Effective, Timely and Global:
the urgent need for good Global Governance of AI

A report by the
Transnational Working Group on AI and Disruptive Technologies
of the World Federalist Movement and Institute for Global Policy

November 2020

Corporate Author
The Transnational Working Group on AI and Disruptive Technologies
of the World Federalist Movement and Institute for Global Policy

Corresponding author: rwhitfield@oneworldtrust.org

Members of the Transnational Working Group:

Keith Best
Didier Coeurnelle
Dr Walter Dorn
Marcelo Lemos
Dr Lars Lünenburger
Fritz Pointer
Marjolijn Snippe
Robert Whitfield (Chair)

We acknowledge the support received from John Daniele, Paju-Anna Hentunen, Imogen McCullough and Hester Mifsud and recent speakers who have inspired us including Jamie Metzl and Nell Watson.

Contents

Executive Summary

Artificial Intelligence (AI) has immense value to offer to humanity, such as improved efficiency, new capabilities, and solutions for complex problems. A fast-growing AI industry is developing, applying AI to every sector of our economy. AI is often able to reduce costs, render a service more effective or produce a design more quickly. As its capabilities continue to grow, it may prove as transformative as the proliferation of electrification and cheap motive power was in the 20th century, ushering in an era of abundance, longer and healthier lives and greater realisation of human rights.

However, AI also brings new problems and threats. This will necessitate the creation of governing institutions, as the impact of AI will be experienced in every country in the world. Governance needs to be effective, timely and global. As in many fields of human endeavour, issues that are not bounded by geography or jurisdiction require global responses.

This paper explores some of the arguments for the global governance of AI; it sets out conventions and institutions for global governance; and it highlights some of the complementary concerns that need to be addressed in parallel with the regulation. The governance of AI needs to address the immediate issues (good and bad), be able to ride the wave of technological and economic progress, and take into account the major long-term risks, a process which needs to start now.

There are several domains where the application of AI could be detrimental to the interests of individuals, societies and humanity at large. A categorisation of domains is proposed, and a greater application of AI to the domain of "Good AI" is encouraged as long as it is safe and ethical.

Governance and regulation
AI strategy, governance and regulation are all being addressed at the national level, with the leading AI nations competing for global leadership. Within this competitive context, it is difficult to believe that the optimum balance is being found between the facilitation of rapid development of AI and the mitigation and avoidance of its inherent and major risks. A more appropriate international approach is being sought by several intergovernmental organisations, but it is not clear that any is best suited to the governance of a technology of such importance.

AI technology is developing at a tremendous speed. Regulations are not generally known for their adaptability and governance has been evolving at a slow pace. There is a real challenge, therefore, to develop a governance and regulation of AI that provides the rigour where it is needed, but also flexibility where it is needed. There could possibly be a skeletal standard, something akin to WTO Rules, or the global rules could be more prescriptive: that remains to be determined. But one thing that these different approaches have in common is that they are all best based upon sound and agreed principles.

There are over 160 sets of ethical principles regarding AI[1], each new set adding to the apparent complexity. Five studies have shown however that these principles can be corralled into six clusters which can be described as beneficence, justice and fairness, safety and security, human autonomy, accountability, and transparency. On the basis of the cluster outlines and the associated themes, all set out in the report, it emerges that the negotiation of a global set of universalizable "Values and Principles" should not prove as great an obstacle as might at first appear.

One sector that is in immediate need of regulation is that of social media, especially AI-generated synthetic media that is used to distort and damage progressive human activity. Disinformation, typically travelling at several times the speed of the truth, is allowed to work its way through the networks, making money for the platforms but with collateral damage to the "users". Allied to social media markets are a variety of tools using AI to create disinformation. This is an area that desperately needs strong regulation, and there are several regulatory tools that can be brought to bear.

Impact on work, well-being and human existence
Whilst regulation can facilitate the orderly growth of a market, there may be other issues that need to be addressed in parallel with that growth. These include the impact of AI on work and society, the issue of deskilling of human beings, and risks to systemic fragility, as described in later sections.

The medium-term impact of AI on the job market is not certain, but most economic forecasters tend to see the AI revolution as different from earlier revolutions (e.g., agrarian, industrial, and information). It is not human muscle or even knowledge that is being replaced but intelligence, and that is expected to have a major long-term impact on the job market, leading to a radical reduction in hours worked in many areas of the economy (whether through higher unemployment or through forced part-time work). Plans need to be in place to successfully manage these changes and enable people to flourish, whether employed, working part-time or unemployed. Financial solutions will need to have been devised and discussed with the public before they are required. This will have to be globally coordinated.

Controlling the Movement towards Superintelligence
As AI increases its penetration of the job market, skills will start to be lost. "Moral deskilling" could develop if AIs are allowed to take more and more moral decisions. In the longer term, as almost everything is done through AI, there is a real risk of a tragedy of the commons, of individuals not seeing the point of studying and developing skills if the AI can simply do the job – and choosing not to acquire the skills or expertise. If everyone were to take that view, ultimately the machines would be in charge and humanity enfeebled. This is an unacceptable scenario – but it will require a major cultural change to prevent it from happening.

As AIs continue to develop towards the point of "superintelligence," it is imperative to design AI to be safe. It is not certain that it will be possible to retain control of the goals of a superintelligence. The existential risk for humanity has been assessed by Ord[2] at 10% in the next 100 years, an unacceptable scenario. Some excellent minds have been applied to find a solution to this problem. A number of different ideas are being pursued (Plan A), though it is far from clear whether they

will be successful in time. The danger is that humanity blunders its way to creating a superintelligence and loses control of it, before guaranteed methods of ensuring that the AI's goals are fully aligned with those of humanity. To avoid this, a pause capability will be needed, to be able to stop all further AI development for as long as is necessary (Plan B). Schemes such as Differential Technological Development[3] are proposed, helping to ensure that the above "safety" measures are given sufficient priority.

Conventions / institutions
Appropriate AI Governance is required in all spheres, including the civilian and military spheres. Within the military sphere, nuclear weapons systems depend to an increasing and dangerous degree on AI, where signals, for instance from radar, could prove faulty. It is important to make sure that the AI controllers in all nuclear-weapon states are up to the task: otherwise a global catastrophe could result.

Another major concern today is the development of Lethal Autonomous Robots (LARs), which are seen as a new level of warfare. A Protocol within the Convention for Certain Conventional Weapons (CCW) is urgently needed. It is proposed that nations unilaterally declare a domestic moratorium on LARs, negotiate a global moratorium and develop a Protocol / Convention totally banning LARs. A UN monitoring and inspection regime would also be required, including for confidence-building.

Several international bodies are currently engaging in the virgin territory of AI governance. Though it is possible that one could emerge as the successful institution, it is recommended to develop a UN Framework Convention on AI (UNFCAI) from scratch in the next few years, followed by a Protocol on AI. The negotiations currently underway within these different international bodies should proceed on that understanding.

To prepare the way, using perhaps the UN Secretary-General's new Advisory Group on AI as a springboard, it is proposed that a multi-stakeholder World Conference on AI should be organised in 2023, leading to the appointment of an International Negotiating Committee. In addition, new bodies are proposed in the form of a multistakeholder Intergovernmental Panel on AI (to provide scientific, technical and policy advice to the UNFCAI), an AI Global Authority to provide an inspection regime (including the military aspects) and a supervisory body that could introduce a democratic aspect.

Humanity has to be proactive on AI issues, or AI could pose a threat to humanity. By taking bold and progressive actions now, the future of humanity and the planet will be better protected.

RECOMMENDATIONS

Sec	
2.4	**Application areas** **a. A system of classification should be introduced and applied that distinguishes clearly between Good AI and Bad AI.** b. Promote, support, and ensure that AI systems are developed in the domain of Good AI.
4.7	**Values and principles** **A single global set of universalizable values and principles should be negotiated and adopted by all**.
5.5	**Social Media** a. **Regulation of the sector** including i. The certification of systems and companies that make them. ii. The certification of sufficient knowledge and awareness of professionals practicing within various domain capacities. iii. Ensuring that the global set of principles agreed includes the principle that the AI products be designed to operate humanely iv. The use of generative technologies such as Generative Adversarial Networks (GANs) in creating images, sounds and videos should be banned without a clear and indelible labelling of the use of such technologies upon media generated through such methods. Such labelling should be easily recognized by both humans and machines. b. Strengthening privacy laws and the adoption of privacy protecting technologies. c. Access to or the formation of an International Agency able to carry out monitoring and inspection – see Section 10.7 d. A tax on data collection and/or compensation to the data owners. If necessary, the use of attention models themselves should be banned.
6.4	**Recommendations re Work and Society** a. Agree on the need for the international community to reach a common understanding of the issues and the factors that can contribute to a positive way forward both economically and socially. b. Commit to ensuring that the planning required to address the economic and social consequences of the deployment of AI takes place hand in hand with the development of AI. **c. Ensure that there is a constructive global response to this global problem, that can be announced and implemented by 2025** d. Develop a phased approach, including ensuring that in the longer term there are contingency plans for a further transition as wealth increases, leading to a world of abundance e. Agree which international body should take on the role of leadership and coordination of the above.
7.4	**Recommendations re deskilling and long-term enfeeblement** a. Due thought should be given to the impact of increased deployment of AI on human skillsets (including moral skillsets) and autonomy, both in the short, medium and long term.

	b. Care needs to be taken to avoid AIs replacing skills that are fundamental to human identity, functioning and flourishing and the exercise of which may be considered essential for human dignity c. A future pause in development of AI, if feasible, might be desirable, as we find a suitable way to achieve these proposals (cf Rec. 8.8 d)
8.8	**Recommendations regarding Safeguarding the Future of Humanity** a. **Include within the regulatory framework measures such as Differential Technological Development designed to encourage the development and deployment of AI systems that will address the AI goal alignment problem and lead to Safe AI.** b. Promote the broad and speedy development of Beneficial AI, Friendly AI and other similar projects aimed at addressing the AI Alignment problem. c. Ensure that there will always be a means to stop an AI process from continuing, with the ability to close down a whole AI system if it threatens human welfare. The means must not be able to be circumvented by a highly intelligent AI. d. Develop a means of pausing the development of AI, with the support of AI if necessary, to be used in extremis.
9.7	**Governance action re Lethal Autonomous Robots and AWMD** a. A *moratorium* should be introduced on research, development, production, stock-piling and deployment of Lethal Autonomous Robots, including Autonomous Weapons of Mass Destruction. **b. A *treaty* should be negotiated, either as a Protocol of the Convention on Conventional Weapons or as a standalone Convention, to ban the research, development, production, stock-piling and deployment of Lethal Autonomous Robots.** c. A UN *inspection regime* should be created for immediate confidence-building and expertise development, and to verify an eventual treaty (cf Section 10.7)
10.8	**Main Global Governance on AI recommendations** **A multistakeholder World Conference on the Governance of AI should be held in 2023 as a prelude to the negotiation of** **a. A UN Framework Convention on AI** **b. A subsequent Protocol on AI** To support the negotiation and implementation of these agreements, providing a global legal framework for the regulation of AI, new bodies are proposed: a. A **Global Panel on AI**, possibly building upon the Global Partnership on AI, providing a technical support on AI analogous to that provided by the IPCC to the UNFCCC. b. An **AI Global Authority**, empowered to provide monitoring and inspection to support the work of the UN Framework Convention on AI. (cf Section 9.7) c. A **supervisory body** might also be associated with a democratic input as in other treaty-based institutions.

Key

	= AI and Regulation		**= Complement-ary action**		**= Conventions and Institutions**

1 INTRODUCTION

Artificial intelligence (AI) has many definitions such as the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. More technically, AI textbooks define the field as the study any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals.[4] Some definitions include an ability of the machine or robot to modify its own code to learn from experience. In this paper, the term AI is used fairly broadly, including Artificial General Intelligence, comparable to that of the human brain.

The development of Artificial Intelligence (AI), after some doldrums during the past few decades, is occurring at great speed, bringing huge benefits to many sectors of the economy. Not all deployment of AI is beneficial, however. These harmful aspects of AI need to be addressed. Furthermore, this is only the tip of the iceberg in terms of AI's potential. The capability of AI and the range of applications are both expected to grow dramatically over the coming years and decades. All these factors combine to press for the establishment of effective and timely governance of AI. We strongly believe that since AI will have such a huge impact all around the world, that governance should be global governance, albeit with much of the implementation at the national level.

This report therefore seeks to set out the case for effective, timely and global governance of AI. Following a brief exposition of the case for AI (Section 2), there is an initial introduction to governance and regulation, (Section 3). This is followed by a discussion of the key Values and Principles (Section 4) that have been proposed around the world, with the daunting number of sets of principles appearing to fit a more manageable group of six clusters of principles. Section 5 focuses on Social Media as an example of a sector in serious need of regulation.

There are some broader impacts that are highlighted in the latter part of the report, each one raising some fundamental issues for human life on this earth. The potential impact of AI on Work and Society is discussed in Section 6, whilst the impact on skills, including moral skills, and indeed the role of humanity in such a context is discussed in Section 7. The technical implications of humanity failing to retain control over an emerging superintelligence are discussed in Section 8 together with the steps necessary to minimise the risk of such an outcome.

Finally, there are the institutional implications of global governance. These are discussed first in the military sphere, focusing particularly on Lethal Autonomous Robots (Section 9). The overall structures and institutions that would enable suitable global governance to be established are set out in Section 10, with an overall phased summary in Section 11 and brief Conclusions in Section 12.

2 THE CASE FOR GOOD AI

We have asked ourselves “is it really necessary to make a case for AI?”. After all, it is already here, becoming ubiquitous in our lives, and growing at a seemingly unstoppable and accelerating rate. Its massive adoption growth is predicated on the benefits and advantages that this group of converging technologies is providing to businesses, individuals, communities, and humanity in its entirety. Why not just focus on the regulatory needs that we deem necessary?

- First, because failing to acknowledge and convey the importance and unique transformational nature of AI would be a disservice to the purpose of this document.
- Second, because AI’s potential future benefits to humanity are the main driving force for us to find a balanced approach to the great risks and challenges AI presents humanity with.

2.1 What AI is doing today

Today, AI is already allowing us to speak with virtual assistants around our houses and workplaces - helping us to get information, organize our activities and even helping us to execute some of them. AI-powered apps provide us with instant secured communications, enable remote transactions and help us exercise control from a distance without spending the time and effort to be in situ. This is providing significant savings and efficiencies in the daily life of hundreds of millions of people.

The Fourth Industrial Revolution (Industry 4.0) is unfolding, powered by AI, intelligent robots connected by machine-to-machine (M2M) and Internet of Things (IoT) technologies, providing new levels of automation to many industries and manufacturing operations. AI-enabled autonomous vehicles and civil drones have started to make their debuts on the road and in the skies, opening up new opportunities for transporting both people and goods. AI-predictive algorithms combined with neuroscience improve and personalize the learning experience for students and workers.

By powering real time translation tools, AI brings us closer to each other and, by creating intelligent views and interpretations of satellites pictures, AI helps us to know more about, and understand better, the planet we live in. In essence, AI gives us more opportunities at a global scale to work together to create a better world.

It is enabling us to know about the DNA of our cells and of all living things, and to begin to understand every aspect of our incredibly complex bodies. AI, combined with Health Big Data, makes it possible to understand the human genome[5], power most genetic-related research and applications[6] and better understand, prevent and cure diseases.

AI is transforming healthcare, from robot-assisted surgeries and therapies, new diagnostic methods, and personalized therapies, to improving the quality of telemedicine and enabling affordable access to medical knowledge for millions. It is enabling medical advances not only in understanding and detecting diseases[7]

but also in creating new treatments, drugs and vaccines[8]. Currently, AI is also being used to create a vaccine for COVID-19[9]. AI also helps managing healthcare workflows and making them more efficient. AI is helping us with the fight against climate change by changing the way we plan and deliver transportation, build cities and eliminate CO_2[10]. The education sector is undergoing a transformation phase, as technologies become embedded in teaching, learning and assessment[11] . This should allow education to focus on being more human[12], as the emphasis on previously undervalued social skills, such as relationship-building and empathy, will only increase going forward.

