noisy Autonomous Agents Improve Global Human Coordination in Network Experiments. *Nature*. 545(7654), 370–374.

Citron, D. K., & Pasquale, F. A. (2014). The scored society: due process for automated predictions. *Washington Law Review*, 89, 1–33.

CNN. (2018). Self-driving electric bus propels Swiss town into the future. *CNN.* Available from: https://edition.cnn.com/2018/06/27/sport/trapeze-self-driving-autonomous-electric-bus-switzerland-spt-intl/index.html

COMEST (2017). *Report of COMEST on Robotics Ethics*. UNESCO. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000253952

Conn, A. (2018) AI Should Provide a Shared Benefit for as Many People as Possible, Future of Life Institute, 10 Jan 2018 [online] Available at: https://futureoflife.org/2018/01/10/shared-benefit-principle/ [Accessed 12 Aug. 2019].

Corbe-Davies, S., Pierson, S., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of KDD '17*, Halifax, NS, Canada, August 13-17, 2017, 10 pages. DOI: 10.1145/3097983.3098095

Council of Europe (2019a). Ad Hoc Committee on Artificial Intelligence – CAHAI. [online] Available at: https://www.coe.int/en/web/artificial-intelligence/cahai [Accessed 29 Oct. 2019].

Council of Europe (2019b). Council of Europe's Work in progress. [online] Available at: https://www.coe.int/en/web/artificial-intelligence/work-in-progress [Accessed 29 Oct. 2019].

Consultative Committee of the Convention for the Protection of Individuals with regard to the Processing of Personal Data (2019) Guidelines on Artifical Intelligence and Data Protection. Available from: https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8

Cummings M. (2004). Automation bias in intelligent time critical decision support systems. In AIAA: *1st Intelligent Systems Technical Conference*. *AIAA 2004, 20-22 September 2004*, *Chicago, Illinois.* pp. 6313.

Curtis, J. (2016). Schocking dashcam footage shows Tesla 'Autopilot' crash which killed Chinese driver when futuristic electric car smashed into parked lorry. Daily Mail. https://www.dailymail.co.uk/news/article-3790176/amp/Shocking-dashcam-footage-shows-Tesla-Autopilot-crash-killed-Chinese-driver-futuristic-electric-car-smashed-parked-lorry.html [accessed 30/8/19].

Danaher, J. (2017). Robotic rape and robotic child sexual abuse: Should they be criminalised? *Criminal Law and Philosophy*, 11(1), 71–95.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Available from: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scrapssecret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Datta, A., Tschantz andM.C., Datta, A. (2015). Automated Experiments on Ad Privacy Settings – A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*. 1, 92–112, DOI: 10.1515/popets-2015-0007

Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J., and Hajkowicz, S. (2019). *Artificial Intelligence: Australia's Ethics Framework.* Available from: https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/supporting_documents/ArtificialIntelligenceethicsframeworkdiscussionpaper.pdf

De Angeli, A. (2009). Ethical implications of verbal disinhibition with conversational agents. *Psychology Journal*, 7(1), 49–57.

De Angeli, A., & Brahnam, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with Computers.* 20(3), 302–310

De.digital. (2018). *The Federal Government's Artificial Intelligence Strategy*. [online] Available from: https://www.de.digital/DIGITAL/Redaktion/EN/Standardartikel/artificial-intelligence-strategy.html. [Accessed 10 May 2019].

Delvaux, M. (2017). 'With recommendations to the Commission on Civil Law Rules on Robotics' *European Commission 2015/2103(INL)*.

Die Bundesregierung (2018) *Strategie Künstliche Intelligenz der Bundesregierung*.

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology,* 20: 1.

Digital Poland Foundation (2019). *Map of the Polish AI*. Digital Poland Foundation..

Duckworth, P., Graham, L., Osborne andM.AI (2019). Inferring Work Task Automatability from AI Expert Evidence. *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*. University of Oxford.

Dutton, T. (2018). An Overview of National AI Strategies. [online] *Medium*. Available at: https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd [Accessed 4 April 2019].

Ethics Commission (2017). Ethics's Commission's complete report on automated and connected driving. *Federal Ministry of Transport and Infrastructure.* Available from: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html?nn=187598

Etzioni, A. and Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology*, 18(2), 149–156

European Commission (2012) Special Eurobarometer 382: Public Attitudes towards Robots. Eurobarometer Surveys [online] Available at: https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/SPECIAL/surveyKy/1044/p/3

European Commission (2017) Special Eurobarometer 460: Attitudes towards the impact of digitisation and automation on daily life [online] Available at: https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/SPECIAL/surveyKy/2160

European Commission (2018a). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe*. Available from: https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

European Commission (2018b). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions - Coordinated Plan on Artificial Intelligence* (COM(2018) 795 final). Available from: https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence

European Commission (2018c). High-level expert group on artificial intelligence: Draft ethics guidelines for trustworthy AI. *Brussels.* [online] Available from: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_draft_ethics_guidelines_18_december.pdf [Accessed 15/03/2019].

European Commission (2018d). EU Member States sign up to cooperate on Artificial Intelligence. [online] Available at: https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence [Accessed 30 Oct. 2019].

European Commission High-Level Expert Group on Artificial Intelligence (2019) *Ethics Guidelines for Trustworthy AI.* Available from: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477

European Commission High-Level Expert Group on AI (2019b) Policy and Investment Recommendations for Trustworthy AI. Available from: https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

European Parliament, Council and Commission, (2012). Charter of Fundamental Rights of the European Union. *Official Journal of the European Union*

European Parliament, 2017. EP Resolution with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). Available at: http://www.europarl.europa.eu/

Europol. (2017). *Serious and organised crime threat assessment*. Available from: https://www.europol.europa.eu/socta/2017/.