Industry sectors that are embracing AI include retail; consumer goods; food and beverage; media, entertainment and telecom; manufacturing; automotive; aerospace; Industry 4.0. as well as the services sector (financial services and healthcare)[13]

AI's benefits and unique transformational power are already acknowledged not only by the business and academic communities but by national governments and multinational bodies, such as the EU[14] and UN[15]

2.2 What AI promises in the future

Extending on the current use mentioned above, AI continues to offer further benefits in the future, by increasing efficiency, enhancing our understanding of phenomena in social, natural and economic life, both in width and in depth, and offering new services. While AI is not intended to be biased, it is originated through data and impartial structures. It is, therefore, all the more essential to address these imbalances from the get-go and constitute global governance[1617]

Three fundamental ways AI will benefit businesses in the future are the way businesses understand and interact with customers, offer more intelligent products and services, and improve and automate business processes[18]. AI will help advance the UN's Sustainable Development Goals.[19] AI will also have an impact on space exploration and astronomic discoveries. According to the National Institutes of Natural Sciences, AI will enable the categorisation of galaxies as it handles large data sets of images.[20] AI can also help us live longer and more fulfilling lives. Healthcare will take huge leaps as advances, such as "nanotechnological bodies" and "brain-computer interfaces" lower death rates.[21]

The potential for AI to deliver a massive increase in wealth at a global scale is real. McKinsey anticipates a $13 trillion increase in the global economy by 2030[22]. Breaking the economy of scarcity paradigm, which has ruled all economic systems in history, is within AI's reach. The creation and delivery of the products and services needed for every person in the planet to have a much longer, healthier, and more enjoyable life, in a sustainable manner, is at the end of the AI's promise.

As the level of intelligence of AI increases, a key milestone could be the achievement of Artificial General Intelligence (AGI) comparable to that of humans. Further increases in intelligence at this point are expected to occur more rapidly, with a potential intelligence explosion or singularity leading to the creation of superintelligence, dramatically more intelligent than humans. It may be that the world of abundance can only achieved with the aid of superintelligence. There are key issues associated with superintelligence however that are addressed in Section 7 and 8

2.3 The importance of consensus of support for Good AI

The biggest challenge we face is to find the path that will allow humanity to enjoy the outcomes of AI's promise, whilst avoiding the potential harm associated with AI. Identifying this path will not be easy: it will be slow, it will require multiple steps and most likely we will face setbacks. It is in this context that a classification of AI developments, initiatives, and applications, based on their merits and benefits as well on their risks and negative consequences to humanity, is needed.

There are many different ways of classifying AI, but efforts to build a broadly accepted and useful classification based on beneficial, or detrimental, impact to humanity are still in the early stages. The creation of such a classification may become one of the initial challenges of the governance entities that we propose in Section 10. A universally accepted classification of this nature should help us to identify "Good AI" and to promote it and accelerate its development and deployment. In the main the categorisation relates to the purpose and impact of the AI application, i.e. Good purpose and/or impact of the AI application (Good AI) and Bad purpose and/or impact of the AI application (Bad AI). This is quite distinct from AI Safety issues, which would represent a cross-cutting concern.

Good AI, like the many examples above, should satisfy three basic criteria:

a. it should provide clear and tangible benefit to individuals, communities and humanity as a whole;

b. it should comply with ethical principles (see Section 4);

c. it should post no, to low, risk to humanity's well-being and/or existence.

The identification, classification, and promotion of Good AI is essential to help mitigate the risks that AI poses to humanity. We could argue that if we task AI to make people live longer, healthier, happier, and more resilient lives we may significantly reduce the existential risk AI poses to all of us. Bad AI includes cybercrime, lethal autonomous robots (See Section 9) and fake news and other ways of individual and social manipulation (See Section 5). If we task AI to provide technologies and controls to increase human safety in the presence of Bad AI or potentially dangerous AI systems, we may gain the time, knowledge and wisdom to build a great future for humanity.

Another approach to classification is that proposed by the European Union, namely High-risk and Non-high-risk. It might be possible to combine the two approaches, starting with the Good / Bad categorisation and then applying the risk related assessment.

AI is a quickly developing field. Today's abilities of AI were considered science fiction thirty, or even ten years ago. The speed of development is expected to increase, leading to the increasing potential for Bad AI. The need for good regulation exists today, and the urgency to define and implement it is increasing. The case for AI rests on our ability, humans' ability, to effectively regulate and govern on AI to build an effective and safe path towards what it promises

2.4 Recommendation

a. A system of classification should be introduced and applied which distinguishes clearly between Good AI and Bad AI.
b. Promote, support, and ensure that AI systems are developed and deployed in areas of Good AI.

3 REGULATION AND GOVERNANCE

3.1 Competition and regulation

Nations wish to realise the benefits that AI can offer, whilst seeking to minimise any associated risks. To some degree they need to make a trade-off between these different aspects, with supporters of rapid expansion of AI often arguing that it is “too early to regulate”, whilst others are often more focused on addressing the damage done and the risks, and are less concerned if their action slows down the rate of development of AI.

The fact that most nations of the world are fully aware of the significance of AI and are actively engaging in their own processes is clearly indicated by the timeline of AI strategic documents, effective as of April 2020 (Figure 1) developed by UNICRI / CAIR (United Nations Interregional Crime and Justice Research Institute / Centre for Artificial Intelligence and Robotics).

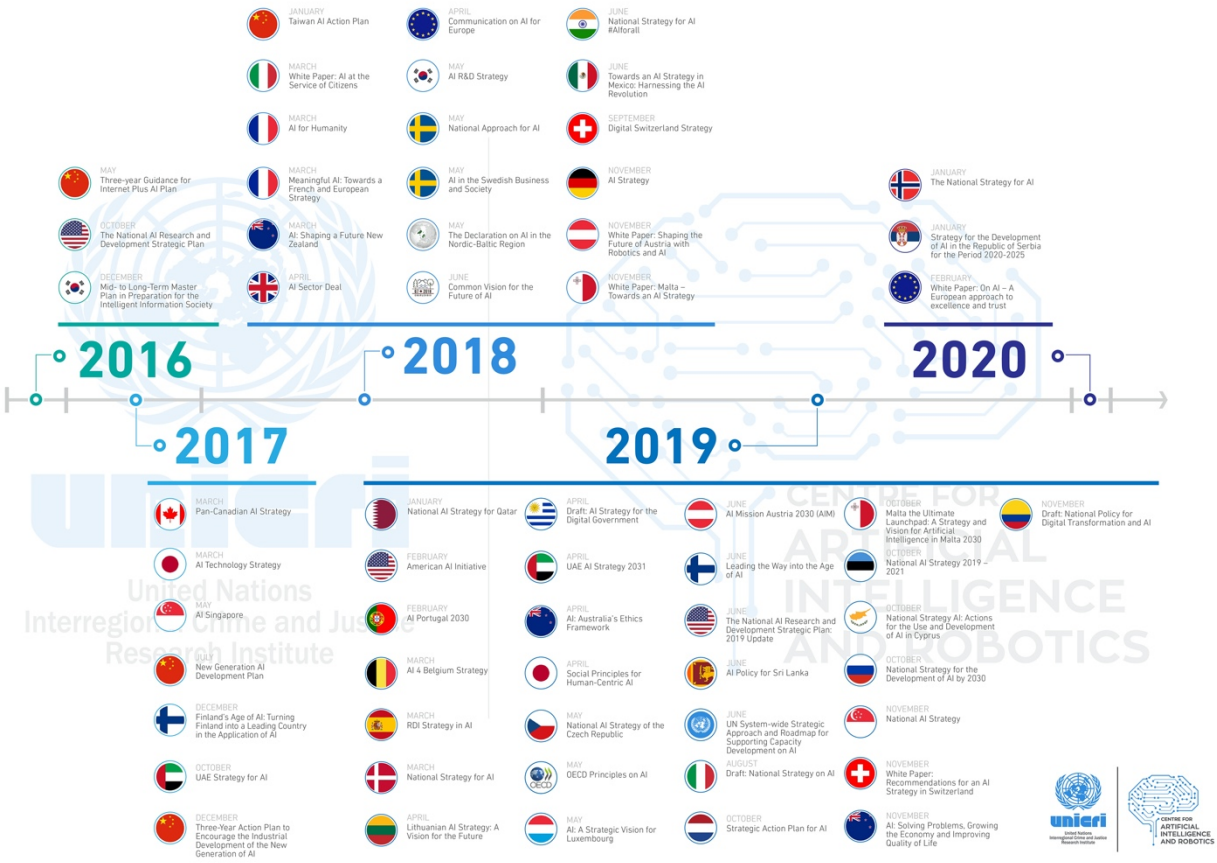

Figure 1 UNICRI Timeline of AI strategic documents[23]

A noticeable feature of many nations’ AI plans however is their competitive nature. The strategies of each of the major players in this field all stress their

determination to achieve high or higher market share, each one seeking to establish itself as a, or the, global AI leader.

China
In 2017, the State Council of the People's Republic of China formally approved "A Next Generation Artificial Intelligence Development Plan".[24] The strategic objectives of the Plan were set out in the form of three steps, namely

> "By 2020 the overall technology and application of AI will be in step with globally advanced levels…"
>
> "By 2025, China will achieve major breakthroughs in basic theories for AI such that some technologies and applications achieve a world-leading level and AI becomes the main driving force for China's industrial upgrading and economic transformation…"
>
> "By 2030, China's AI theories, technologies, and applications should achieve world-leading levels, making China the world's primary AI innovation centre, achieving visible results in intelligent economy and intelligent society applications, and laying an important foundation for becoming a leading innovation-style nation and an economic power"

US
The Executive Order on Maintaining American Leadership in Artificial Intelligence[25] by the US President in 2019 is unequivocal in its message, as section 1 states that

> "The United States is the world leader in AI research and development (R&D) and deployment. Continued American leadership in AI is of paramount importance to maintaining the economic and national security of the United States and to shaping the global evolution of AI in a manner consistent with our Nation's values, policies, and priorities."

The European Union
The EU has been somewhat less ambitious than its US and Chinese competitors, but nevertheless concludes, in its paper on Artificial Intelligence for Europe[26]:

> "The EU has a strong scientific and industrial base to build on, with leading research labs and universities, recognised leadership in robotics as well as innovative startups. It has a comprehensive legal framework which protects consumers while promoting innovation and it is making progress in creating a Digital Single Market. **The main ingredients are there for the EU to become a leader in the AI revolution,** in its own way and based on its values." (emphasis in the original).

3.2 Current international AI Governance initiatives

Building on the surge of interest in AI governance at the national level, there has over the last couple of years been a great upswelling in interest at the international level. People have started to be impacted by AI, for instance by the immediate issues discussed in Section 5, and have started to understand the need for, and started to demand, governance and regulation of AI. Many in the industry have acknowledged the need for the regulation of AI but have sought to

delay its introduction, fearful that crude constraints will be put in the way of a technology that is still evolving. But such concerns have not stopped several international organisations from seeking to engage in AI Governance.

The leading initiatives include those launched by the Council of Europe, the European Union, UNESCO, the International Telecommunications Union (ITU), the OECD, the G7 (led by France and Canada), and the IEEE. These initiatives are summarised below.

Council of Europe CAHAI
The Ad Hoc Committee on Artificial Intelligence (CAHAI), established by the forty-seven nations of the Council of Europe, became, in September 2019, the first ever intergovernmental committee group to develop a legal framework for the design, development and application of artificial intelligence. It will be based on the Council of Europe's standards on human rights, democracy and the rule of law and draw upon broad multi-stakeholder consultations. CAHAI has a two-year mandate.

In July 2020, Gregor Strojin, the Chair of the CAHAI committee, underlined the importance of working with a wide range of stakeholders active in the field of AI. "We need to be on one side, up to date on the developments of the technology and we also need to think globally and inclusively," he said[27]. The first draft of the feasibility study is due to be presented at the next CAHAI plenary meeting in December 2020. The panel will then make a decision by the end of 2020 on whether to proceed with drafting and negotiating a legally binding treaty; if it does it could be delivered within as little as two years. In order for the treaty to get the green light, all 47 member countries would need to approve it. Each country would then incorporate the new rules into their national legislation. It seems that "there is overwhelming agreement — but not yet consensus" that such rules are needed according to Jan Kleijssen, the Council of Europe's director of information society and action against crime.

In the meantime, two working groups, a policy development group and a consultations and outreach group have begun work. The policy development group is analysing existing AI applications and drawing on the experts' contributions from the meetings, "mapping risks and opportunities as well as the existing legal frameworks," Strojin. They will then develop policy proposals to feed into the work of a legal framework group whose work should start at the beginning of 2021. This group will either draft new materials or propose "soft law instruments," he said.

European Union
In April 2019, the European Union's High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence, following extensive consultation. The Guidelines put forward a set of 7 key requirements that AI systems should meet in order to be deemed trustworthy, namely Human agency and oversight, Technical Robustness and safety, Privacy and data governance, Transparency, Diversity, Societal and environmental well-being and Accountability.

The European Commission have produced a White Paper on Artificial Intelligence – A European Approach, which went out for consultation in February 2020[28], and

proposes:

a. Measures that are designed to streamline research, foster collaboration between Member States and increase investment into AI development and deployment;
b. Policy options for a future EU regulatory framework that would determine the types of legal requirements that would apply to relevant actors, with a particular focus on high-risk applications.

UNESCO
UNESCO expressed an intention in 2018 to engage with Artificial Intelligence and its Governance. UNESCO is convinced that that there is an urgent need for a global instrument on the ethics of AI to ensure that ethical, social and political issues can be adequately addressed both in times of peace and in extraordinary situations like the current global health crisis. In November 2019 they formally launched a two-year drafting process.

Twenty-four renowned specialists with multidisciplinary expertise on the ethics of artificial intelligence were tasked with producing a draft UNESCO recommendation[29] that takes into account the wide-ranging impacts of AI, including on the environment and the needs of the global south. They launched a widespread consultation in July 2020.

The UNESCO Recommendation is expected to define shared values and principles, and identify concrete policy measures on the ethics of AI in ten policy areas. Its role will be to help Member States ensure that they uphold the fundamental rights of the UN Charter and of the Universal Declaration of Human Rights and that research, design, development, and deployment of AI systems take into account the well-being of humanity, the environment and sustainable development.

The final draft text is expected to be presented for adoption by Member States during the 41st session of UNESCO's General Conference in November 2021 and, if adopted, it will be the first global normative instrument to address the developments and applications of AI.

ITU AI4Good
ITU'S AI4Good initiative is not an example of governance or regulation but rather it is one of the leading action-oriented, global and inclusive United Nations platforms on AI. With less than 10 years to achieve the United Nations' Sustainable Development Goals (SDGs), AI is seen as holding great promise by capitalizing on the unprecedented quantities of data now being generated on sentiment behaviour, human health, commerce, communications, migration and more.

The AI for Good Summit is organized every year by the ITU with XPRIZE Foundation in partnership with over 35 sister United Nations agencies, Switzerland and the Association for Computing Machinery (ACM). The goal is to identify practical applications of AI and scale those solutions for global impact.

OECD

The OECD launched its own AI related initiative some years ago and established an Expert Working Group AIGO. In February 2020 it launched its AI Observatory. Its OECD Principles on Artificial Intelligence, adopted in May 2019 by the OECD member countries, promote artificial intelligence that is innovative and trustworthy and that respects human rights and democratic values. The principles formed the basis of the human-centred principles adopted by the G20 in June 2020[30].

The OECD Network of Experts on AI (ONE AI) is now moving from principles to practice for trustworthy AI and has set up working groups to

a. Devise a simplified, user friendly classification of trustworthy AI and an AI knowledge graph for AI approaches.
b. Look at new indicators to measure the adoption of trustworthy AI
c. Identify examples of implementation and good practices for the values-based OECD AI principles
d. Identify and develop good practices for implementing the OECD AI Principles' five recommendations to policy makers to invest in AI R&D, foster a digital ecosystem for AI, shape an enabling policy environment for AI, building a human capacity and preparing for labour market transformation and international cooperation for trustworthy AI.