Everett, J., Pizarro, D. and Crockett, M, (2017). Why are we reluctant to trust robots? *The Guardian.* Available from: https://www.theguardian.com/science/head-quarters/2017/apr/24/why-are-we-reluctant-to-trust-robots

Ezrachi, A., & Stucke, M. E. (2016). Two artificial neural networks meet in an online hub and change the future (of competition, market dynamics and society). *Oxford Legal Studies Research Paper, No. 24/2017; University of Tennessee Legal Studies Research Paper, No. 323.*

Farmer, J. D., & Skouras, S. (2013). An ecological perspective on the future of computer trading. *Quantitative Finance*. 13(3), 325–346

Felton, R. (2017). Limits of Tesla's Autopilot and driver error cited in fatal Model S crash. *Jalopnik.* Available from: https://jalopnik.com/limits-of-teslas-autopilot-and-driver-error-cited-in-fa-1803806982#_ga=2.245667396.1174511965.1519656602-427793550.1518120488

Felton, R. (2018). Two years on, a father is still fighting Tesla over autopilot and his son's fatal crash. *Jalopnik.* Available from: https://jalopnik.com/two-years-on-a-father-is-still-fighting-tesla-over-aut-1823189786

Ferrara, E. (2015). *Manipulation and abuse on social media*

Floridi, L. (2016). Tolerant paternalism: Pro-ethical design as a resolution of the dilemma of toleration. *Science and Engineering Ethics*. 22(6), 1669–1688.

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*,374(2083).

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford review.* 5. Oxford, Oxford University Press.

Ford, M. (2009) *The Lights in the Tunnel: Automation, Accelerating Technology, and the Economy of the Future.*

Foundation for Law & International Affairs (2017) China's New Generation of Artificial Intelligence Development Plan. *FLIA*. [online] Available FROM: https://flia.org/wp-content/uploads/2017/07/A-New-Generation-of-Artificial-Intelligence-Development-Plan-1.pdf

Frey, C, B. and Osborne, M, A. (2013). The Future of Employment: How Susceptible Are Jobs to Computerisation? *Oxford Martin Programme on the Impacts of Future Technology*.

Furman, J & Seamans, R. (2018). AI and the Economy. *NBER working paper no.24689*

Future of Life Institute (2019). National and International AI Strategies.*Future of Life Institute*. [online] Available from: https://futureoflife.org/national-international-ai-strategies/ [Accessed 28 Apr. 2019].

G7 Canadian Presidency (2018). *Charlevoix Common Vision for the Future of Artificial Intelligence*.

G20 (2019) G20 Ministerial Statement on Trade and Digital Economy: Annex. Available from: https://www.mofa.go.jp/files/000486596.pdf

Gagan, O. (2018) Here's how AI fits into the future of energy, World Economic Forum, 25 May 2018 [Online] Available at: https://www.weforum.org/agenda/2018/05/how-ai-can-help-meet-global-energy-demand [Accessed on 13 Aug. 2019].

Garfinkel, S. (2017). Hackers are the real obstacle for self-driving vehicles. *MIT Technology Review*. Available from: https://www.technologyreview.com/s/608618/hackers-are-the-real-obstacle-for-self-driving-vehicles/

Gibbs, S. (2017). Tesla Model S cleared by safety regulator after fatal Autopilot crash. *The Guardian*. Available from: https://www.theguardian.com/technology/2017/jan/20/tesla-model-s-cleared-auto-safety-regulator-after-fatal-autopilot-crash

Gillespie T. (2014). The relevance of algorithms. In Gillespie, T., Boczkowsi, P. J., Foot, K. A. (eds.) (2014). *Media technologies: essays on communication, materiality, and society*.Cambridge, MA: MIT Press. pp. 167-194.

Gogarty, B., & Hagger, M. (2008). The laws of man over vehicles unmanned: The legal response to robotic revolution on sea, land and air. *Journal of Law, Information and Science*, 19, 73–145.

Goldhill, O. (2016). Can we trust robots to make moral decisions? *Quartz*. Available from: https://qz.com/653575/can-we-trust-robots-to-make-moral-decisions/

UK Government Office for Science (2015) Artificial intelligence: opportunities and implications for the future of decision making. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf [Accessed 13 Aug. 2019].

GOV.UK. (2018a). *AI Sector Deal.* [online] Available from https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal [Accessed 10 May 2019].

GOV.UK. (2018b). *Centre for Data Ethics and Innovation (CDEI).* [online] Available from: https://www.gov.uk/government/groups/centre-for-data-ethics-and-innovation-cdei [Accessed 10 May 2019].

GOV.UK (2019). The UK's Industrial Strategy. *GOV.UK*. [online] Available from: https://www.gov.uk/government/topical-events/the-uks-industrial-strategy [Accessed 10 May 2019].

Government Offices of Sweden (2018). National approach to artificial intelligence. *Ministry of Enterprise and Innovation*.

Graetz, G. and Michaels, G. (2015). Robots at Work. *Centre for Economic Performance Discussion Paper No. 1335*.

Gray, M. L. and Suri, S. (2019). *Ghost Work,* Houghton Mifflin Harcourt.

Greene, J. D., Sommerville, R. B., Nystrom, L., Darley, J., and Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105–2108. doi:10.1126/science. 1062872.

Guiltinan, J. (2009). Creative destruction and destructive creations: Environmental ethics and planned obsolescence. *Journal of Business Ethics*. 89 (1). pp.1928.

Gurney, J. K., (2013). Sue My Car, Not Me: Products Liability and Accidents Involving Autonomous Vehicles. unpublished manuscript

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2016). The off-switch game. In: *IJCAI-ECAI-2018: International Joint Conference on Artificial Intelligence. IJCAI-ECAI-2018, 13-19 July 2018, Stockholm, Sweden.*

Hallaq, B.,, Somer, T., Osula, A., Ngo, K., & Mitchener-Nissen, T. (2017). Artificial intelligence within the military domain and cyber warfare. In: 16th European Conference on Cyber Warfare and Security (ECCWS 2017), 29-30 June 2017, Dublin, Ireland.Published in: Proceedings of 16th European Conference on Cyber Warfare and Security.

Hallevy, G. (2010) The Criminal Liability of Artificial Intelligence Entities (February 15, 2010). Available at SSRN: https://ssrn.com/abstract=1564096 or http://dx.doi.org/10.2139/ssrn.1564096

Hancock, J. T., Curry, L. E., Goorha, S., and Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes.* 45(1): 1–23.

Harambam, J., Helberger, N., and Van Hoboken, J. (2018). Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133).

Hardt, M. (2014). *How Big Data is Unfair*. *Medium.* [online] Available from [accessed 9 Apr. 2019]

Hart, R, D. (2018). Who's to blame when a machine botches your surgery? *Quartz.* Available from: https://qz.com/1367206/whos-to-blame-when-a-machine-botches-your-surgery/

Hawkins, A. J. (2019). California's self-driving car reports are imperfect, but they're better than nothing. *The Verge.* Available from: https://www.theverge.com/2019/2/13/18223356/california-dmv-self-driving-car-disengagement-report-2018

Hawksworth, J. and Fertig, Y. (2018) What will be the net impact of AI and related technologies on jobs in the UK? PwC UK Economic Outlook, July 2018.

Hern, A. (2016). 'Partnership on AI' formed by Google, Facebook, Amazon, IBM and Microsoft. *The Guardian*. Available from: https://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms

Hess, A,. (2016). On Twitter, a Battle Among Political Bots. *The New York Times*. Available from: https://www.nytimes.com/2016/12/14/arts/on-twitter-a-battle-among-political-bots.html

Human Rights Watch. (2018). 'Eradicating ideological viruses': China's campaign of repression against Xinjiang's Muslims. *Technical report,* Human Rights Watch.

IEEE (2019). *Homepage* [online] Available from: https://www.ieee.org [Accessed 11 Mar2019].

Iglinski, H., Babiak, M. (2017). Analysis of the Potential of Autonomous Vehicles in Reducing the Emissions of Greenhouse Gases in Road Transport. *Procedia Eng*.192, 353–358.

International Telecommunication Union (2018). *AI for Good Global Summit 2018* [online] Available from: https://www.itu.int/en/ITU-T/AI/2018/Pages/default.aspx [Accessed 14 May 2019].

International Telecommunication Union (2018). United Nations Activities on Artificial Intelligence [online]. Available from: http://www.itu.int/pub/S-GEN-UNACT-2018-1 [Accessed 12 November 2019]

Isaac, M. (2016). Self-driving truck's first mission: a 120-mile beer run. *New York Times.* Available from: https://www.nytimes.com/2016/10/26/technology/self-driving-trucks-first-mission-a-beer-run.html

Israel Innovation Authority (2019*). Israel Innovation Authority 2018-19 Report*. [online] Available from: https://innovationisrael.org.il/en/news/israel-innovation-authority-2018-19-report [Accessed 10 May 2019].Iyengar, S., Sood, G., and Lelkes, Y. (2012). Affect, not ideology: Social identity perspective on polarization. *Public Opinion Quarterly*. 76(3),405.

Jacobs, S. B. (2017) The Energy Prosumer, 43Ecology L. Q.519.

Japanese Strategic Council for AI Technology (2017). *Artificial Intelligence Technology Strategy*. Available from: https://www.nedo.go.jp/content/100865202.pdf

Johnson, A., and Axinn, S. (2013). The Morality of Autonomous Robots. *Journal of Military Ethics.* 12 (2), 129-141

Johnston, A. K. (2015). Robotic seals comfort dementia patients but raise ethical concerns. *KALW*. Available from: https://www.kalw.org/post/robotic-seals-comfort-dementia-patients-raise-ethical-concerns#stream/0

JSAI (2017). *Ethical Guidelines.* [online] Available from: http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf [Accessed 7 May19].

JSAI (2019). *Overview: Inaugural Address of President Naohiko Uramoto, Artificial Intelligence expanding its scope and impact in our society*. [online] Available from: https://www.ai-gakkai.or.jp/en/about/about-us/ [Accessed 11 May 2019].

Kayali, L. (2019). *Next European Commission takes aim at AI*. [online] POLITICO. Available at: https://www.politico.eu/article/ai-data-regulator-rules-next-european-commission-takes-aim/ [Accessed 27 Aug. 2019].

Kenyan Wall Street (2018). Kenya Govt unveils 11 Member Blockchain & AI Taskforce headed by Bitange Ndemo. *Kenyan Wallstreet*. [online. Available from: https://kenyanwallstreet.com/kenya-govt-unveils-11-member-blockchain-ai-taskforce-headed-by-bitange-ndemo/ [Accessed 6 May 2019].

Khakurel, J.,Penzenstadler, B., Porras, J., Knutas, A., and Zhang, W. (2018). The Rise of Artificial Intelligence under the Lens of Sustainability. *Technologies*. *6*(4), 100.

Khosravi, B. (2018). Autonomous cars won't work – until we have 5G. *Forbes.* Available from: https://www.forbes.com/sites/bijankhosravi/2018/03/25/autonomous-cars-wont-work-until-we-have-5g

King, T.C., Aggarwal, N., Taddeo, M. et al. (2019). Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Sci Eng Ethics.* pp.1-32

Kingston, J. K. C. (2018) Artificial Intelligence and Legal Liability. Available at: https://arxiv.org/ftp/arxiv/papers/1802/1802.07782.pdf [Accessed 17/08/19].

Kitwood, T. (1997). *Dementia Reconsidered: The Person Comes First*. Buckingham, Open University Press.

Knight, W. (2019). *The World Economic Forum wants to develop global rules for AI*. [online] MIT Technology Review. Available at: https://www.technologyreview.com/s/613589/the-world-economic-forum-wants-to-develop-global-rules-for-ai/ [Accessed 20 Aug. 2019].

Kroll, J.A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133).

Lalji, N. (2015). Can we learn about empathy from torturing robots?This MIT researcher isgiving it a try. *YES! Magazine.* Available from: http://www.yesmagazine.org/happiness/should-we-be-kind-to-robots-katedarling.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*. 10, (1096)

LaRosa, E., & Danks, D. (2018). Impacts on Trust of Healthcare AI. In: *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. AEIS: 2018, 1-3 February, 2018, New Orleans, USA.*

Larsson, S., Anneroth, M., Felländer, A., Felländer-Tsai, L., Heintz, F., Cedering Ångström, R. (2019). Sustainable AI report. *AI Sustainability Centre.* Available from: http://www.aisustainability.org/wp-content/uploads/2019/04/SUSTAINABLE-AI.pdf

Lashbrook, A. (2018). AI-driven dermatology could leave dark-skinned patients behind. *The Atlantic.* Available from: https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/

Leggett, T. (2018) Who is to blame for 'self-driving car' deaths? BBC Business News. 22 May 2018. Available at: https://www.bbc.co.uk/news/business-44159581

Le Miere, J. (2017). Russia is developing autonomous 'swarms of drones' it calls an inevitable part of future warfare. [online] *Newsweek*. Available at: https://www.newsweek.com/drones-swarm-autonomous-russia-robots-609399 [Accessed 26 Apr. 2019].