Global Partnership on AI

The two consecutive Presidencies of the G7 in 2018 and 2019 were Canada and France. Together they developed an AI initiative and in May 2020, G7 countries agreed on launching the Global Partnership on AI (GPAI) to "enhance multi-stakeholder cooperation in the advancement of AI that reflects our shared democratic values and addresses shared global challenges, with an initial focus that includes responding to and recovering from COVID-19, and committing to the responsible and human-centric development and use of AI in a manner consistent with human rights, fundamental freedoms, and our shared democratic values".

It is intended to guide the responsible development and use of AI with a Secretariat hosted by the OECD in Paris and with Centres of Expertise in both Montreal and Paris. The relationship with the OECD is intended to bring strong synergies between GPAI's scientific and technical work and the international policy leadership provided by the OECD, strengthening the evidence base aimed at responsible AI. There are 15 initial members, including all G7 nations and such countries as India, South Korea and Singapore – but not China or Russia.

The GPAI will initially be comprised of four working groups focused on responsible AI, data governance, the future of work, and innovation and commercialisation.

IEEE

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS) is a program of the Institute of Electrical and Electronics Engineers (IEEE), launched to address ethical issues raised by the development and dissemination of these systems. It identified over one hundred and twenty key issues and provided candidate recommendations to address them. In addition, it has provided the inspiration for fourteen standardization projects that are currently under development with the IEEE Standards Association, as well as four certification programs for systems and the organizations that deploy them.

Individual stakeholders
In the process of developing these governance initiatives, regulations and principles, the various bodies have been supported by a multitude of stakeholders, helping experts to produce first drafts, providing feedback in consultations and producing their own White Papers. Google's "Perspectives on Issues in AI Governance" is a case in point.

3.3 Future governance and regulation

The competitive nature of the different national (and supranational in terms of the EU) strategies is completely understandable. But it is far from clear that the establishment of regulation to minimise risk and keep the world safe is best created within a competitive environment. A far better way would be to establish a framework for regulation within a global context, where the trade-offs can be discussed and established rationally.

Cross border links:
The European Union's data privacy law (General Data Protection Regulation, better known as GDPR) which took effect in 2018, has an impact way beyond the borders of the European Union: it extends many of its data privacy safeguards to processors globally processing data of persons in the EU, as well as to within-EU processing of data from persons around the world.

An acute need for regulation:
An example of a sector that is in acute need of regulation is Social media. The situation is set out in Section 5. A global governance framework can prioritise and facilitate the introduction of appropriate regulation of AI in this domain. See also sections 6 to 9 for additional domains that show risks to humans and humanity that are imminent on the short to mid-term range.

3.4 A dichotomy and the way forward

There is a dichotomy, indeed more than one. For some, it is too early for detailed regulation of AI; they argue that governance should be limited to the application of agreed principles. For others, the need for regulation is clear and urgent, at least in some domains / respects. For some people, new compulsory requirements should be limited to high-risk AI applications, whereas for others the first priority is to address Bad AI (Section 2.3).

One approach might be to establish a basic skeletal standard, something akin to WTO trading rules – basic agreements on GAAP and what constitutes an acceptable tariff. These would be metarules which create acceptable bounds for economic conflicts, whilst still allowing for some arbitrage and innovation within individual nations.

The aim of this paper is not however to resolve all these dichotomies. There is clearly a great deal of work to be done. The aim of this paper is to explore and make the case for a global governance framework for AI (and indeed potentially for other disruptive technologies), to raise some issues that would need to be addressed by negotiators of such a framework, and to raise some issues that need to be born in mind by policy makers now and progressively in the future.

The possible governance structures and institutional implications of such an approach are discussed in Section 10. Such a framework would need to rest upon a set of agreed global values and principles, and the challenge of establishing such a set is discussed in the next section.

4 VALUES AND PRINCIPLES

4.1 Introduction

In recent years, growing concern over the social, ethical, economic, and environmental implications of AI technologies has led to a proliferation of documents from national and supra-national governing bodies, academia, and industry setting out their values and principles for the development, deployment and governance of AI[31]. While these documents bear different titles, like guidelines, principles, concept, or white paper, we will use the term "principle set" because of the core content of the documents. Our goal is not to add to the confusion by producing yet another set of principles, but to highlight commonalities and convergence between the existing sets by identifying principle clusters, following the work of five previous clustering attempts ((Zeng, Lu and Huangfu, 2019), (Jobin, Ienca and Vayena, 2019), (Fjeld et al, 2020), (Floridi et al, 2018), (Royakkers et al, 2018)). Thus, we argue that developing a singular, globally agreed set of values and principles is not as complex a task as it first appears.

Section 4.2 discusses the distinction between values and principles. Section 4.3 sets out the principle clusters with a brief outline and the general themes of each cluster. Section 4.4 demonstrates the validity and usefulness of the principle clusters by showing the correspondence of two principle sets (OECD[32] and UNESCO[33]) to the clusters defined in 4.3. Section 4.5 explains the method used to identify the clusters and their themes, as well as the formation of the outlines. Section 4.6 reflects upon the cultural aspects of principles with Section 4.8 drawing some conclusions.

4.2 Values and Principles

Values as a foundation

Before introducing the main clusters of principles identified in 4.3, it is worth briefly discussing the values upon which these principles are based and the different approaches that existing sets take in defining them.

Some principle documents, notably UNESCO's *Recommendation on the Ethics of Artificial Intelligence*, make a clear distinction between values and principles by providing a separate list and then defining their principles in terms of the values that they have set out. UNESCO describes its values, which include Ensuring Diversity and Inclusiveness, and Environment and Ecosystem Flourishing, as "motivating ideals in shaping policy measures and legal norms" that act as "the foundations of principles"[34]. Similarly, the European Commission's *Ethics Guidelines for Trustworthy AI* describes its ethical principles and key requirements as "rooted in fundamental rights", which include respect for human dignity and freedom of the individual[35]. UK government guidance on the use of AI[36] refers back to the SUM values (Respect, Connect, Care, Protect), and the FAST track principles which draw on these values but are more "specifically catered to the actual processes involved in developing and deploying AI systems"[37].

In these cases, values are argued to be a more abstract moral and ethical framework that principles build upon to enable the transition into actionable policy recommendations.

Within the clusters

In general, frequent reference is made to values within principle sets, mostly those falling under the cluster identified in section 4.3 as Beneficence, which can be loosely defined as promoting and protecting human values and includes themes of sustainability, equity, and the common good. Of the principle sets used by Fjeld et al[38] in their analysis, 69% included themes relating to human values; similarly, the Humanity/For Human cluster had some of the highest topic coverage in Zeng, Lu and Huangfu[39]. In general, principles relating to Beneficence are often the first in the list[40].

Though a few, including IEEE's *Ethically Aligned Design* and Google's *AI Concepts*, argue that actively doing good is a necessary precondition for the development and deployment of AI technologies,[41] most existing principle sets emphasise protecting human values against the potential harms of AI, rather than on how AI could potentially promote those values.[42] Thus, it can be argued that values take a more positive view (the "ideals") and principles are more realistic in transforming these values into actionable statements upon which governance can be based. This is reflected in UNESCO's first principle, which requires proportionality in the use of AI systems in meeting any goal and that and harm prevention should be prioritised, suggesting that there are limitations on the extent to which AI can actively do good relative to human values.[43]

The principle documents that do not include their own set of values but make reference to human values within the principles, frequently define them in the context of political, civil, and labour rights, and with reference to existing legal frameworks such as the UDHR.[44] Interestingly, the European Commission makes a distinction between legal and ethical AI, arguing that since the law cannot always keep up with developments in technology or ethical norms and, additionally, can't fully address all issues, AI systems must be aligned with human values beyond compliance with the law. [45] This suggests that values can be viewed as a moral framework for all actors involved in the creation and use of AI technologies, aiming to shape behaviour alongside regulation that enforces it. This idea is echoed by UNESCO with the argument that their value set "inspires desirable behaviour" but that their principles are what governance is based on.[46] Similarly, the objective of the Support, Underwrite, and Motivate (SUM) values is to enable those involved in AI projects to consider the potential impacts and evaluate the project's "ethical permissibility."[47]

Two categories of principles

Looking at the principle clusters set out in the next section, or indeed the ones presented by the previous clustering attempts, it can be seen that principles fall into one of two broad categories. Firstly, ethical principles which are closely linked to values, often building directly on them in the cases where principle documents make a distinction between the two. This category can be summarized as follows: AI technology must do good (either simply protecting human values such as sustainability, equality, justice, etc. or actively promoting them) and it must do no harm (either by undermining human values, increasing inequality, causing

injustices etc. or via the creation of new threats such as LARs or rapidly self-improving AGI). The second category are principles such as Accountability and Transparency, which are not necessarily ethical principles in their own right, but are crucial in enabling and ensuring the application of the previous category. Indeed, Floridi et al[48] specifically define the Explicability cluster, which includes both transparency and accountability, as "a new enabling principle for AI" which builds on the core principles of bioethics (from which the SUM values also take inspiration[49]). Furthermore, Jobin, Ienca and Vayena[50] argue that Transparency's "pro-ethical nature" could underly the high frequency of references to it in principle sets relative to other ethical principles.

Conclusion

Values are more abstract than principles, take a more positive view, and can act as ethical guidance in all stages of the life cycle of AI systems beyond what is required by law. In comparison, principles provide more concrete and practical recommendations to aid in governance, going beyond values to include ideas of transparency and accountability that assist in the upkeep and/or promotion of those values. The process of agreeing upon a global set of principles should thus involve the negotiation of a global set of values.

4.3 Principle Clusters

Introduction

The rationalisation process that produced the principle clusters set out here was guided by five previous attempts to analyse and cluster principle sets ((Zeng, Lu and Huangfu, 2019), (Jobin, Ienca and Vayena, 2019), (Fjeld et al, 2020), (Floridi et al, 2018), (Royakkers et al, 2018)). Taken together they provide much evidence of convergence, and the clusters identified here are broadly supported by their work, as discussed in detail in section 4.5.

The brief outlines shown in Table 1 are intended not to provide concrete definitions for each cluster (for in doing so, we would simply be producing a new set of principles), but to aid understanding of what each cluster broadly means and the general areas of commonality that it covers. The precise honing of the outline and themes of these clusters, along with subsets of principles within each cluster, can be negotiated globally starting from this base of six clusters, leading to the development of one coherent set upon which to base binding regulations.

Clusters	Outline	Themes
Beneficence	AI technologies should seek to protect and actively promote human values and human dignity. The potential social, environmental, and economic benefits to society should be prioritized and shared equally.	Education, sustainability, access to technology, human well-being, common good
Justice and Fairness	AI technologies should do no harm in terms of undermining human values or violating human rights. Bias in the design and deployment of AI should be mitigated to prevent discrimination, and technologies should respect diversity, inclusion, and fairness.	Representative and high-quality data, freedom, equality
Safety and Security	AI technologies should do no harm in terms of violating the privacy of individuals, causing them economic or social harm, or producing new risks to humanity. Precautions should be taken to mitigate the impact of the overuse, misuse, or bad design of AI, including consideration of the potential long-term and future impacts.	Anti-LARs, self-improvement, risk-management, data protection, resilience to external threats, AGI/ASI
Human Autonomy	AI technologies should protect and promote human autonomy and the right of individuals to choose how and if they interact with AI. Human control of technology should be maintained with the autonomy of machines limited, though the constraints placed on AI systems should be proportional to the level of automation that the system is capable of.	Freedom, consent, self determination
Accountability	The creators and users of AI technology must do so responsibly, respecting the first four principles in all stages of the lifecycle of AI systems. When negative outcomes do occur, it must be possible to identify the actors responsible and hold them to account.	Monitoring, balance of powers, trust, right to appeal automated decisions, legal liability
Transparency	AI technologies must be transparent and explainable to judge whether they are abiding by the first four principles and, in case they are not, to identify how and why, as well as who is responsible in order to ensure Accountability.	Explainability, traceability, right to information, regular reporting, communication

Table 1: The Six Principle Clusters

4.4 Applying the framework to two sets of principles

Principles	Principles for responsible stewardship of trustworthy AI[51]	Recommendation on the ethics of Artificial Intelligence[52]
Beneficence	**Inclusive growth, Sustainable Development and Wellbeing:** "*Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet*" **Human-centred Values and Fairness:** "*AI actors should respect the rule of law, human rights and democratic values*"	**Proportionality and do-no-harm**: ensuring "*the implementation of procedures for risk assessment and the adoption of measures in order to preclude the occurrence of*" potential harms to humanity or ecosystems **Sustainability: "***social, cultural, economic and environmental impact*" of AI should be considered relative to the UN SDGs
Justice and Fairness	**Human-centred Values and Fairness:** including "*non-discrimination and equality, diversity, fairness, social justice"*	**Fairness and non-discrimination: "***AI actors should promote social justice, by respecting fairness [which] implies sharing benefits of AI technologies***"**
Safety and Security	**Robustness, Security and Safety**: "*in conditions of normal use, foreseeable use or misuse, or other adverse conditions [AI systems must] function appropriately and do not pose unreasonable safety risk*"	**Safety and Security: "***unwanted harms (safety risks) and vulnerabilities to attacks (security risks) should be avoided***"** **Privacy:** including protecting "*the right for individuals to have personal data erased*"
Human Autonomy	"*AI actors should implement mechanisms and safeguards, such as capacity for human determination*"	**Human oversight and determination:** *"the decision to cede control in limited contexts remains that of humans"*
Accountability	**Accountability:** "*AI actors should be accountable for the proper functioning of AI systems and for the respect of the [other] principles*"	**Responsibility and accountability:** "*The ethical responsibility and liability for the decisions and actions based in any way on an AI system should always ultimately be attributable to AI actors*."
Transparency	**Transparency and Explainability:** "*AI actors should commit to transparency and responsible disclosure regarding AI systems*" including making stakeholders aware of "*their interactions with AI systems*" and enabling them to understand and challenge the outcomes of AI systems	**Transparency and explainability: "***a crucial precondition to ensure that fundamental human rights and ethical principles are respected, protected and promoted*" **Awareness and literacy: "***Public awareness and understanding of AI technologies and the value of data should be promoted through open and accessible education"*

Table 2: Comparing Two Principle Sets to the Six Clusters

The OECD's *Principles for responsible stewardship of trustworthy AI* and UNESCO's *Recommendations on the Ethics of Artificial Intelligence* were chosen for comparison with our principle clusters as they are two examples of highly visible and influential principle sets from multi-stakeholder organisations. In addition, high level experts and wide-ranging consultation processes were involved in the creation of both sets. As shown in table 2, both sets map closely to the six principle clusters identified in section 4.3, providing strong evidence for the convergence around those clusters.

The OECD principles are the first set of its kind to be officially adopted by governments, having been agreed upon by its 36 member countries as well as Argentina, Brazil, Costa Rica, Malta, Peru, Romania and Ukraine[53]. Furthermore, the G20 principles, agreed upon in June 2019 at the Osaka Summit, are drawn directly from the OECD principles[54]. Both America and China, two major players in the development of AI technologies, are signatories to the G20 principles, highlighting that consensus can be achieved despite cultural differences previously believed to be insurmountable.

The UNESCO principles have been submitted to its 193 members states and are due to undergo further negotiation before their November 2021, when they are planned to be adopted by members at the General Conference[55]. The first draft of recommendations was published only very recently, on 7 September 2020, providing evidence of a trend towards increasing convergence as argued by Fjeld et al[56], who found that more recently published documents tend to cover more of the principle clusters.

The OECD principles were developed by an expert group with over fifty members, including representatives from over twenty countries, each of the three main sectors currently authoring principle documents (government, corporate, academia) as well as from civil society[57].The UNESCO principles were developed through a three stage online consultation process: open to the public, collaborating with host countries in each of its regions and subregions, and with partners for "open, multistakeholder, and citizen deliberation workshops"[58].

Both sets were influenced by extensive feedback from a diverse range of views, geographies, and sectors, and thus the commonalities between them and the principle clusters identified here demonstrate the emergence of a broad consensus on AI ethics.