Leontief, Wassily,. (1983). National Perspective: The Definition of Problems and Opportunities.. *The Long-Term Impact of Technology on Employment and Unemployment*. Washington, DC: The National Academies Press. doi: 10.17226/19470.

Lerner, S. (2018). NHS might replace nurses with robot medics such as carebots: could this be the future of medicine? *Tech Times.* Available from: https://www.techtimes.com/articles/229952/20180611/nhs-might-replace-nurses-with-robot-medics-such-as-carebots-could-this-be-the-future-of-medicine.htm

Levin, S. (2018). Video released of Uber self-driving crash that killed woman in Arizona. *The Guardian*. Available from: https://www.theguardian.com/technology/2018/mar/22/video-released-of-uber-self-driving-crash-that-killed-woman-in-arizona

Li, H., Milani, S., Krishnamoorthy, V., Lewis, M., & Sycara, K. (2019). Perceptions of Domestic Robots' Normative Behavior Across Cultures. In: *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. AEIS: 2019, 27-28 January, 2019, Honolulu, Hawaii, USA.* Available here: http://www.aies-conference.com/2019/wp-content/papers/main/AIES-19_paper_232.pdf

Li, S., Williams, J. (2018). Despite what Zuckerberg's testimony may imply, AI Cannot Save Us. *Electronic Frontier Foundation*. Available from: https://www.eff.org/deeplinks/2018/04/despite-whatzuckerbergs-testimony-may-imply-ai-cannot-save-us

Lim, D., (2019). Killer Robots and Human Dignity. In: AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. *AEIS: 2019, 27-28 January, 2019, Honolulu, Hawaii, USA.*

Lin, P. (2014). What if your autonomous car keeps routing you past Krispy Kreme? *The Atlantic*. Available from: https://finance.yahoo.com/news/autonomous-car-keeps-routing-past-130800241.html;_ylt=A2KJ3CUL199SkjsAexPQtDMD?guccounter=1&guce_

Lin, P., Jenkins, R., & Abney, K. (2017). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press.

Lin, T. C. W. (2017). The new market manipulation. *Emory Law Journal*, 66, 1253.

Loh, W. & Loh, J. (2017). Autonomy and responsibility in hybrid systems. In P. Lin, et al. (Eds.), *Robot ethics 2.0*. New York, NY: Oxford University Press: 35–50.

Lokhorst, G.-J. and van den Hoven, J. (2014) Chapter 9: Responsibility for Military Robots. In *Robot Ethics: The Ethical and Social Implications of Robotics* edited by Lin, Abney and Bekey (10 Jan. 2014, MIT Press).

Malta AI (2019). *Malta AI: Towards a National AI Strategy* [online] Available at: https://malta.ai [Accessed 10 May 2019].

Manikonda, L., Deotale, A., & Kambhampati, S,. (2018). What's up with Privacy? User Preferences and Privacy Concerns in Intelligent Personal Assistants. In: *AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. AEIS: 2018, 1-3 February, 2018, New Orleans, USA.*

Marda, V,. (2018). Artificial intelligence policy in India: a framework for engaging the limits of data-driven decision-making. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 376 (2133).

Marshall, A. and Davies, A. (2018). Lots of lobbies and zero zombies: how self-driving cars will reshape cities. *Wired.* Available from: https://www.wired.com/story/self-driving-cars-cities/

Martinho-Truswell, E., Miller, H., Nti Asare, I., Petheram, A., Stirling, R., Gómez Mont, G. and Martinez, C. (2018). *Towards an AI Strategy in Mexico: Harnessing the AI Revolution*.

Mattheij, J. (2016) 'Another Way Of Looking At Lee Sedol vs AlphaGo'. Jacques Mattheij: Technology, Coding and Business. Blog. 17th March 2016.

Matthias, A. (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, Sept 2004, Vol. 6, Issue 3, pp.175-183.

Mazzucato, M. (2018) Mission-Oriented Research & Innovation in the European Union. European Commission: Luxebourg.

Mbadiwe, T. (2017). The potential pitfalls of machine learning algorithms in medicine. *Pulmonology Advisor.* Available from: https://www.pulmonologyadvisor.com/home/topics/practice-management/the-potential-pitfalls-of-machine-learning-algorithms-in-medicine/

McAllister, A. (2017). Stranger than science fiction: The rise of AI interrogation in the dawn of autonomous robots and the need for an additional protocol to the UN convention against torture. *Minnesota Law Review.* 101, 2527–2573.

McCarty, N. M., Poole, K. T., and Rosenthal, H. (2016). *Polarized America: The Dance Of Ideology And Unequal Riches.* Cambridge, MA: MIT Press, 2nd edition.

Meisner, E. M. (2009). *Learning controllers for human–robot interaction*. PhD thesis. Rensselaer Polytechnic Institute.

México Digital (2018). Estrategia de Inteligencia Artificial MX 2018. [online] *gob.mx*. Available from: https://www.gob.mx/mexicodigital/articulos/estrategia-de-inteligencia-artificial-mx-2018 [Accessed 6 May 2019].

Millar, J. (2016). *An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars*. 30(8), 787-809.

Min, W. (2018) Smart Policies for Harnessing AI, OECD-Forum, 17 Sept 2018 [online] Available from: https://www.oecd-forum.org/users/68225-wonki-min/posts/38898-harnessing-ai-for-smart-policies [Accessed 12 Aug. 2019].

Ministry of Economic Affairs and Employment of Finland (2017). Finland's Age of Artificial Intelligence. Available from: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf
Ministry of Economic Affairs and Employment of Finland (2018a). *Artificial intelligence programme*. [online] Available from: https://tem.fi/en/artificial-intelligence-programme [Accessed 26 Apr. 2019].

Ministry of Economic Affairs and Employment of Finland (2018b). *Work in the Age of Artificial Intelligence.* Available from: https://www.google.com/search?client=safari&rls=en&q=work+in+the+age+of+artificial+intelligence&ie=UTF-8&oe=UTF-8

Mizoguchi, R. (2004). The JSAI and AI activity in Japan. *IEEE Intelligent Systems* 19 (2).

Moon, M., (2017). Judge allows pacemaker data to be used in arson trial. *Engadget*. Available from: https://www.engadget.com/2017/07/13/pacemaker-arson-trial-evidence/

National Science & Technology Council (2019) The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update. Available from: https://www.whitehouse.gov/wp-content/uploads/2019/06/National-AI-Research-and-Development-Strategic-Plan-2019-Update-June-2019.pdf

NTSB (2018) Preliminary Report Released for Crash Involving Pedestrian, Uber Technologies, Inc., Test Vehicle. National Transport Safety Board News Release. May 24, 2018. Available at: https://www.ntsb.gov/news/press-releases/Pages/NR20180524.aspx

Nemitz, P,. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 376 (2133).

Newman, E. J., Sanson, M., Miller, E. K., Quigley-McBride, A., Foster, J. L., Bernstein, D. M., and Garry, M. (2014). People with easier to pronounce names promote truthiness of claims. *PLoS ONE*.9(2).

NITI Aayog (2018). *National Strategy for Artificial Intelligence #AIFORALL*.

Nevejans, N. et al. (2018). *Open letter to the European Commission on Artificial Intelligence and Robotics*.

New America. (2018). Translation: *Chinese government outlines AI ambitions through 2020*. [online] Available from: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-government-outlines-ai-ambitions-through-2020/ [Accessed 27 Apr. 2019].

NHS Digital. (2019). *Widening Digital Participation*. NHS Digital. Available from: https://digital.nhs.uk/about-nhs-digital/our-work/transforming-health-and-care-through-technology/empower-the-person-formerly-domain-a/widening-digital-participation

NHS' Topol Review. (2019). *Preparing the healthcare workforce to deliver the digital future.* Available from: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf

Nordic cooperation (2018*). AI in the Nordic-Baltic region*. [online] Available from: https://www.norden.org/en/declaration/ai-nordic-baltic-region [Accessed 26 Apr. 2019].

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C.and Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America.*,115, E5716–E5725.

O'Carroll, T. (2017). Mexico's misinformation wars. *Medium.* Available from: https://medium.com/amnesty-insights/mexico-s-misinformation-wars- cb748ecb32e9#.n8pi52hot

O'Connor, T. (2017). Russia is building a missile that can makes its own decisions. [online] *Newsweek*. Available from: https://www.newsweek.com/russia-military-challenge-us-china-missile-own-decisions-639926 [Accessed 26 Apr. 2019].

O'Donoghue, J. (2010). E-waste is a growing issue for states. *Deseret News.* Available from: http://www.deseretnews.com/article/700059360/E-waste-is-a-growing-issue-for-states.html?pg=1

O'Kane, S (2018). Tesla defends Autopilot after fatal Model S crash. *The Verge.* Available from: https://www.theverge.com/2018/3/28/17172178/tesla-model-x-crash-autopilot-fire-investigation

O'Neil, C,. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Crown Publishers.

O'Neill, S. (2018). As insurers offer discounts for fitness trackers, wearers should step with caution. *National Public Radio.* Available from: https://www.npr.org/sections/health-shots/2018/11/19/668266197/as-insurers-offer-discounts-for-fitness-trackers-wearers-should-step-with-cautio?t=1557493660570

OECD (2013) Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data [OECD/LEGAL/0188]

OECD (n.d.) OECD initiatives on AI [online] Available at: http://www.oecd.org/going-digital/ai/ [Accessed 13 Aug. 2019].

Ori.(2014a). If Death by Autonomous Car is Unavoidable, Who Should Die? Reader Poll Results. *Robohub.org*. Available from: http://robohub.org/if-a-death-by-an-autonomous-car-is-unavoidable-who-should-die-results-from-our-reader-poll/.

Ori. (2014b). My (autonomous) car, my safety: Results from our reader poll. *Robohub.org*.. Available from: http://robohub.org/my-autonomous-car-my-safety-results-from-our-reader-poll

Orseau, L. & Armstrong, S. (2016). Safely interruptible agents. In: *Uncertainty in artificial intelligence: 32nd Conference (UAI). UAI: 2016, June 25-29, 2016, New York City, NY, USA.* AUAI Press 2016

Ovanessoff, A. and Plastino, E. (2017). How Artifical Intelligence Can Drive South America's Growth. *Accenture.*

Oxford Insights (2019) Government Artificial Intelligence Readiness Index. Available from: https://ai4d.ai/wp-content/uploads/2019/05/ai-gov-readiness-report_v08.pdf

Pagallo, U. (2017). Apples, oranges, robots: four misunderstandings in today's debate on the legal status of AI systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 376 (2133).

Pariser E. (2011). *The filter bubble: what the Internet is hiding from you*. London, UK, Penguin.

Park, M. (2017). Self-driving bus involved in accident on its first day. *CNN Business.* Available from: https://money.cnn.com/2017/11/09/technology/self-driving-bus-accident-las-vegas/index.html

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA, Harvard University Press.

Personal Data Protection Commission Singapore (2019). *A Proposed Model Artificial Intelligence Governance Framework*. Available from: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/A-Proposed-Model-AI-Governance-Framework-January-2019.pdf

Pfleger, P. (2018). Transportation workers form coalition to stop driverless buses in Ohio. *WOSU Radio.* Available from: https://radio.wosu.org/post/transportation-workers-form-coalition-stop-driverless-buses-ohio#stream/0

Pham, T., Gorodnichenko, Y. and Talavera, O. (2018). Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection. NBER Working Papers w24631. The National Bureau of Economic Research; Cambridge, MA.

Piesing, M. (2014). Medical robotics: Would you trust a robot with a scalpel? *The Guardian.* Available at: https://www.theguardian.com/technology/2014/oct/10/medical-robots-surgery-trust-future

Plantera, F. (2017). Artificial Intelligence is the next step for e-governance in Estonia, State adviser reveals.[online] *e-Estonia.* Available from: https://e-estonia.com/artificial-intelligence-is-the-next-step-for-e-governance-state-adviser-reveals/. [Accessed 28 Apr. 2019].