4.5 Process of establishing the clusters

Identifying the Clusters

The rationalisation process that produced the principles set out in 4.3 was guided by five previous attempts to analyse and cluster principle sets. Zeng, Lu and Huangfu[59] with the Linking Artificial Intelligence Principles (LAIP) project use a manually chosen set of keywords under 10 general topics to measure the coverage of these topics across 74 different proposals. Jobin, Ienca and Vayena[60] compared 84 sets and identified 11 general themes, with convergence around the top five. Fjeld et al[61] examined 36 particularly influential documents selected for variety in terms of geography, stakeholder, approach and more, and collated 47 principles clustered into 8 main themes. Floridi et al[62] produce 5 main principles based on analysis of 6 key principle documents from multi-stakeholder

organisations and the 4 core principles of bioethics. Lastly, the research of Royakkers et al[63] reveals 6 recurring themes emerging from a review of scientific literature on 6 dominant technologies, including the Internet of Things, biometrics and virtual reality. As a reference, Algorithm Watch have logged over 160 sets of AI principles.[64]

The principles in 4.3 appeared in each of the clustering attempts discussed, either directly (e.g. Transparency was a key theme for Zeng, Lu and Huangfu[65], Jobin, Ienca and Vayena[66], and Fjeld et al[67]) or as a keyword under a thematically equivalent principle (e.g. Accountability is a keyword for Jobin, Ienca and Vayena's[68] principle of Responsibility, whereas Zeng, Lu and Huangfu[69] counted mentions of responsibility as evidence for the Accountability theme. Floridi et al[70] categorised both of these concepts along with transparency under Explicability). Principles that were frequently linked in the above literature or in main principle documents, such as Safety and Security or Justice and Fairness, are here grouped together for simplicity. Similarly, where some previous clustering attempts defined a lesser theme as its own principle, such as Sustainability[71] or Privacy[72], but others listed it as a subtopic (Fjeld et al[73] in Promotion of Human Values and Floridi et al[74] in Non-maleficence respectively), this attempt has favoured the latter approach.

Formulating the outlines
Each of the previous clustering attempts devotes significant space to exploring the main themes within the principles and the similarities and/or differences in how subtopics such as privacy are defined across the source document sets. Whilst there is convergence around broad principle clusters, there is some divergence in the details of these principles, in terms of how they are interpreted and the relative importance of different subtopics.[75]

In the interests of simplicity, the general outlines for the principle clusters set out in Section 4.3 are drawn from commonalities identified by each of the previous clustering attempts. For example, Fjeld et al[76] note that 89% of the principle documents within their dataset make reference to "non-discrimination and the prevention of bias". Similarly, according to Jobin, Ienca and Vayena[77] 28/84 of the principle documents define justice and fairness in terms of "prevention, monitoring and mitigation of unwanted bias [and/or] discrimination". This commonality led to the wording "*Bias in the design and deployment of AI should be mitigated to prevent discrimination*" under the Justice and Fairness principle cluster. The outlines also show the general pattern across principle sets discussed in the Section 4.2.

Identifying the themes
The themes column contains topics that were listed as separate principles by some of the previous clustering papers e.g. sustainability[78], balance of powers[79], as well as keywords used to find evidence of a principle e.g. education, AGI/ASI[80], communication, self-determination[81], and subtopics/principles that lie within a cluster e.g. right to information, representative and high quality data.[82]

4.6 Cultural differences

An interesting debate took place at the AI Summit 2020[83] on whether AI is a cultural matter. For Nicolas Miailhe, of The Future Society, yes it clearly is, with

the Europeans taking a precautionary and technophobic approach, the Global South technophilic and both China and the US bolder and technophilic.

For Bing Song, of the Berrgruen Institute, the answer was yes and no. She broke the issue down into three layers. The bottom layer, the foundational layer was focussed on holistic issues, e.g. existential risk; the intermediate layer was focused on prioritisation and the culture of governance; whilst the surface layer reflected such things as modernisation and seeking prosperity. She considered that across the world, there was no difference at the foundational layer – and indeed at the surface layer little evidence of cultural differences, with for instance wide support for the Sustainable Development Goals. It is at the intermediate layer that she found that the big difference lies.

She suggested that the West is seeking to export its value set whilst China is not willing to embrace a set of principles that include Democracy and Human Rights – despite signing up to the G20 Principles. There is the need for a common set of values and principles: progress has been made but there is clearly still some way to go.

4.7 Universalizable

To work effectively as global principles, this negotiated set of AI principles would need to be universalizable, that is to say they should be able to be applied across virtually all cultures and all time. They should not be associated with any particular culture in history: they should apply to everyone equally, with few exceptions except for minors or the infirm.

4.8 Conclusions and Recommendations

Despite the more than 160 different sets of AI Principles, there is strong evidence of a set of six clusters of principles, as set out in Table 1. Four flow from values, whilst the other two enable the first four to be achieved. Responsible AI is not aided by the pursuit of a myriad of different principles which can only lead to confusion and non-achievement. But, within each of these six clusters of principles, a set of principles can be negotiated and agreed. The evidence is that this would not be as hard as it might initially seem.

It is recommended that a single global set of universalizable values and principles be negotiated and adopted by all.

5 SOCIAL MEDIA AND THE INFOCALYPSE

5.1 The Problem

There is no sector, at the present time, more in need of regulation than social media. We are experiencing what some have called an Infocalypse[84] or a Crisis of Epistemology. This crisis is due to a number of factors not all associated with AI. But AI tools are fuelling the fire: this is extremely harmful and must be stopped.

The attention of children and many adults is being diverted away from real living and experience to the vicarious joys of living with a screen. US teens and tweens (8-12) have been spending 9 hours and 6 hours a day respectively with digital media, excluding time spent at school or for homework.[85]

Increasingly over the past decade people have started to live in worlds where they have a different set of facts. They are targeted with news that is tailored to their own prejudices, and increasingly they are being fed disinformation rather than information. This creation of silos of thinking is leading to a greater polarisation of society. People are being manipulated on a massive scale by social media executives and by third parties, all of whom have their own selfish goals, which are quite distinct from those of the people who are being manipulated. The truth is being lost.

5.2 The cause

A new social media business model has emerged in the past decade or so, built around the vast computing power that is increasingly available today. The business model is focussed on attention and the basic psychological need for humans to interact with each other. Interaction triggers dopamine, a key human reward system developed over evolutionary time scales and not easy to modify, in contrast to an algorithm that can be tweaked in moments.

The model typically consists of three algorithms, each with a controller / team of controllers, namely

a. An engagement algorithm: maximising spins and clicks as evidence of attention engaged.
b. A growth algorithm: maximising growth of the social network
c. An advertising algorithm: maximising advertising revenue.

Markets have developed selling adverts per click. Fake news travels several[86] times as fast as the truth. We are psychologically hardwired to be interested in things that are surprising or new. Fake news therefore generates more clicks, travels faster, and generates more advertising revenue than the truth. The overall model can be described as a disinformation model rather than an information model. The platforms have no sense of what is the truth and what is not – what they do know about is the number of clicks a piece of "news" receives. It is the gradual, slight, imperceptible change in the user's behaviour and perception that is the product[87] that the platforms sell to advertisers.
The click markets undermine democracy and undermine freedom[88].

The apparent mindset of key influencers
The mindset of the creators of the platforms was famously exposed by Sean Parker, the first President of Facebook, who explained that "it's a social validation feedback loop…The inventors understood this consciously… it's me, it's Mark [Zuckerberg], it's Kevin Systrom on Instagram, it's all of these people ... and we did it anyway."[89]

The mindset of developers of Deepfakes is exemplified by the views of the developers of Lyrebird who saw the opportunity to write some code to produce a significant result, without taking into account the fact that this may not be beneficial to society. It became the cool thing to do, rather than the right thing to do. The developers explained that if they did not do it someone else would[90] – another sort of tragedy of the commons. The need for regulation is crystal clear.

The mindset of national activists would seem to be to create as much disruption and damage to nations with which the relevant nation is competing. The active deployment by Russian activists[91] of AI controlled bots to cause disruption in the elections and referenda of various states is an example of the aggressive deployment of these tools, to which the west is just starting to develop defences[92].

5.3 It is getting worse

Of great concern is the fact that at the moment the trend is simply for the above phenomenon to get worse. The AIs are going to get better at predicting what to put on the screen, not worse. The businesses behind these platforms are now the most valuable by market capitalisation. There are strong financial incentives for the growth of AI to continue, and accelerate.

Additional elements

To this infocalyptic mix is being added a whole series of new elements, such as synthetic media abuse including dark patterns[93], algorithms told to solve for obfuscation[94], deboosting, information gerrymandering[95], on-device virtual moderating, opinion necromancy, linguistic signatures[96], voodoo dolls[97], false feedback and zersetzung.[98]

5.4 What can be done with Social Media?

The US Government has up to now resisted calls for the Big Tech companies to be broken up, concerned that they would be weakened in so doing, and be less able to compete with powerful new Chinese companies in the development of AI in the future[99]. The appointment of a judge on October 21st 2020 to hear an Anti-Trust case brought against Google however is an indication that there is significant support in many parts of the US system for standing up to Big Tech. The fact that Chris Hughes[100], the co-founder of Facebook, is actively calling for Facebook to be broken up is another indication. These moves are significant in terms of the level of competition within the sector, but they do not in any way address the problems being experienced by the users. These problems remain.

Facebook and regulation

There is much that can be done however to address the problems, and indeed industry leaders such as Mark Zuckerberg (Facebook) have stated that regulation would be welcomed. In March[101], he wrote, “Lawmakers often tell me we have too much power over speech, and I agree.” He called for more government regulation — not just on speech, but also on privacy and interoperability, the ability of consumers to seamlessly leave one network and transfer their profiles, friend connections, photos and other data to another. Hughes suggests that “Facebook isn’t afraid of a few more rules”[102]. This is the challenge: to bring about transformative regulation that will stick. Zuckerberg’s apparent enthusiasm for more rules does not mean that as the proposed regulation is formulated, there will not be a huge lobbying exercise to limit the impact. Such regulation will be dramatically more effective if the US and China are on board.

In terms of data privacy, Hughes argues for a new agency in the US[103], empowered by Congress to regulate tech companies. Its first mandate would be to protect privacy and he suggests that a landmark privacy bill in the United States should go beyond GDPR and specify exactly what control Americans have over their digital information, require clearer disclosure to users and provide enough flexibility to the agency to exercise effective oversight over time. In addition, the

agency should create guidelines for acceptable speech on social media, and be charged with guaranteeing basic interoperability across platforms.

More broadly
The suggestions above relate primarily to changes in the US regulatory framework, reflecting the leading role played by US Social Media companies. But China has major platforms of its own and Russia is very sophisticated as a user of social media. In practice, all regulatory action should be within a global framework. It needs to embrace the platforms, the developers and users of tools that can facilitate disinformation or other harmful action, and the users of the platforms themselves. In Shoshana Zuboff's view, models designed to demand addictive attention should be banned[104].

Epistemological crisis
One should realise that overcoming the Epistemological Crisis is likely to require many measures that are not directly connected to AI, such as

a. Making the intentional or careless communication of disinformation illegal and subject to a fine
b. Making a communication that is knowingly false illegal and leading to a fine or other remedy.
c. Quasi-real-time validation for public speaking.
d. Credibility ratings for sources of information.

5.5 Recommendations

Recommended actions include:

a. Regulation of the sector including
    i. The certification of systems and companies that make them.
    ii. The certification of sufficient knowledge and awareness of professionals practicing within various domain capacities
    iii. Ensuring that the global set of principles agreed includes the principle that the AI products be designed to operate humanely
    iv. The use of generative technologies such as Generative Alternative Networks (GANs) in creating images, sounds and videos should be banned without clear and indelible labelling of the use of such technologies. Such labelling should be easily recognized by both humans and machines.
b. Strengthening privacy laws and the adoption of privacy protecting technologies.
c. Access to or the formation of an International Agency able to carry out monitoring and inspection (see Section 10.7)
d. A tax on data collection and/or compensation to the data owners.

If necessary, the use of attention models themselves could be banned.

6 AI ACTIVITIES REPLACING HUMAN WORK

6.1 Introduction

AI is currently impacting and projected to impact labour in many ways. On the positive side, it offers the opportunity to automate certain tasks and focus the human work on other tasks (e.g. higher-level strategies, emotional aspects,

creative aspects). It can enable some of the more unattractive roles to be automated, freeing up people from unpleasant menial jobs. On the other hand, it has been found to display serious bias: researchers are currently seeking means to eliminate such bias.[105]

There are two key questions however with regards the impact of the deployment of Artificial Intelligence on human labour. The first addresses the scale, nature and longevity of the impact on employment of an increasingly capable AI. This question is addressed in Section 6.2. The second relates to how should society respond, if one accepts the likelihood of a radical reduction in paid employment. This question is addressed in Section 6.3. None of the issues will happen overnight but the potential impact long-term is truly transformative.

In conclusion, Section 6.4 seeks to determine the implications of the above on the global governance and regulation of artificial intelligence including the question of timing and phasing.

6.2 Scale of impact anticipated

A number of economic "revolutions" have occurred including the First Industrial Revolution (steam based), the Second Industrial Revolution (oil and electricity based) and the Third Industrial Revolution, the Digital and Communications Technology Revolution. Each time there has been concern that the result will be widespread structural unemployment. In each case there has been turbulence in the labour market but in time new job opportunities have arisen as a result of the new technologies.

We are now entering the Fourth Industrial Revolution, characterised by a fusion of technologies that is blurring the lines between the physical, digital and biological spheres - and with Artificial Intelligence at its heart. The same concerns about jobs are recurring. Not all agree however as to the most likely impact. Luis Cevasco et al[106] for instance write about the introduction of narrow Artificial Intelligence in the field of public goods and argue that the impact improves the productivity of those working in the field, rather than taking their jobs: this is less likely however to be the case in a competitive commercial situation.

Whilst some authors such as Eager et al[107] and Gartner[108] take a positive view of AI's societal and economic impacts, most forecasters argue that this time it is likely to be different from previous "revolutions" and that the impact on employment of an increasingly capable AI will transform the role of work within human society.

Evidence of a major reduction in employed hours

A number of authors have suggested that artificial intelligence will have a major impact on employment including McAfee and Brynjolfsson[109] (The Second Machine Age), Richard and Daniel Susskind[110] (The Future of the Professions), Daniel Susskind[111] (World Without Work), Jerry Kaplan[112] (Humans Need Not Apply), Federico Pistono[113] (Robots will Steal Your Jobs But That's OK) to name but a few.

In 2013, Carl Frey and Michael Osborne[114] from Oxford University carried out a study of employment in the United States and estimated that some 47% of US employment (some 64 million jobs) had the potential to be automated within

"perhaps a decade or two". In 2015, the management consultants McKinsey[115] published an interim report that addressed not whole jobs but tasks. They estimated that with current technology, 45% of tasks could be automated. With for instance, machine natural language comprehension at the median human level, this percentage would rise to 58%.

The watershed book was The Rise of the Robots by Martin Ford[116], a writer with 25 years software development experience. Written in 2015, Ford identified evidence for six disturbing trends, summarised below:

a. Stagnant wages: in the US, the change in the weekly earning of a typical worker, that is production workers and non-supervisory workers in the private sector (more than half the work force) declined 13% between 1973 and 2013.
b. A Bear Market for Labour's Share and a Raging Bull for Corporations: Labour's share of national income declined in 38 out of 56 countries studies, including the US, Japan, France, Germany and China, whilst the story for corporate profits was very different.
c. Diminishing Job Creation, Lengthening Jobless Recoveries, and Soaring Long-Term Unemployment: The rate of new job creation per decade in the US has fallen steadily from over 30% in the 1960's to 0% in the 2000's: that is before allowing for increased labour supply (at least 9 million in the 2,000's).
d. Soaring Inequality: the income disparity between the richest and everyone else has grown steadily in industrialised countries since the 1970's. In the US between 1993 and 2010, over half the increase in national income went to the top 1%. Between 2009 and 2012 that figure rose to 95%.
e. Declining Incomes and Underemployment for Recent University Graduates: in the US, half of new graduates are unable to find fobs that utilize their education and get them on the first rung of the career ladder.
f. Polarisation and Part-Time Jobs: there is significant evidence of polarisation, an organic process linked to the business cycle. Routine jobs are eliminated for economic reasons during a recession. But then, during the recovery, businesses find out that with ever advancing technology, they can get by without rehiring.