Polonski, V. (2017). #MacronLeaks changed political campaigning. Why Macron succeeded and Clinton failed. *World Economic Forum*. Available from: https://www.weforum.org/agenda/2017/05/macronleaks-have-changed-political-campaigning-why-macron-succeeded-and-clinton-failed

Press Association (2019). Robots and AI to give doctors more time with patients, says report. *The Guardian*. Available from: https://www.theguardian.com/society/2019/feb/11/robots-and-ai-to-give-doctors-more-time-with-patients-says-report

ProPublica (2016). Machine Bias. *ProPublica. Available from:* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology.* 20: 5. https://doi.org/10.1007/s10676-017-9430-8

Ramchurn, S. D. et al. (2013) AgentSwitch: Towards Smart Energy Tariff Selection. Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

Reuters (2019). *G7 urges tight regulations for digital currencies, agrees to tax digital giants locally*. [online] VentureBeat. Available at: https://venturebeat.com/2019/07/19/g7-urges-tight-regulations-for-digital-currencies-agrees-to-tax-digital-giants-locally/ [Accessed 27 Aug. 2019].

Riedl, M.O., and Harrison, B. (2017. Enter the matrix: A virtual world approach to safely interruptable autonomous systems. *arXiv*. preprint arXiv:1703.10284

Roberts, S. (2016) 'Digital Refuse: Canadian Garbage, Commercial Content Moderation and the Global Circulation of Social Media's Waste'. Media Studies Publications.

Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schultz, J., Hale, T. M., and Stern M.J. (2015) Digital Inequalities and Why They Matter. *Information, Communication & Society*. 18 (5), 569-592. http://dx.doi.org/10.1080/1369118X.2015.1012532

SAE International. (2018). SAE International releases updated visual chart for its 'levels of driving automation' standard for self-driving vehicles. *SAE International.* Available from: https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-'levels-of-driving-automation'-standard-for-self-driving-vehicles

Sage, A. (2018). Waymo unveils self-driving taxi service in Arizona for paying customers. *Reuters.* Available from: https://www.reuters.com/article/us-waymo-selfdriving-focus/waymo-unveils-self-driving-taxi-service-in-arizona-for-paying-customers-idUSKBN1O41M2

Saidot (2019). *About us* [online] Available from: https://www.saidot.ai/about-us [Accessed 3 May 2019].Salvage, M. (2019). Call for poor and disabled to be given fitness trackers. *The Guardian.* Available from: https://www.theguardian.com/inequality/2019/may/04/fitbits-nhs-reduce-inequality-health-disability-poverty

Sample, I. (2017). Computer says no: why making AIs fair, accountable and transparent is crucial. *The Guardian*. Available from: https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial

Sample, I. (2017). Give robots an 'ethical black box' to track and explain decisions, say scientists. *The Guardian*. Available from: https://www.theguardian.com/science/2017/jul/19/give-robots-an-ethical-black-box-to-track-and-explain-decisions-say-scientists

Santos-Lang, C. (2002). Ethics for Artificial Intelligences. In Wisconsin State-Wide technology Symposium 'Promise or Peril?'. *Reflecting on computer technology: Educational, psychological, and ethical implications.* Wisconsin, USA.

Sarmah, H. (2019). Looking East: How South Korea Is Making A Strategic Move In AI. [online] *Analytics India Magazine*. Available from: https://www.analyticsindiamag.com/looking-east-how-south-korea-is-making-a-strategic-move-for-ai-leadership/ [Accessed 28 Apr. 2019].

Sathe G. (2018). Cops in India are using artificial intelligence that can identify you in a crowd. *Huffington Post.* Available at: https://www.huffingtonpost.in/2018/08/15/facial-recognitionai-is-shaking-up-criminals-in-punjab-but-should-you-worry-too_a_23502796/.

Sauer, G. (2017). A Murder Case test's Alexa's Devotion to your Privacy. *Wired*. Available from https://www.wired.com/2017/02/murder-case-tests-alexas-devotion-privacy/

Scherer, M. U. (2016) Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies, *29 Harv. J. L. & Tech. 353 (2015-2016)*

Scheutz, M. (2012). The inherent dangers of unidirectional emotional bonds between humans and social robots. In: Lin, P., Abney, K. and Bekey, G. (eds.). Robot Ethics: *The Ethical and Social Implications of Robotics*, MIT Press, pp.205-221.

Schmitt, M.N., (2013). *Tallinn manual on the international law applicable to cyber warfare.* Cambridge University Press.

Schönberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology* 27 (2), 171–203. https://doi.org/10.1093/ijlit/eaz004

Selbst, A. D. and Barocas. S. (2018). The intuitive appeal of explainable machines. *87 Fordham Law Review 1085 Preprint*, available from: https://ssrn.com/abstract=3126971

Selbst, A. D. and Powles, J. (2017) Meaningful information and the right to explanation. *Int. Data Privacy Law* 7, 233–242. (doi:10.1093/idpl/ipx022)

Selinger, E. and Hartzog, W. (2017). Obscurity and privacy. In: Pitt, J. and Shew, A. (eds.). *Spaces for the Future: A Companion to Philosophy of Technology*, New York: Routledge.

Servoz, M. (2019) The Future of Work? Work of the Future! On How Artificial Intelligence, Robotics and Automation Are Transforming Jobs and the Economy in Europe, 10 May 2019. Available at: https://ec.europa.eu/epsc/publications/other-publications/future-work-work-future_en [Accessed 13 Aug. 2019].

Seth, S. (2017). Machine Learning and Artificial Intelligence Interactions with the Right to Privacy. *Economic and Political Weekly*, 52(51), 66–70

Sharkey, A., and Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology.* 14 (1): 27-40.

Sharkey, N., Goodman, M., & Ross, N. (2010). The coming robot crime wave. *IEEE Computer Magazine.* 43(8), 6–8.

Shepherdson, D. and Somerville, H. (2019) Uber not criminally liable in fatal 2018 Arizona self-driving crash – prosecutors. Reuters News. March 5, 2019. Available from: https://uk.reuters.com/article/uk-uber-crash-autonomous/uber-not-criminally-liable-in-fatal-2018-arizona-self-driving-crash-prosecutors-idUKKCN1QM2P4

Shewan, D. (2017). Robots will destroy our jobs – and we're not ready for it. *The Guardian.* Available from: https://www.theguardian.com/technology/2017/jan/11/robots-jobs-employees-artificial-intelligence.

Smart Dubai (2019a). *AI Ethics*. [online] Available from: https://www.smartdubai.ae/initiatives/ai-ethics [Accessed 10 May 2019].