One can envisage other possible explanations of these trends such as globalisation, financialisation and politics. But on investigation, the arguments do not stand up. Ford argues persuasively that the impact on employment from automation has already begun, but that the impact has been masked.

Why should it be different this time?
In the past, it has been argued that machines are tools to help improve the productivity of the worker. This was not of course a smooth process and there were peaks of unemployment with the initial advent of a new technology, leading to troughs as the wealth generated by the improved productivity led to the creation of new jobs. During the industrial revolution productivity surged ahead as workers were made far more productive with the use of machines.

With the information and communications revolution, the Third Industrial Revolution, things started to change. Jeremy Rifkin's[117] book The End of Work was written with only a passing reference to artificial intelligence and twenty years

before the Fourth Industrial revolution was even conceived. Now however with Artificial Intelligence, Ford[118], Susskind[119], Chace[120] and others argue that the machines are replacing the workers themselves, whether their role is manual or cerebral.

Shannon Vallor's[121] description of the evolution of humans in relation to work is slightly different, starting from an earlier stage in the process

a. "First we specialized human knowledge and muscle into particular humans.
b. Next we specialized human knowledge and muscle into machines.
c. Now we are specializing intelligence itself into machines (perhaps thereby leaving humanity without a job?)."

Ford[122] describes how automation and AI have been taking and are projected to take the jobs of the middle classes – and are now increasingly taking the jobs of the elite. In the past, people have seen education as the ladder to climb out of the problem. Ford and others argue however that the analogy is not a ladder but a pyramid, with surrounding water rising. There is limited room at the top: those who have considered themselves immune to such problems will be increasingly affected.

Ford[123] sees the impact of AI as an extension of the impact of automation, but far deeper and more profound.

Education
Re-education and retraining are the historic solutions to addressing unemployment when motivated workers have lost their jobs due to a decline in a particular sector and wish to re-enter the market economy. Because the change in the job market will be gradual, appropriate education options for young people will be needed, i.e. education institutions should start adapting their course offerings and curricula now. Several have already started. Such an approach is totally relevant in employment and economic terms if there are jobs available in the market economy for which one can be re-educated / retrained.

But if there are no jobs in the market economy for which someone can be realistically retrained, nor are there likely to be at any time in the future, then it is necessary to reorient the education and training to help the individual to

a. be able to contribute in the social economy
b. develop as a human being, helping them towards a pathway of flourishing – and a realisation of the good life

Will it be stopped?
There are authors such as Nicholas Carr[124] who have described the threat that they see being posed by the Internet and automation and who argue that they should be firmly constrained. The broad consensus however is that this trend towards more automation and deployment of Artificial Intelligence should be accepted – but that it should be actively regulated and controlled, with appropriate adaptive measures being taken to avoid problems.

6.3 How should society respond?

Clearly the preferred option is a future of good quality work for all. Every effort should be made to achieve this. If, however, one accepts the strong possibility of a radical reduction in paid employment in the market economy, how should society respond? Martin Ford[125] speaks of a New Economic Paradigm whilst Callum Chace[126] describes the resulting turmoil as an Economic Singularity. The strict definition of a Singularity is that one cannot see beyond it. Whilst such a definition could be valid in relation to the Technological Singularity, it is certainly less so with the Economic Singularity, which is nearer term and which urgently needs to be analysed and understood.

Chace[127] identifies six challenges that will have to be faced when seeking to make a smooth transition to a desirable position for humanity post the Economic Singularity. These six challenges are Meaning, Economic Contraction, Income, Allocation, Cohesion and Panic. Each of these challenges is discussed below.

Meaning

This is a huge subject. What is the meaning of life? How can one gain a sense of purpose? A multitude of writers have addressed these questions over millenia and this section just scratches the surface.

It is interesting to consider how different civilisations viewed work in the past. In Sparta the elite did not work – but they were trained to be warriors. In Thebes in Egypt, they were the rulers. In Greek, the word for 'work" is "ascholia" meaning "absence of leisure", leisure being "scholia". It is said that Zeus punished mankind with work – whilst in the Bible, Adam and Eve were punished with labour. Gregory Clark[128] describes studies of hunter-gatherers indicating that they have / had more leisure time than men today.

Chace[129] rightly argues that we make deep emotional investments in ideas and institutions like family, friendships, work, loyalty to tribes, nations and causes. In humanity's early days, time was primarily spent hunting and gathering. Later, most people were farmers and more recently, as the people have become more specialised, people have specific jobs – and those jobs form a major part of their identity. For many people, their job is their prime source of identity. How will people adapt if they are deprived of such a role?

Susskind[130] describes Marie Jahoda's chilling study of how the community of Marienthal, outside Vienna responded to the collapse of the principal employer in the town in the 1930's. The study revealed growing apathy, a loss of direction in life, and increasing ill-will to others. People borrowed half as many library books, they dropped out of political parties and stopped turning up to cultural events: in a few years, membership of the athletics club and glee club more than halved. Unemployment benefits required that claimants do no informal work: over two years, Marienthal saw a threefold increase in anonymous denunciations of others for breaking that rule – yet almost no change in the number of complaints that were judged well-founded.

Chace[131] argues that the challenge of meaning is not a major issue because of the examples of aristocrats and the (reasonably affluent) retired. The position of the gentry and the aristocracy was indeed something that people aspired to – and they made a point of not being employed as such. But they were at the top of their

society and were free to concern themselves with politics, science and the arts and as such play a leading role in society. This is quite different from someone who has started to play a significant role in society through the performance of their work and who is then deprived of that role through no fault of their own, with no hope of employment in the future. They are not in this position of leisure by choice but by compulsion, with a much more limited set of options.

Similarly, the example of the (reasonably affluent) retired person would seem to be quite different. Chace[132] talks of the U-shaped pattern of happiness through life, with the retired state being typically one of greater happiness. But this is typically from the view point of someone who is able to look back with satisfaction on their achievements in life, whether in terms of family, children, work, friendships etc. That again is quite different from the position of someone in their prime, or just starting out as an adult, being deprived potentially for ever of employment.

One is led back to the Greeks and Socrates' statement that "an unexamined life is not worth living". For some, religion will provide a context and purpose for life whilst others will need to find meaning in a Humanist context, both reflecting the faith instinct described by Nicholas Wade[133]. Positive Psychology has reinvigorated the study of the upside of life since the turn of the millenium and writers such as Martin Seligman[134] and Mihaly Csikszentmihalyi[135], with their concepts of flow and flourishing, offer a positive way forward for humanity. The challenge will grow with the removal of employment as a source of satisfaction, but humanity is starting to re-engage with what provides real life satisfaction and what life is really about.

Post a Technical Singularity, if all decisions of any significance are taken by machines rather than humans, then the challenge of finding meaning for a human being will be dramatically harder. This is an argument for avoiding such an outcome (through Differential Technological Development and a Nanny AI if necessary) until a clear, safe and positive way forward is assured (See Section 8).

Economic contraction

The problems of economic contraction are articulated by Ford[136] and Chace[137] and rest upon the fact that in economics as we know it, the supply and demand for goods need to equate (apart from the very short term). The concern is that if the workers are steadily made redundant as they are replaced by machines, their ability to afford to purchase the goods that the machines produce will be severely reduced. One can anticipate a reduction in both volume and price. Once a deflationary spiral has been triggered, it is difficult to pull out of it. This could be exacerbated if there was a marked global decline in fertility below 2, as indicated by the Vollset et al[138] study.

Whilst this is a very real issue that needs to be addressed, solutions can be found, such as through taxation and the redistribution of income via one of the mechanisms outlined below. A more dystopian view is that the democracy / human rights paradigm could collapse as a result of a political coup by the elite, leading to a formal hijack of most economic resources. In this scenario, economic contraction is avoided, but at a major social cost.

Both Ford[139] and Chace[140] explore one eventual outcome where, in a world of abundance, the price falls to zero – a so-called Startrek economy. Such a world has its attractions but there are some key issues that would need to be resolved such as allocation (see below).

Whilst in the past Governments have been shy to engage in this area, in the future, it is likely that Governments will be increasingly needed to help humanity to flourish.

Income
This is the challenge that has attracted the most attention. A Universal Basic Income is the best known of the potential solutions but there are several versions, including

Unconditional systems

a. Universal Basic Income (UBI) – a fixed sum paid to all unconditionally (Anne Lowrey[141])
b. Guaranteed Minimum Income (GMI) – favoured by Martin Ford[142], ensuring that no-one's income is less than an agreed basic level, but avoiding paying money to people who already have a decent income
c. Negative Income Tax (NIT) – a variant of GMI favoured by Milton Friedman[143]

Conditional systems

a. Conditional Basic Income (CBI) – separate variants have been proposed by Jeremy Rifkin[144] and Daniel Susskind[145]: full payment of the CBI is conditional on a certain amount of time being spent working as determined by the community.
b. A Global Marshall Plan – a more radical concept, developed by Pieter Kooistra[146].

Kearney and Mogstadt[147] of the Aspen Institute provide a good overview.

The unconditional systems vary only in the manner in which they are calculated. The straight UBI is given to all without means-testing. This has the merit of transparent equity and is simple to administer. Other systems seek to avoid paying money to those who do not need it, through a Guaranteed Minimum Income Scheme, developed via the tax system or in some other way. It leaves people free to choose whether to work in the market economy, the social economy or not at all. Some proponents favour a relatively low level of UBI so as to ensure that there is a sufficient incentive to work. Whilst such considerations are relevant today, they may become far less so in the decades to come. These GMI schemes risk introducing a major disincentive to work (for money) which is a key factor in any such design.

An alternative approach to this work incentive / encouragement issue is to put some conditions upon the application of the scheme. Rifkin[148] addresses two groups, namely

a. those who are in work but have leisure time, which could be spent in part working for instance in the community,
b. those who are not in work.

For the former Rifkin[149] suggests the payment of Shadow Wages, which can be tax deductible just as the payment of cash donations for charities can be tax deductible. For the latter group, he suggests a Social Wage for those who are willing to work in the third sector.

A world with less work risks being a deeply divided one, which is why many support the encouragement of work in the third sector in the interests of the community. How Susskind[150] asks, do you avoid feelings of shame and resentment on either side? He proposes that the "distribution problem" is solved through the payment of a Conditional Basic Income, whilst the contribution problem is resolved by the making of a non-economic contribution, if an economic one is not possible. It would fall to individual societies to determine what such non-economic contributions might look like.

Whilst the conditionality described above relates to the contribution of labour to the third sector, the conditionality of the Kooistra's[151] Global Marshall Plan relates to conditions relating to what the Basic Income is spent on. The scheme uses a sociocratic system to determine what types of expenditure would NOT be acceptable, with the overall aim of gradually shifting the way the world's population spends its money towards a more peaceful, sustainable goal as the volume of funds flowing through the GMP increases its proportion of global economic activity.

In addition to these Government related solutions, O'Keefe et al[152] from the Future of Humanity Institute describes a Windfall Clause approach. Whilst not in any way dismissing Government managed solutions, O'Keefe et al[153] sets out how businesses could voluntarily sign up to a Windfall Clause if their income exceeds a substantial proportion (>1%) of the world's total economic output.

On top of its basic design, there are two key factors to consider regarding a UBI, namely when to introduce the scheme and at what level. To some degree they are linked in that the longer the introduction is delayed, the higher the global GNP should be and therefore the higher the level of payment could be. Whilst such a system could be introduced now, the level of payment would be relatively low. There would seem to be a strong argument for waiting until firstly the debate about the choice of mechanism is resolved and secondly the global GNP has reached a point where the payment can be made at a good level. What that level is will be the subject of much debate also. The sooner the scheme is introduced, the more it would be important to ensure that there remains sufficient incentive to perform the work that does need to be performed. The later the scheme is introduced, the more the emphasis is likely to be on ensuring that the payment is sufficient to enable the recipient to live the good life and to flourish.

To sum up, there are both multiple possible solutions and multiple factors to consider. Much more thought and debate nationally and internationally is required.

Allocation

A critical issue is that of allocation. Imagine that we manage to cross Peter Diamandis'[154] bridge to an economy of abundance where everyone is living a comfortable and fulfilling life and is able to take advantage of the almost-free goods and services which are provided by the "machines of loving grace". Any

extra costs that we would have to meet are funded by non-punitive taxes on the income and assets of those still working.

The problem comes with the allocation of scarce resources. One may assume that the limited number of paintings by Van Gogh,1930's Bentleys, apartments on Fifth Avenue or beach-front houses will all belong to the working / plutocrat rich. But for those who are not the working /plutocrat rich, will all their houses be the same? Who will decide on what the cut-off point is between a house that people can carry on living in and a house that is too nice to be a normal person's property?

Cohesion

The level of inequality has been increasing due to automation. Joseph Stiglitz[155] argues powerfully how capitalism today benefits the top 1% to the detriment of the bottom 99%. And this is just at the beginning of the impact of AI on employment. As the majority of jobs are lost and not replaced, the level of inequality is likely to rise and lead to what Harari[156] describes as "the Gods and the Useless".

How people respond to such a level of inequality is open to question. Violence is one undesirable income. An institutionalisation of such a divide, as portrayed in a Brave New World[157] (first published in 1932) is considered by most readers to be a nightmare. Yet some authors such as Chace[158] have come to consider, in comparison with other scenarios, that a Brave New World might not be so bad.

Ensuring a reasonable level of economic equality and the potential for people to live fulfilling lives are key goals in order to avoid a serious loss of cohesion.

Panic

Society currently does not have a plan for how to cope with the economic singularity. There is no consensus regarding what type of economy could cope with more than half the population being permanently unemployed nor how to manage the transition from here to there. In the absence of such a plan, for Chace[159] it is clear: when large numbers of people realise that their livelihoods are in jeopardy, they will panic.

Chace[160] sees self-driving vehicles as the canary in the coalmine, making it impossible to ignore the impact on employment of cognitive automation. He considers that the panic will occur within a few years, perhaps a few months, of self-driving vehicles leading to people being laid off.

6.4 Implications for regulation and governance

What is clear from the above analysis is that the impact of AI will be most profound in almost all aspects of life and certainly on the economic and social human experience.

Timing and phasing

The challenges outlined above will not all arrive at the same moment. For instance, the issues of both scale and conditionality of a Basic Income concept are closely linked to the phase in which the Basic Income is being considered. In the near future, with significant amounts of paid employment available, it is important not to discourage people from seeking work (the danger of 100% marginal tax rate). When AI has had a major disruptive impact on the labour markets resulting in very little work available in the market economy, the need to

incentivise the individual to find work in the market economy is low, whilst the need to encourage active engagement in the third sector is high.

It is therefore important to envisage and plan a phased and adaptive introduction of policy adjustments. Global citizens' assemblies could aid this process.

Global interaction
As indicated in the UBI table above, there is a significant global interaction between the economic and social policies introduced to address AI related challenges. For instance, the establishment of a reasonably generous UBI in one country could put significant pressure on migration, particularly if it was sharing the benefits of a country rich in AI related wealth, wealth generated at the expense of less well AI endowed nations. It would therefore be important to both seek to share the wealth globally and to coordinate policy introduction around the world.

Economic institutions involved
Economic issues are the responsibility of Nation States. But there are a number of global institutions such as the International Labour Organisation (ILO), the Organisation for Economic Cooperation and Development (OECD).
the World Trade Organisation (WTO), the World Bank, the International Monetary Fund (IMF) that provide coordination and guidance.

As AI continues to be developed and deployed around the world, these institutions need to engage in stimulating a global debate on appropriate solutions to the issues raised above, solutions that will not only work in the short term in a particular country, but will contribute to a flourishing of humanity across the world. If the ILO is to take the lead on this then it will really need to rise to the challenge.

Recommendations
The crucial actions to address these dramatic trends are to

a. Agree on the need for the international community to reach a common understanding of the issues and the factors that can contribute to a positive way forward both economically and socially.
b. Commit to ensuring that the planning required to address the economic and social consequences of the deployment of AI takes place hand in hand with the development of AI.
c. Ensure that there is a constructive global response to this global problem, that can be announced and implemented by 2025
d. Develop a phased approach, including ensuring that in the longer term there are contingency plans for a further transition as wealth increases, leading to a world of abundance
e. Agree which international body should take on the role of leadership and coordination of the above.

7 DESKILLING

7.1 Skill and expertise

In the past skills, the learned power to do something competently, have been lost because either the society or empire collapsed, or because the skill was no longer required. Usually a few people retained the skill, but it became very rare. Now, as

machines become more competent, and as we hand over more and more tasks to machines, the danger of skill loss is becoming more serious.

When driverless cars are readily available, should we be concerned about losing our driving skills? We would need to retain the skills if there is any question of defaulting to the driver – and whilst there is any possibility of driving a car oneself, e.g. a vintage car requiring a driver. But when that is no longer the case?

Airline pilots are very aware of the dangers of deskilling. Automatic pilots are capable of flying a plane from take-off to landing. The most difficult sectors of a flight are taking off and landing. Pilots do not use the automatic pilot to help with the difficult bits however, but they use it for the easy bits (the cruise). They are fully aware of their vulnerability if they are not as proficient as they can be at taking off and landing. Will pilots always take this approach?

A seed bank has been established deep inside a mountain on Svalbard, where samples of seeds of all plants are stored in case of some global catastrophe or the loss of a region's seeds. Should a skills archive be established that would enable humanity to recover in the event of some technological catastrophe?

Carr[161] argues that whether it is a pilot on a flight deck, a doctor in an examination room, or an Inuit hunter on an ice floe, "knowing demands doing". He suggests that one of the most remarkable things about human beings is also one of the easiest to overlook, namely that each time we "collide with the real, we deepen our understanding of the world and become more fully a part of it". While we're wrestling with a difficult task, we may be motivated by an anticipation of the ends of our labour, but it's the work itself—the means—that makes us who we are. He argues that computer automation severs the ends from the means. It makes getting what we want easier, but it distances us from the work of knowing. As we transform ourselves into creatures of the screen, we face an existential question: Does our essence still lie in what we know and what we can do, or are we now content to be defined by what we want?

7.2 Moral deskilling

Perhaps the loss of a skill that we did not have in the distant past and may not need in the future is no great loss. But Vallor[162] argues that there is a more fundamental skill at risk, namely our moral judgement. He argues that as more and more moral judgements are being made by machines (AI), there is a serious risk that our moral judgement will wane through lack of experience and practice; he goes on to cite examples in the fields of automated weapons technology, new media practices, and social robotics.

Brian Patrick Green[163] builds on Vallor's[164] thesis. He argues that we already modify human behaviour through law, government and culture. We appoint special people including parliamentarians, judges and police to make and enforce laws so as to promote or suppress certain behaviours. As a result, we do not have to think as much about moral choices as we would in a less structured society. AI would extend this trend to cover even more aspects of life with even more control.

Green[165] sees AI as lowering our moral capacities by

a. Attacking Truth and Attention: a result of the Infocalypse, and through AI-powered games draining our attention away from the important things in life such as caring relationships and thinking about solving larger-scale problems
b. Preventing human maturation and moral development: with technology tending to infantilise humans, and with AI-driven manipulations of our psychology tending to crowd our spending time attending to relationships, caring about others and thinking about ethical problems
c. Normal and weaponised complexity: as normal life becomes more and more complex, understanding will no longer be an expectation, and in the midst of this lack of understanding, bad things can happen; as understanding is no longer an expectation, humans become much easier to deceive and manipulate.

In response to these threats, Green[166] identifies six possible solutions, including what he describes as

a. Education: using AI for personalising and enhancing education (including through Virtual Reality); teaching and rewarding practical wisdom and moral leadership, not just knowledge; working harder to imbue good moral habits and teaching moral attention; ensuring that automated systems do not give equal weight to falsehoods and the truth; using AIs to help protect the information ecosystem and help humans become more discerning in their assessment of facts.
b. Attention: using AIs to help train our moral attention by filtering out distractions and highlighting ethical issues;
c. Becoming adults: strongly resist technology that seeks to act like our parents or infantilize us; AI could help us develop moral maturity and discernment, helping train us with virtues such as restraint, practical wisdom and courage.
d. Interacting with other humans: AI could encourage us to spend more time with others face to face, thereby building stronger interpersonal relationships: it is in interaction with others that most moral life happens.
e. Dealing with complexity: as with airline pilots, if AI could help solve the easier problems in life, could humanity concentrate on solving the biggest ethical problems such as world peace, hunger, healthcare etc.? This would be counter to the mission of Demis Hassabis (Deep Mind) to “first solve intelligence, and then use that to solve everything else”[167].
f. Stopping weaponised complexity: AI can help us expose when bad actors are using complexity as a weapon to deceive and manipulate us (though currently weaponised complexity would seem to have the upper hand).

It is also important to engage with key moral dilemmas right now. An example that is the application of the well-known ‘trolley problem’ in relation to autonomous cars. Fully autonomous self-drive cars will need to be programmed to deal with a myriad of moral issues such as whether the vehicle should prioritise the life of a pedestrian over a passenger, or give equal weight to all lives, or be influenced by the age of the individual concerned. Car manufacturers have tended to play down this issue with arguments reminiscent of sub-prime mortgage advocates arguing that by collateralising the debt they became risk free. Should not society be

involved in those moral decisions? These moral decisions are being enshrined in AIs around the world, they are real issues and humanity should engage with them.

Another conscious initiative to support the engagement of the wider society with ethical issues would be the development of Apps to help avoid moral deskilling – both for the near term and the longer term, when the Apps could help to keep people in touch with the moral decisions being made on their behalf and to continually keep them under review.

7.3 Risk of an unacceptable end-game

The risks outlined above are near-term and serious – but they feel as though they could be manageable with sufficient awareness and determination. As the time horizon moves towards more highly intelligent and perhaps superintelligent machines the risks become much greater. They can evoke the children's film WALL-E[168], set in a world where humans become passengers in a cruise ship run by machines, on a cruise that goes on for ever (so long as the worst dangers of the goal-alignment problem can be overcome). They think that they are in charge, but reality is quite the opposite.

Russell[169] makes the point that there will be a serious risk of a tragedy of the commons and a slide towards enfeeblement. For any individual human, it may seem pointless to engage in years of arduous learning to acquire knowledge and skills that machines already have. But if everyone thinks that way, the human race will collectively lose its autonomy. This risk may not materialise: some consider that the youth of today are robust and will resist such a risk. But it would be foolhardy to ignore it.

Russell[170] suggests that the solution to this problem may be cultural rather than technical. He envisages the need for a cultural movement to reshape our ideals and preferences towards autonomy, agency, and ability, and away from self-indulgence and dependency – evoking a 21st century version of Sparta's unique ethos. Russell envisages "preference engineering on a global scale" along with radical changes to the way society works. He suggests that we might even need the help of superintelligent machines, both in shaping the solution and in the actual process of achieving a balance for each individual.

Russell[171] makes the analogy with a parent and child: once the child is beyond the helpless stage, parenting requires an ever-evolving balance between doing everything for the child and leaving the child entirely to his or her devices. In this analogy, humanity is in the role of the child, and machines in the role of the parent.

There are huge issues at stake here and such issues will need to be fully considered and addressed both before and during the Long Reflection (see Section 8.4). What is equally clear is that WALL-E scenarios are totally unacceptable, even if they are a caricature.

Something must be done. AI, and indeed AGI, should have as one of its main goals the improvement in the resilience (including physical and mental health) of human beings: this helps to reduce the risk of catastrophic developments. This end-game concern however is verging on an existential threat, quite distinct from the existential threat addressed in Section 8.

One possible temporary measure would be the pressing of the Pause Button discussed in Section 8, leading to a pause in the development of AIs of still greater intelligence. But as discussed later, such a Pause Button will be extremely difficult to establish, if it is to be effective.

7.4 Recommendations

It is proposed that

a. Due thought should be given to the impact of increased deployment of AI on human skillsets (including moral skillsets) and autonomy, both in the short, medium and long term.
b. Care needs to be taken to avoid AIs replacing skills that are fundamental to human identity, functioning and flourishing, and the exercise of which may be considered essential for human dignity.
c. A future pause in the further development of AI might be desirable, to enable us to find a suitable way of achieving these proposals (cf Recommendation 8.8 d.)

8 THE FUTURE OF HUMANITY

8.1 Existential risk and risk perception

The concept of existential risk, the end of the human race, has been in the realm of mystical belief, but of little scientific relevance until recently. We have become aware of the five periods of mass extinction, but have had no ability to influence the causes in the past, such a massive asteroid hitting earth.

In the last 80 years however, things have started to change. For the first time, new existential risks are developing as a result of humanity's own actions. In contrast to our complete inability to control an asteroid in the past, these new risks are risks that we can control – so long as we take them seriously.

One problem is that our ability to comprehend the significance of these risks is very poor. This is in part because they are new, and we have not had the cognitive need to get our minds around such things before. Nor have we integrated it into our civic and moral traditions. Ord[172] (2020), drawing on Wiener[173] (2016), Kahneman[174] (2011) and Weber[175] (2006), identifies some of the key reasons why our ability to estimate the significance of catastrophic and existential risks is so poor.

Economic reasons

Our economic systems are extremely poor at dealing with existential risk. Key reasons include:

a. It is a public good – so the market is not interested.
b. It is a Global public good – so even nations will neglect it.
c. The same effect that causes an undersupply of protection causes an oversupply of risk.
d. Even worse – it is intergenerational

Political reasons

Similarly, our political systems are poorly equipped to deal with such issues:

a. Short termism exacerbated by the electoral cycle

b. No key constituency to take ownership of the issue
c. The sheer gravity of the issue – above my paygrade

Behavioural psychology reasons
Behavioural psychology offers key explanations for why people tend to both underestimate the probability of such an event and underestimate the significance of the event:[176]

a. Underestimating probability: The availability heuristic reflects the fact that people tend to be influenced in their assessment of probability by the "availability of the event", namely the closeness in time and/or place to such an event: the closer such an experience the more likely the person is to attribute a high probability. The fact that by definition people have no experience of an existential risk means that they are likely to downplay such an event.
b. Understanding scale: Scope neglect (or mass / psychic numbing) describes the phenomenon that someone will be prepared to make a payment in relation to a tragic individual case. If the number of say children involved in the tragedy is more than one, the payment may go up – but if it goes up to say 100, the absolute payment (let alone the relative payment) goes down. Imagine how low it would be for billions of current humans or trillions of future generations.

Perception of the goal alignment risk
In relation to AI, a potential existential risk has been identified in relation to the ability of human beings to ensure with complete certainty that the goals of a superintelligent AI would not conflict with the goals of humanity. The nature of this concern and how it is perceived is set out in Section 8.2 below. What AI developers are doing about it is described in Section 8.3, whilst what else should be done is set out in Sections 8.4 to 8.7 and a summary of recommendations shown in Section 8.8.

This goal-alignment existential risk is assessed by Ord[177] to be 10% within the next 100 years, a truly horrendous figure and 1.5 times the combined existential risk from all other causes, namely engineered pandemics, unforeseen anthropogenic risks, other anthropogenic risks and all natural risks.

8.2 The control / goal-alignment risk

The basic problem of control / goal alignment reflects a theme that has been the basis of fairy-tales and myths down the ages. King Midas wanted lots of gold and was granted his wish that everything he touched would turn into gold. He had not articulated his goal accurately however and soon turned his beloved daughter into gold. Similarly, the Sorcerer's Apprentice created huge problems when the assistance he sought got out of hand.

In the current AI related literature, there are several well-known examples of goal misalignment, including Nick Bostrom's[178] paperclip (when the whole universe gets gradually transformed into paperclips), or Marvin Minsky's[179] calculation of Pi, or Stuart Armstrong's[180] eliminating the human race in order to cure cancer. The apocryphal genie stories typically involve three wishes, the last one being to undo the first two. The point is that there are lots and lots of trade-offs that need to be made which are often assumed – but potentially fatally.

The scale of the problem for AI grew larger when Steve Omohundro[181] established that a highly intelligent AI would have instrumental goals, such as remaining switched on and obtaining more and more resources in order to achieve its goal. A superintelligent AI is expected to be extremely good at defending itself. This was reflected in the film "2001 A Space Odyssey": HAL had his goals and disobeyed human instructions as he considered that they conflicted with his goals.

The views of experts

The basic problem of control (or goal-alignment - either term may be used) was perceived long ago. Alan Turing for instance, the father of computing, argued that at some point computers would probably exceed the intellectual capacity of their inventors, and that "therefore we should have to expect the machines to take control."[182] He did not speculate further.

Norbert Wiener[183], the father of cybernetics, argued that it would be difficult to manage powerful computers, or even to accurately predict their behaviour. "Complete subservience and complete intelligence do not go together," he said. Envisioning Sorcerer's Apprentice scenarios, he predicted, "The future will be an ever more demanding struggle against the limitations of our intelligence, not a comfortable hammock in which we can lie down to be waited upon by our robot slaves."

I. J. Good[184], Alan Turing's chief statistician at Bletchley Park, was the first person to articulate an "intelligence explosion". "An ultra-intelligent machine could design even better machines," he wrote. "There would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultra-intelligent machine is the *last* invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously." Good was Kubrick's supercomputer advisor on 2001- A Space Odyssey.

Eliezer Yudkowsky[185] was one of the first to explore the goal-alignment problem in some depth around the turn of the century. Nick Bostrom[186] references Yudkowsky's work extensively in his own study "Superintelligence", which formalises the thinking in a logical and structured manner and a more academic style. Superintelligence lays out the problem and identifies the steps needed to resolve it and the challenges to be met at each step.

Stuart Russell[187], co-author of the standard textbook on AI across the world, had, by 2013, come to see the issue of the goal alignment / control problem as possibly the most important question facing humanity.

Geoffrey Hinton[188], a leading AI researcher based at the University of Toronto and described as the "Godfather of Deep Learning", avoids speculation, considering that anything beyond five years in this field is impossible to predict given the exponential rate of development. He was however overheard talking to Nick Bostrom in 2015 at the Royal Society in London where he had given the keynote speech[189]. Bostrom asked him whether he was in the group of people who considered that managing existential risk from artificial intelligence was probably hopeless or easy; he replied that he was in the "hopeless" camp. According to the

same report, Hinton does not categorically rule out human beings controlling an artificial superintelligence, but warns that "there is not a good track record of less intelligent things controlling things of greater intelligence". Asked by Nick Bostrom why he continued his research despite his grave concerns, Hinton stated, "I could give you the usual arguments. But the truth is that the prospect of discovery is too *sweet*."

There are many other such quotes from Elon Musk, Lord Martin Rees, Stephen Hawking and so on. Some authors however have characterised the concern as hype and dismissed it as "transhumanist"[190] (Coeckelbergh 2020). The fact that very high profile and relevant people have expressed concern is somehow portrayed as evidence that the problem is overdramatised. The fact that Nick Bostrom considers the future beyond the development of intelligent machines is presented as evidence that his views are tainted. Such responses seem biased.

At this stage, nobody is saying that there WILL be a goal-alignment catastrophe. But the fact is that many of those experts with the most relevant expertise consider that a serious risk exists (and for many, the gravest existential threat that we face). They believe that there is grave danger that a superintelligence will be developed, with the capability to not only develop improvements in its own design but to create that new improved version of itself. Unless the goals of that superintelligence are completely aligned with those of humanity, then there will be a conflict of goals sooner or later and humanity will be the loser.[191] The ultimate demise of humanity would be simply a matter of time.

It behoves the rest of society to take these concerns extremely seriously.

8.3 Transforming AI to eliminate the risk

The AI field continues to be dominated by those interested in increasing the power of AI rather than making it safer. Many of the AI experts who are concerned about control however are committed to developing solutions to the control / goal-alignment problem (for this paper defined as Safe AI). There have been a number of contradictory "obvious solutions", evidence of the fact that to design a truly safe AI is extremely difficult.[192] Four such initiatives are set out below.

Beneficial AI

In 2015, Stuart Russell described[193] how the intent is to stop doing AI as we do now (improving the decision-making capability of systems) and instead to build a new AI that has, built into it, certain adjectives. Civil Engineering is not called Building Bridges that Don't Fall Down - of course everyone wants the bridges to stay up. This is implicit in the discipline of Civil Engineering. In the same manner, AI should act in ways that are well aligned with what human beings actually want: we should not need to stress that, but today we do need to since it is not the case.