Smartdubai.ae. (2019b). *AIEthics Self Assessment*. [online] Available from: https://www.smartdubai.ae/self-assessment [Accessed 12 May 2019].

Smith, A., & Anderson, J. (2014). *AI, Robotics, and the Future of Jobs*. Pew Research Center

Smith, B. (2018). Facial recognition technology: The need for public regulation and corporate responsibility. *Microsoft on the Issues*. Available from: https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/

Snaith, E. (2019). Robot rolls into hospital ward and tells 97-year-old man he is dying. *The Independent*. Available from: https://www.independent.co.uk/news/world/americas/robot-grandfather-dying-san-francisco-hospital-ernesta-quintana-california-a8815721.html

Solon, O. (2018). Who's driving? Autonomous cars may be entering the most dangerous phase. *The Guardian*. Available from: https://www.theguardian.com/technology/2018/jan/24/self-driving-cars-dangerous-period-false-security

Sparrow, R,. (2002). The march of the robot dogs. *Ethics and Information Technology*. 4 (4), 305–318.

Sparrow, R., and Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*. 16, 141-161.

Spatt, C. (2014). Security market manipulation. *Annual Review of Financial Economics*, 6(1), 405–418.

Stahl, B.C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*. 86, 152-161.

Stilgoe,J. and Winfield, A. (2018). Self-driving car companies should not be allowed to investigate their own crashes. *The Guardian*. Available from: https://www.theguardian.com/science/political-science/2018/apr/13/self-driving-car-companies-should-not-be-allowed-to-investigate-their-own-crashes

Strubell, E., Ganesh, A. and McCallum, A. (2019) Energy and Policy Considerations for Deep Learning in NLP, arXiv:1906.02243

Swedish AI Council. (2019). *Swedish AI Council*. [online] Available from: https://swedishaicouncil.com [Accessed 10 May 2019].

Taddeo, M. (2017). Trusting Digital Technologies Correctly. *Minds & Machines*. 27 (4), 565.

Taddeo, M. and Floridi, L. (2018) How AI can be a force for good. *Science* vol. 361, issue 6404, pp.751-752. DOI: 10.1126/science.aat5991

Task Force on Artificial Intelligence of the Agency for Digital Italy (2018*). White Paper on Artificial Intelligence at the service of citizens.*

Tesla. (nd). Support: autopilot. *Tesla*. Available from: https://www.tesla.com/support/autopilot

The Danish Government (2018). *Strategy for Denmark's Digital Growth*. Ministry of Industry, Business and Financial Affairs. Available from: https://eng.em.dk/media/10566/digital-growth-strategy-report_uk_web-2.pdf

The Danish Government (2019). *National Strategy for Artificial Intelligence*. Ministry of Finance and Ministry of Industry, Business and Financial Affairs. Available from: https://eng.em.dk/media/13081/305755-gb-version_4k.pdf

The Foundation for Responsible Robotics (2019). About us: *Our mission* [online] Available from: http://responsiblerobotics.org/about-us/mission/ [Accessed 11 Mar2019].

The Future of Life Institute (n.d.) AI Policy Challenges and Recommendations. Available at: https://futureoflife.org/ai-policy-challenges-and-recommendations/#top [Accessed 12/08/19].

The Future of Life Institute (2019). *Background: Benefits and Risks of Artificial Intelligence*.[online]. Available from: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence [Accessed 19 Mar.2019].

The Future Society (2019). *About us* [online] Available from: https://thefuturesociety.org/about-us [Accessed 11/03/2019].

The Institute of Electrical and Electronics Engineers (IEEE) (2017). *Ethically Aligned Design: First Edition. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. *(EADv2)*.

The Institute of Electrical and Electronics Engineers (IEEE) (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (EAD1e)*

The Institute for Ethical AI & Machine Learning (2019). *Homepage* [online] Available from: https://ethical.institute/index.html [Accessed 11 Mar.2019].

The Partnership on AI (2019). *About us* [online] Available from: https://www.partnershiponai.org/about/ [Accessed 11 Mar.2019].

The White House (2016) *Artificial Intelligence, Automation, and the Economy* [online] Available from: https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF [Accessed 12 Aug. 2019].

The White House (2019a). *Accelerating America's Leadership in Artificial Intelligence.* [online] Available from: https://www.whitehouse.gov/articles/accelerating-americas-leadership-in-artificial-intelligence/ [Accessed 28 Apr. 2019].

The White House (2019b). *Artificial Intelligence for the American People* [online] Available from: https://www.whitehouse.gov/ai/. [Accessed 28 Apr. 2019].

Thiagarajan, K. (2019). The AI program that can tell whether you may go blind. *The Guardian.* Available from: https://www.theguardian.com/world/2019/feb/08/the-ai-program-that-can-tell-whether-you-are-going-blind-algorithm-eye-disease-india-diabetes

Thielman, S. (2017). The customer is always wrong: Tesla lets out self-driving car data – when it suits. *The Guardian.* Available from: https://www.theguardian.com/technology/2017/apr/03/the-customer-is-always-wrong-tesla-lets-out-self-driving-car-data-when-it-suits

Thomson, J. (1976). Killing, letting die, and the trolley problem. *The Monist*. 59, 204–217.

Thurman N. (2011). Making 'The Daily Me': technology, economics and habit in the mainstream assimilation of personalized news. *Journalism.* 12, 395–415.

Tindera, M. (2018). Government data says millions of health records are breached every year. *Forbes*. https://www.forbes.com/sites/michelatindera/2018/09/25/government-data-says-millions-of-health-records-are-breached-every-year/#209fca3716e6

Torres Santeli, J. and Gerdon, S. (2019). *5 challenges for government adoption of AI*. [online] World Economic Forum. Available at: https://www.weforum.org/agenda/2019/08/artificial-intelligence-government-public-sector/ [Accessed 27 Aug. 2019].

TUM (2019). *New Research Institute for Ethics in Artificial Intelligence [Press Release]*. Available from: https://www.wi.tum.de/new-research-institute-for-ethics-in-artificial-intelligence/ [Accessed 11 Mar.2019].

Turkle, S., Taggart, W., Kidd, C.D. and Dasté, O.,(2006). Relational Artifacts with Children and Elders: The Complexities of Cyber companionship. *Connection Science*, 18 (4) pp 347-362.