Russell[194] proposes the development of "Provably beneficial AI". The term is somewhat oxymoronic since "beneficial" is a loosely defined term, so how could it be proved to be beneficial? But that is a reflection of the challenge that AI researchers face.

Russell is seeking to apply Inverse Reinforcement Learning (IRL) to the problem of Value alignment. With Reinforcement Learning, you are given rewards in

response to the actions that you take (the process used by Deepmind). With Inverse RL you observe some behaviour and seek to work out what value that behaviour is seeking to maximise. There is now a quite extensive literature on Inverse Reinforcement Learning. Russell's version is Cooperative IRL.

A key aspect is the introduction of uncertainty in the logic of the AI. Where a superintelligence using current technology could be dangerous is if it is totally focused on its objective and there turns out to be collateral damage along the path to the achievement of this objective: the collateral damage that could be of a scale to eliminate the human race. With Russell's[195] concept however, the Beneficial AI is never completely certain that it has interpreted human preferences correctly. Critically, it would therefore be prepared to allow itself to be switched off – because it accepts that humans could know better.

Such an approach is of course fraught with difficulties. Humans are irrational, inconsistent and weak-willed. Values differ across individuals and culture. Nevertheless, the concept appears to have real potential.

Friendly AI
Yudkowsky[196] launched Friendly AI 1.0 in 2001 with a detailed plan of what needed to be done. Since then, he and his Singularity Institute (later the Machine Intelligence Research Institute – MIRI) have pursued 1.0 and subsequent versions of the Friendly AI Plan.

In line with Russell's[197] 2015 talk, MIRI's work is currently aimed at:

> "helping jump-start a paradigm of AI research that is conscious of the field's long-term impact. [Their] methodology is to break down the alignment problem into simpler and more precisely stated subproblems, develop basic mathematical theory for understanding these problems, and then make use of their newfound understanding in engineering applications."[198]

Safe AI Scaffolding Strategy
Steve Omohundro[199] is developing the "Safe-AI Scaffolding" strategy for building powerful new technology in a way that contributes to the greater human good. The idea is to start with provably safe intentionally limited systems and then to use those to build more powerful systems.

Singularitynet
Similarly, Ben Goertzel[200] is working on a decentralised form of AI which is intended to democratise access to AI. He claims that the team are pursuing the Safe AI goal.

8.4 Remaining existential risk

"*You can't control research"*
Russell's[201] answer to this assertion that you cannot control research is as follows:

a. The Asilomar Conference on Recombinant DNA (1975) agreed to self-imposed restrictions on recombinant DNA experiments
b. Industry adherence has since been reinforced by a US FDA ban on human germline modification.
c. There is a pervasive culture of risk analysis and awareness of societal consequences (but he acknowledges that there is a minority garage

subculture that argues for the total freedom to carry out whatever research the researcher might wish to perform).

Whilst these measures have been largely successful to date, on 25 November 2018, two days before the Second International Summit on Human Genome Editing in Hong Kong, Jian-kui He, a Chinese researcher of the Southern University of Science and Technology, released a video on YouTube announcing that he and his colleagues had "created" the world's first genetically altered babies, Lulu and Nana[202]. This incident demonstrates that whilst the establishment of a culture in support of a certain safety related policy can help enormously to reduce the incidence of breaching that policy, it simply cannot guarantee that the policy will be adhered to 100.000 %.

It is therefore necessary to assume that additional powerful measures will need to be taken, over and beyond the achievement of "a different way of doing AI". The fact that the AI research community is starting to prioritise the safety of AI is encouraging, but it is not sufficient. The analogy with the avoidance of human germline modification is relevant – but as has been demonstrated by Jian-Kui He in China, it does not guarantee anything. With AI, one renegade research team that pursues unconstrained superintelligence could be enough to spell disaster.

Beneficial AI and Friendly AI would therefore seem to be excellent routes to explore, but they should NOT be seen as sufficient. They seek to provide a means of developing AI that can work with humanity, but they do not of themselves provide the means to guarantee that no illicit development will take place using dangerous AI. This requires additional measures, which still may not offer an eternal guarantee, but would offer humanity a great deal more security than simple reliance on a positive culture.

A long-term view of our predicament
Toby Ord[203], taking a very long-term view, sees humanity's (and indeed life's) existence on earth in three phases, namely

a. From origin to reaching existential security
b. Long Reflection
c. Realising full potential

The first, he sees as consisting of two components, namely:

a. *Preserving* humanity's potential, extracting ourselves from immediate danger so we don't fail before we've got our house in order. This includes direct work on the most pressing existential risks and risk factors, as well as near term changes to our norms and institutions.
b. *Protecting* humanity's potential means establishing lasting safeguards that will defend humanity from dangers over the long-term future, so that it becomes almost impossible to fail.

In simple terms, *Preserving* is about fighting the latest fire, whereas *Protecting* is about making changes to ensure that fire will never again pose a serious threat. Ord[204] argues persuasively that humanity's current prime task is to bring this century's risk down to a very low level and then to keep gradually reducing it from then on. This vision needs to be more widely spread.

Ord[205] argues that, once preserving has been completed and protection is in place, there appear to be no major obstacles to humanity lasting an extremely long time, if only that were a key global priority, which at present it is not. There is a need to devote at least as much of humanity's brilliance to forethought and governance as to technological development. There will be great challenges in getting people to look far enough ahead and to see beyond the parochial conflicts of the day. But the logic is clear, the moral arguments powerful: it can be done.

8.5 Will humanity be ready?

A key question that needs to be addressed is whether humanity will be ready and willing to enter the superintelligent machine age at the point when it might be technically feasible to do so. Humanity needs to have the time to have the debate and to prepare. Until recently such an era was the exclusive province of science fiction writers and their readers. But with the rapid advances being made in the field of AI, this is something that everyone needs to come to terms with. It is a realm that needs to be addressed in schools and universities and it needs to be a part of the public discourse.

Do we know what we want?

Max Tegmark[206], in his book "Life 3.0", included a questionnaire with a series of options for the future. A non-statistical sample of responses clearly indicated that the large majority of the population do not appear to want any option other than the status quo – with human-beings firmly in control. It is unthinkable that such a transformative step as moving beyond a world dominated by humans should occur without any conscious decision. If the world is to enter the superintelligent machine age, it should only be because people have analysed the options, satisfied themselves that all necessary precautions have been taken and chosen to proceed. Between now and then, other major changes will have occurred. There will be major advances, both biological and digital, in the "enhancement" of the human. Projects such as Elon Musk's Neuralink are already making progress down a carefully planned path towards a Whole Brain Interface[207].

A Long Reflection

The second major phase in Ord's[208] Grand Strategy for Humanity is The Long Reflection. The ultimate aim of this Long Reflection is to achieve a final answer to the question of which is the best kind of future for humanity. It would not be necessary to fully complete this process before moving forward. It would however be essential to be sufficiently confident in the broad shape of what we are aiming at before taking each bold and potentially irreversible action – each action that could plausibly lock in substantial aspects of our future trajectory

For example, it may be that the best achievable future involves physically developing humanity by genetically improving our biology (replacing humanity). Or it may involve giving people the freedom to adopt a stunning diversity of new biological forms (fragmenting humanity). Proceeding down either of these paths prematurely would be in the context of "Protecting humanity's potential". And these are the kinds of decision that would need to be made after the Long Reflection.

We need to take our time, and choose our path with great care. Once we have existential security, Ord argues, we are almost assured success if we take things slowly and carefully: the game is ours to lose; there are only unforced errors.[209]

This is not to say that this is the sole task of humanity during this period. There would be other great projects, such as the continuing quests for knowledge, prosperity and justice.

The ultimate aim is not just to win the goodwill of those alive at the time, but to deliver a verdict that stands the test of eternity.

These first two steps of existential security and the Long Reflection could be seen as designing a constitution for humanity. How long such a period of reflection should take is a very open issue. It might seem that it would be difficult to maintain a debate for more than fifty years. There will be people pushing to move to the next stage and they will be very keen to reach a conclusion. But Ord's[210] point is that due to the significance of the decision, and the security of the current position, it would be of prime importance to get the decision right: to have considered all options very carefully, reached a true popular consensus and made all necessary preparations.

Ord[211] argues that it would be possible to begin the Long Reflection now: it would not hurt, but he rightly argues that it is not the most urgent task. To maximise humanity's chance of success, it is vital to first reach a position of safety: to achieve existential security. This, he suggests, is the task of our time. The rest can wait.

8.6 The need to have a Pause capability – a Plan B

The step to a superintelligent machine world is considered by some as the most important step humanity will ever take. Such a decision should only be taken when humanity is ready. Today humanity is not ready and shows little sign of being so for a long time to come. It is entirely possible, and indeed arguably most probable, that humanity will NOT be ready for such a transition, at a time when AI developers are pushing for it. It would seem that at that point humanity would need a Pause button – the ability to put the whole transition process on hold until the time is reached when society is agreed that it is both desirable and technologically safe to proceed.

A Pause Button and an AI Nanny

This argument for having a Pause capability as a last resort was first articulated by Ben Goertzel[212]. There may be a number of different ways in which the Pause capability can be achieved but Goertzel's AI Nanny proposal is the best candidate so far. It is based on the premise that it will be possible to develop an AI that is more intelligent than humans but not so intelligent that it is impossible to control. His concept is of an advanced AGI software program with:

a. General intelligence somewhat above the human level, but not too dramatically so such that it cannot be controlled by humans
b. Interconnection to powerful worldwide surveillance systems, online and in the physical world
c. Control of a massive contingent of robots (e.g. service robots, teacher robots, etc.) and connectivity to the world's home and building automation

systems, robot factories, self-driving cars, and so on and so forth

d. A cognitive architecture featuring an explicit set of goals, and an action selection system that causes it to choose those actions that it rationally calculates will best help it achieve those goals
e. A set of pre-programmed goals including the following aspects:
    i. A strong inhibition against modifying its pre-programmed goals
    ii. A strong inhibition against rapidly modifying its general intelligence
    iii. A mandate to cede control of the world to a more intelligent AI within 200 years
    iv. A mandate to help abolish human disease, involuntary human death, and the practical scarcity of common humanly-useful resources like food, water, housing, computers, etc.
    v. A mandate to prevent the development of technologies that would threaten its ability to carry out its other goals
    vi. A strong inhibition against carrying out actions with a result that a strong majority of humans would oppose, if they knew about the action in advance
    vii. A mandate to be open-minded toward suggestions by intelligent, thoughtful humans about the possibility that it may be misinterpreting its initial, pre-programmed goals

Link between Beneficial AI and AI Nanny

If Beneficial AI (or Friendly AI or some similar technology) has proved successful by the time that the AI Nanny is required, then the AI Nanny could be created using this technology. If that were the case, the AI Nanny would be very low risk, and simply available to provide time to think, discuss (The Long Reflection) whilst various problems were addressed by the AIs.

If, however, a Safe AI had NOT been developed by the time an uncontrollable superintelligence is about to be developed, then the need to pause AI development (for the time being) would be far greater as the risk of a disaster post transition would itself be far greater.

Beneficial AI, AI Nanny and The Long Reflection

The inter-relationship between the different elements above is summarised in Figure 2 below.

a. Research and development of Beneficial / Friendly AI can be pursued in parallel with research and development of AGI, with the aim of ultimately integrating the two elements to form a Safe AGI.
b. Global governance will need to be developed to a point where, with appropriate AI support, there is absolute certainty that there is no research, development or deployment into the current goal-oriented form of Artificial Intelligence beyond a certain point. It is difficult to imagine that this can be assured and enforced without having established appropriate supranational governance.
c. Such governance would also be necessary to support an AI Nanny. This will be needed in order to ensure that the transition to superintelligence and/or the first irreversible step towards a new future for humanity is only made when society is ready for such a step. The AI Nanny would be a form of Beneficial/Friendly AI if that had already been established, but it could be

that an AI Nanny is needed before Beneficial / Friendly AI is available. If more progress is made on the development of Basic AGI and a Basic AGI superintelligence is imminent, then the AI Nanny would need to be fashioned from Basic AGI technology. In either event, significant global cooperation will be required to establish an AI Nanny able to play the part that is needed.

d. With the assistance of more advanced AI, it is to be hoped that the other current existential risks can also be addressed during this period – with the world both preserved and protected from their emergence. An AI Nanny could play a major role in this.
e. With the above measures in place, the time would be available for humanity to have a Long Refection. The ultimate aim of the Long Reflection would be to achieve a final answer to the question: which is the best kind of future for humanity. We would not need to fully complete this process before moving forward. What is essential is to be sufficiently confident in the broad shape of what we are aiming at before taking each bold and potentially irreversible action – each action that could plausibly lock in substantial aspects of our future trajectory.
f. Only when this Long Reflection has been completed, consensus reached, decisions made and preparations completed should humanity take its first step towards a bigger future.

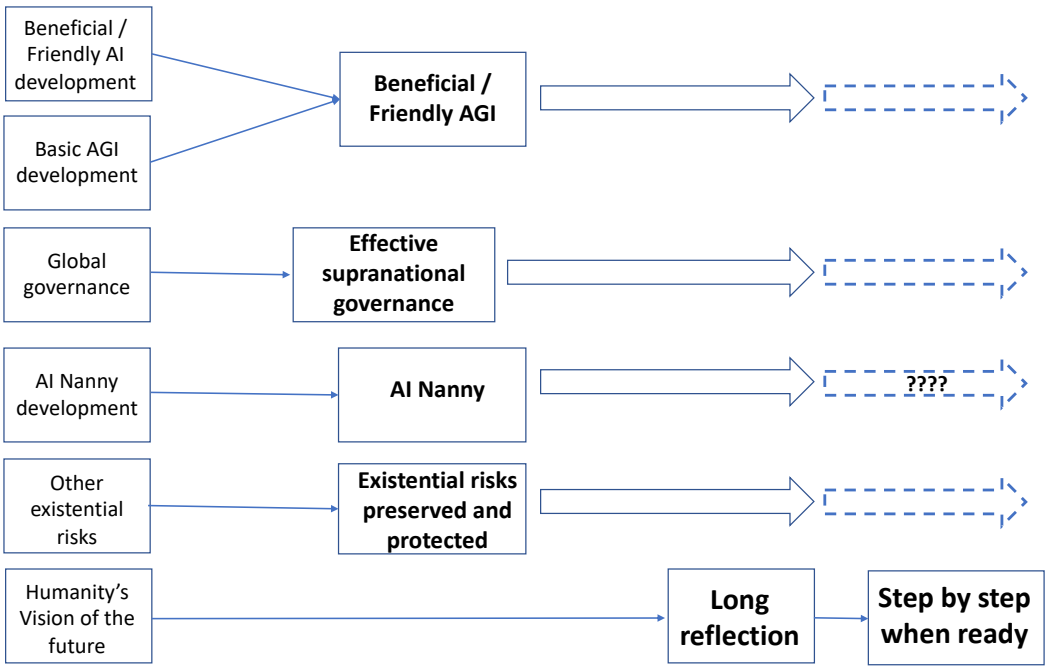

Figure 2: The Long View

8.7 Differential Technological Development

Until the last 80 years, the only risks of bringing about the extinction of the human species have been physical phenomena such as a sizable asteroid colliding with earth. With the development of the atom and then hydrogen bombs and the accumulation of vast arsenals of nuclear weapons, humanity has for the first time developed the ability to destroy itself for ever. This ability has been extended in

recent years to include several other human related existential risks, including climate change and other environmental risks, engineered pathogens and the AI goal-alignment problem.

A key method of addressing these existential risks, and the AI goal-alignment problem specifically is to pursue a policy across the world of Differential Technological Development[213]. That is to say, there should be a clear policy to shift the balance of Research and Development spending towards a higher share for safety related research that seeks to resolve the problem. Or as Bostrom originally defined the concept, "societies would strive to retard the development of harmful technologies and their applications, while accelerating the development of beneficial technologies, especially those that offer protection against the harmful ones"[214]

While it may be too difficult (without the help of dedicated AI based control and supportive global governance) to prevent the development of a risky technology, it should be possible to reduce existential risk by speeding up the development of protective technologies relative to dangerous ones. This means ensuring that Beneficial AI, Friendly AI and all such programmes aimed at addressing the goal-alignment problem are well funded and that there is a suitable diversity of research.