UAE Government (2018). *UAE Artificial Intelligence Strategy 2031*. [online] Available from: http://www.uaeai.ae/en/ [Accessed 28 Apr. 2019].

UCL (2019). *IOE professor co-founds the UK's first Institute for Ethical Artificial Intelligence in Education [Press Release]*. Available from: https://www.ucl.ac.uk/ioe/news/2018/oct/ioe-professor-co-founds-uks-first-institute-ethical-artificial-intelligence-education [Accessed 11 Mar.2019].

UNICRI (2019). *UNICRI Centre for Artificial Intelligence and Robotics* [online]. Available from: http://www.unicri.it/in_focus/on/UNICRI_Centre_Artificial_Robotics [Accessed 14 May 2019].

UK Government Department for Digital, Culture, Media & Sport (2019). *Centre for Data Ethics and Innovation: 2-year strategy.* Available from: https://www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-cdei-2-year-strategy

UNI Global Union (n.d.) *Top 10 principles for Ethical Artificial Intelligence* [online]. Available from: http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

United Kingdom Commission for Employment and Skills, (2014). *The Future of Work: Jobs and Skills in 2030.* Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/303334/er84-the-future-of-work-evidence-report.pdf

Université de Montréal (2017). *Montreal Declaration for a Responsible Development of AI'* [online] Available from: https://www.montrealdeclaration-responsibleai.com/the-declaration [Accessed 11 Mar.2019].

US Department of Defence (2018). *Summary of the 2018 Department of Defence Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity*. Available from: https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

U.S. Department of Education, (2014). *Science, Technology, Engineering and Math.*

Vanian, J. (2019). *World Economic Forum Wants to Help Companies Avoid the Pitfalls of Artificial Intelligence* [online] Fortune. Available at: https://fortune.com/2019/08/06/world-economic-forum-artificial-intelligence/ [Accessed 27 Aug. 2019].

Veale, M., Binns., R & Edwards, L. (2018). Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 376 (2133).

Veruggio, G. and Operto, F. (2006). *The Roboethics Roadmap*. Available from: http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf [Accessed 11 Mar.2019].

Villani, C. (2018*). For a Meaningful Artificial Intelligence: Towards a French and European Strategy*. Available from: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf

Vincent, J. (2017). Google's AI thinks this turtle looks like a gun, which is a problem. *The Verge*. Available from: https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed

Vincent J. (2018). Drones taught to spot violent behavior in crowds using AI. *The Verge*. Available from: https://www.theverge.com/2018/6/6/17433482/ai-automated-surveillance-drones-spotviolent-behavior-crowds.

Viscelli, S. (2018). *Driverless? Autonomous trucks and the future of the American trucker.* Center for Labor Research and Education, University of California, Berkeley, and Working Partnerships USA. Available from: http://driverlessreport.org/files/driverless.pdf

von der Leyen, U. (2019) Political guidelines for the next European Commission: 2019 – 2024. https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf

Wachter S., Mittelstadt B. & Floridi L. (2017). Why a right to explanation of automated decision making does not exist in the general data protection regulation. *Int. Data Privacy Law* 7, 76–99. (doi:10.1093/idpl/ipx005).

Wachter, S., Mittelstadt, B. & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*. 31 (2).

Wagner, A.R. (2018). An Autonomous Architecture that Protects the Right to Privacy. In: AAAI / ACM Conference on Artificial Intelligence, Ethics and Society. *AIES: 2018, 1-3 February, 2018, New Orleans, USA.*

Wallach, W. and Allen, C.,(2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, New York.

Weinburg, C. (2019). Self-driving shuttles advance in cities, raising jobs concerns. *The Information.* Available from: https://www.theinformation.com/articles/self-driving-shuttles-advance-in-cities-raising-jobs-concerns

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. Oxford, W. H. Freeman & Co.

Wellman, M. P. and Rajan, U. (2017). Ethical Issues for Autonomous Trading Agents. *Minds & Machines* 27 (4),609–624.

West, D. M. (2018). *The Future of Work: Robots, AI, and Automation*. Brookings Institution Press Washington DC.

Williams, R. (2017). *Lords select committee, artificial intelligence committee, written evidence (AIC0206).* Available from:

http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/70496.html#_ftn13

Winfield, A.F.T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133).

Winfield, A. F. (2019a). Ethical standards in Robotics and AI. Nature Electronics, 2(2), 46-48.

Winfield, A. (2019b) Energy and Exploitation: AIs dirty secrets, 28 June 2019 [online] Available at: http://alanwinfield.blogspot.com/2019/06/energy-and-exploitation-ais-dirty.html [Accessed 13 Aug. 2019].

Wolfe, F. and Mavon, K. (2017) How artificial intelligence will revolutionise the energy industry [online] Available at: http://sitn.hms.harvard.edu/flash/2017/artificial-intelligence-will-revolutionize-energy-industry/ [Accessed on 13 Aug. 2019].

Worland, J. (2016). Self-driving cars could help save the environment – or ruin it. It depends on us. *Time*. Available from: http://time.com/4476614/self-driving-cars-environment/

World Business Council for Sustainable Development (WBCSD). (2000). *Eco-Efficiency: Creating more Value with less Impact*. WBCSD: Geneva, Switzerland.

World Economic Forum (2018*). The world's biggest economies in 2018*. [online] Available from: https://www.weforum.org/agenda/2018/04/the-worlds-biggest-economies-in-2018/ [Accessed 26 Apr. 2019].

World Economic Forum. (2019a). *World Economic Forum Inaugurates Global Councils to Restore Trust in Technology*. [online] Available at: https://www.weforum.org/press/2019/05/world-economic-forum-inaugurates-global-councils-to-restore-trust-in-technology/ [Accessed 17 Aug. 2019].

World Economic Forum (2019b) White Paper: A Framework for Developing a National Artificial Intelligence Strategy. Available from: http://www3.weforum.org/docs/WEF_National_AI_Strategy.pdf

Yadron, D., Tynan, D. (2016). *Tesla driver dies in first fatal crash while using autopilot mode*. Available from https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences.* 112(4), 1036–1040.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv*. preprint arXiv:1707.09457

Zou, J. & Schiebinger, L. (2018). 'AI can be sexist and racist — it's time to make it fair', *Nature* Available from: https://www.nature.com/articles/d41586-018-05707-8

QA-01-19-779-EN-N

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

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

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