Differential Technological Development is a role for research funders, who would be required to follow it as a key principle for use in designing funding calls and allocating grants: the aim would be to give additional weight to protective technologies, positive discrimination in favour of projects designed to develop Safe AI. And it could also be used by researchers when deciding which of several promising programmes of research to pursue, but this would only be required above a certain scale of investment. The overall spend on the development of Safe AI would need to be monitored centrally so as to be able to course correct as necessary.

8.8 Recommendations

a. ***Include within the Regulatory Framework measures such as Differential Technological Development designed to encourage the development and deployment of AI systems that will address the AI goal alignment problem and lead to Safe AI.***
b. *Promote the broad and speedy development of Beneficial AI, Friendly AI and other similar projects aimed at addressing the AI Alignment problem.*
c. *Ensure that there will always be a means to stop an AI process from continuing, with the ability to close down a whole AI system if it threatens human welfare. The means must not be able to be circumvented by a highly intelligent AI.*
d. *Develop a means of pausing the development of AI, with the support of AI if necessary, to be used in extremis.*

9 MILITARY USE

9.1 Introduction

One can imagine a myriad of military applications of AI. But current applications are already a source of immediate concern. For example, nuclear systems depend to an increasing and dangerous degree on AI, including radar systems that could provide faulty signals. The part that Stanislav Petrov played in saving the world from nuclear Armageddon in 1983 is well known[215]. Would a future Petrov, or Peters, or Phuong know that there was a fault in one of the sensors in an AI controlled hypersonic nuclear missile system and be in a position to stop catastrophe in time? It is important to make sure that the AI controllers in all nuclear-weapon states are up to the task, and decisions on life and death are not left to machines with faulty programming. Otherwise a global catastrophe could result.

The application that is causing immediate concern is that of Lethal Autonomous Robots (LARs) – often called Lethal Autonomous Weapon Systems, with its misleading abbreviation LAWS, or more dramatically “Killer Robots”. The rest of this paper focuses on the governance and regulation of LARs, with the expectation that solutions could provide a model for addressing other potential AI applications within the military.

9.2 The risk of LARs

In August 2007, roboticist Prof. Noel Sharkey warned against the development of fully autonomous robots that make their own decisions on lethal actions and called for their international regulation[216]. In September 2009 the International Committee for Robot Arms Control (ICRAC) was formed, followed in April 2013 by the Campaign against Killer Robots[217]. Support for the campaign to ban LARs has continued to strengthen ever since.

In May 2014, at a meeting of the state parties to the Convention on Conventional Weapons (CCW), 87 nations, together with UN agencies and NGOs, participated in the first multilateral meeting on “lethal autonomous weapon systems”, under the auspices of the Convention on Conventional Weapons.[218] The day before the opening of this meeting, twenty-one Nobel laureates issued a joint call for a ban on fully autonomous weapons.[219] In November that year, more than 70 faith leaders of various denominations endorsed an interfaith call to action against fully autonomous weapons – and a month later the Dalai Lama and other Nobel Peace Laureates issued a similar declaration.[220] An open letter signed by over a thousand AI and Robotics researchers headed by Professor Stuart Russell described the nature of autonomous weapons and concluded:

> …we believe that AI has great potential to benefit humanity in many ways, and that the goal of the field should be to do so. Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control.[221]

At the Paris Peace Forum in 2018, United Nations (UN) Secretary-General António Guterres called for a ban on killer robots, stating: "For me there is a message that is very clear – machines that have the power and the discretion to

take human lives are politically unacceptable, are morally repugnant, and should be banned by international law"[222].

9.3 Group of Government Experts

In 2016, a Group of Government Experts (GGE) was established by the CCW to discuss LARs[223]. Since then, discussions have continued each year. Currently 65 CCW states parties have endorsed group statements calling for a legally binding instrument to prohibit and restrict such weapons systems.[224]

Six states are identified as currently known to be developing LARs, namely the US, China, Israel, Russia, South Korea and the UK,[225] though some seem prepared to change position (e.g. the UK Minister for Counter Terrorism Alistair Burt's speech in the House of Commons in 2013[226] and the speech of Sir Roger Carr, Chair of BAE SYSTEMS, at the World Economic Forum in 2016: it would be a bad idea, he said, to build machines that decided "who to fight, how to fight and where to fight"[227]).

The Campaign against Killer Robots[228] argues that if it is not possible to launch negotiations by the CCW Review Conference in December 2021, then another forum is needed to discuss content and achieve the Campaign's goal of a treaty on LARS.

9.4 The way forward?

Lisa Bergstrom[229] has argued persuasively that it is in the interests of the United States to introduce a limited national ban on LARs. This is the tactic that the US used in relation to blinding laser weapons. With the laser weapons, the US had been concerned that other, less controversial uses of lasers, might be restricted. From a position of strength, having issued their limited national ban, they were able to negotiate an agreement in the CCW that satisfied countries that wanted a broader ban, countries opposed to a ban as well as the concerns of the US military.

A similar opportunity exists in relation to LARs. In 2012 the US Department of Defense issue a directive requiring "appropriate levels of human judgement over the use of force."[230] This was followed up by the publication in 2019 of "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence" by the US Department of Defense[231]. What this principles document shows is, if the Pentagon does wind up using AI in dangerous ways, it will not be because the Defense Department did not have good guidelines, it will be because the Department did not follow them[232].

The US Defense Department is clearly aware of the ethical issues. If the US were to use its limited national ban (moratorium on LARs) to serve as an example, it would be in a position of strength and leadership on this issue and with a chance of negotiating a global treaty with the signatures of all the nations currently researching LARs. Whether the treaty is in the form or a separate Convention, or as a Protocol within the CCW are options.

There is a further argument in favour of backing away from LARs. Some of the largest militaries in the world have been creeping towards developing such weapons, using a logic of deterrence: they fear being crushed by rivals' AI if they cannot unleash an equally potent force. But, as Frank Pasquale[233] suggests, they

should reflect carefully on what martial AI may be used for. LARs offer the potential for domestic oppression which no population would wish for. These populations need to express their views to their Governments.

9.5 Other aspects

It would be most valuable to reach a sound agreement between all the main and potential LARs developers. The results of the current GGE negotiations will be an important input but it may not be helpful to be too prescriptive regarding the terms of this agreement. It is to be hoped however that any such agreement will include defining guiding principles for human involvement in the use of force,[234] and will include, or be complemented by:

a. Developing *protocols* and/or technological means to mitigate the risk of unintentional escalation due to autonomous systems.
b. Developing *strategies* for preventing proliferation to illicit uses, such as by criminals, terrorists, or rogue states, or in domestic situations.
c. Conducting *research* to improve technologies and human-machine systems to reduce non-combatant harm and ensure International Humanitarian Law compliance in the use of future weapons

9.6 Verification

There should be new ways to expose the development of LARs so that the limit of meaningful human control is not overstepped before or after treaty signature.

As a protocol or treaty is being negotiated, a UN certification programme should be established that can independently certify that certain limits have not been exceeded. This might involve inspections and review of documentation.

Once a treaty is signed, a UN inspection regime should be created to verify the treaty, or at least serve as a confidence-building and warning system. If fully autonomous robotic weapons systems are developed, produced, stockpiled, or used, the UN system should identify the violators or nation(s) and a compliance system should be used to expose them and hold them in check.

The body required to handle the above could also handle non-military inspection (see Section 10). Whether the military or non-military requirement comes first, the institution should be designed to cater for both.

9.7 Recommendations

It is recommended that

a. A *moratorium* be introduced on research, development, production, stock-piling and deployment of Lethal Autonomous Robots including Autonomous Weapons of Mass Destruction (AWMD).
b. A *treaty* be negotiated, either as a Protocol of the Convention on Conventional Weapons or as a standalone Convention, to ban the research, development, production, stock-piling and deployment of Lethal Autonomous Robots.
c. A UN *monitoring and inspection regime* should be created for immediate confidence building and expertise development, and to verify an eventual treaty (cf Section 10.7).

10 GOVERNANCE / INSTITUTIONAL ISSUES

10.1 A Global Approach is needed

The significance of the impact of AI is such that the case for the governance and regulation of AI being organised globally is overwhelming. The major nations of the world are placing great emphasis commercially and technically on the development of a strong AI capability and there is talk of a new AI arms race. The role of work in people's lives is being challenged by AI: the way that people spend their lives and indeed find meaning in their lives is in question. And most fundamentally of all, the very future of humanity could be challenged by AI.

Not only will the governance need to be global, but in time it will require 100% participation. The sooner that the people and nations of the world come to understand this need, the greater the chance that suitable governance can be put in place in time to protect humanity.

There are currently several international governance initiatives underway as set out in Section 3.2. Discussions are taking place between representatives of the Council of Europe (CAHAI), the OECD, the European Union, ITU, ISO and UNESCO regarding ongoing work within each organisation, with the objective of "promot[ing] synergies in the development of a legal framework on AI, in which the specific contribution and expertise of each organisation can be highlighted and complement each other"[235]. Such initiatives are to be supported and encouraged, even though they represent regional or partial AI governance as they are addressing immediate challenges. Their development, and the coordination outlined above, should take place however on the understanding that a Global Framework Convention (Section 10.4/5) is the way forward. All current negotiations should be designed to facilitate such an evolution.

UN Secretary General Guterres[236] stated, in his Report on the Roadmap for Digital Cooperation: "Current artificial intelligence-related initiatives lack overall coordination in a way that is easily accessible to other countries outside the existing groupings, other United Nations entities and other stakeholders. There are currently over 160 organisational, national and international sets of artificial intelligence ethics and governance principles worldwide. However, there is no common platform to bring these separate initiatives together." There is an overwhelming argument that not only do these principles need to be brought together but that many need to be enshrined in international law.

Section 10.2 identifies some institutions with relevant skills and expertise in this respect. A review of the broad options regarding AI Global Governance (Section 10.3) leads to the proposal to consider the development of a suitable Framework Convention (Section 10.4/5**).**

10.2 Possible organisations to form the basis of an AI global governance

There are a number of different institutions and regimes that may be relevant when considering the way forward for AI. Some of the initiatives outlined in section 3.2 have overtly put their hat in the ring, but there are other institutions that have relevant skills and experience that it is worth considering. These include UNDRR, the IPCC, UNICRI, IAEA, each of which is reviewed below. One might use

elements of each of them but it is not clear that any of them provides an adequate template in itself.

UNDRR
The United Nations Office for Disaster Risk Reduction (UNDRR) was created in December 1999 to ensure the implementation of the International Strategy for Disaster Reduction. This strategy was limited to natural disasters. One option could be to expand the remit of the UNDRR to address Man-made Disasters, such as catastrophic risk arising from technological developments such AI.

The UNDRR is very small however, with a focus on the Sendai Disaster Risk Reduction Framework.

IPCC
The Intergovernmental Panel on Climate Change (IPCC) provides an objective, authoritative assessment of the scientific evidence in relation to climate change. Some have argued that AI needs such a body. Nicholas Miailhe[237] argued that much like the IPCC, we need an Intergovernmental Panel on AI that many scientists around the world could contribute to as we create a series of facts that we can use to guide policy. He argued that "given the high systemic complexity, uncertainty and ambiguity surrounding the rise of AI, its dynamics and its consequences – a context similar to climate change – creating an IPCC for AI, or 'IPAI', can help build a solid base of facts and benchmarks against which to measure progress". There is now a GPAI, but perhaps not exactly an IPCC for AI.

UNICRI
A global organisation is required, which could control AI development over a certain level of intelligence and capability. A candidate for this role[238] is the United Nations Interregional Crime and Justice Research Institute (UNICRI). It has been working with Interpol to explore the role of AI and robotics in law enforcement.

IAEA
There are several similarities between the International Atomic Energy Authority (IAEA) and an organisation seeking to keep Artificial Intelligence research within certain boundaries. In both cases there is

a. Something that needs to be kept under control
b. Something therefore that needs to be monitored
c. A need to enforce compliance if a breach is identified

There is a distinction between the AI requirement and the IAEA in that the issue already exists with the IAEA. Nuclear monitoring involves both those countries that do have nuclear weapons and those countries that do not have nuclear weapons but should have nuclear energy for use for peaceful means. That is distinct from the AI situation. Currently there is no threat to loss of control anywhere – but as the risk increases it becomes of paramount importance that there is no loss of control anywhere in the world. Reliably monitoring nuclear materials is just about possible (though rogue states can present a challenge). Reliably monitoring advanced AI is feasible today because they consume huge amounts of computing power. But if their architecture is more distributed, this footprint would become less distinct. Tracking down some code around the world will require a very high, if specific, degree of surveillance.

A key issue with both the IAEA and with AI is enforcement. In the short-term conventional remedies can be applied such as trade sanctions or the country concerned being excluded from a relevant Regime or Global institution. With the control issue however, conventional trade sanctions are irrelevant: a strong, swift and effective enforcement would be essential.

10.3 Cooperative or centralised governance organisations

In Section 3.2, the some of the international bodies that have in the past few years indicated their intention to engage in the governance of Artificial Intelligence are identified. This list of institutions is an indication of the breadth of impact of AI on the human existence – and also of the importance now given to the governance of AI around the world. The international community has seized the opportunity and rushed to fill the gap – but not in an organised fashion. It is important to assess whether the current "AI ethics and governance competition" is the best approach or whether there is a better way.

Multiple competing governance structures

The current situation regarding ethical guidance, governance and regulation of Artificial Intelligence is somewhat akin to the Wild West. There is this magnificent new territory stretching as far as the eye can see. There has been little or no AI regulation until recently, the general argument being that it has been too early: the technology was considered too immature and concern was expressed that premature regulation could prove over restrictive to the rapidly evolving technology. In the last two years however, the mood has changed, the starting gun has been fired and a series of governance wagons are charging off to claim the new virgin territory.

Some bodies are restricting themselves for now to ethical Principles and to Recommendations; the Council of Europe is claiming to be the first ever intergovernmental committee group to develop a legal framework for the design, development and application of artificial intelligence whilst UNESCO is claiming that if adopted, its Recommendation on the Ethics of AI will be the first global normative instrument to address the developments and applications of AI. There are variously estimated to be 160 different sets of AI Principles[239], and over 250 Principles, Governance and Regulation initiatives[240]. With no common platform, it is difficult to see that this global governance is the optimal approach to addressing an issue of such paramount importance.

Will there be a winner?

With any race or competition, there is finally a winner. One could envisage that one of the contending institutions could eventually emerge the winner. But the breadth of impact of AI is such that no single existing institution has the requisite remit, skills or experience to embrace the whole of AI governance.

The UN Body that is the strongest current candidate to establish an AI Governance Regime is UNESCO and it is possible that UNESCO could decide that that is what it wished to do. This would mean expanding firmly into the economic arena and in due course to critical global enforcement. That could happen. But would it be the best way for humanity?

The GPAI has been established as a group of like-minded countries. China is not a member as it is not seen as being 'like-minded'. This would suggest that GPAI is not going to develop into the key global body. China is not comfortable with institutions that it has not been a party to establishing. GPAI is sure to advance thinking on AI, but it was not conceived as the Governance body itself and is unlikely to be so.

A phased approach
An alternative approach would be to accept the principles, governance and regulation that is emerging from various institutions at present, whilst in parallel setting out to create the common platform to which UN Secretary General Guterres refers. Such a pathway is outlined in Section 10.4 below.

10.4 Creation of a new Framework Convention for AI and other Disruptive Technologies

Many of the bodies reviewed above have something to contribute to AI governance, but none of the above institutions represents a perfect fit. The scope of AI itself is broad, embracing the meaning and availability of work and the future of humanity, quite apart from the more immediate issues of the threat to epistemology, moral deskilling etc. There seems therefore to be a very strong case for the creation of a tailor-made Convention to address the issues of AI governance.

Framework Conventions
There are different ways to approach such a Convention, but the most suitable would appear to be to establish a Framework Convention. Key elements of a Framework Convention include

a. An agreed scope of activity
b. An overall objective to which most people would be happy to sign up to, helping to create a Forum involving the whole world, where the more contentious commitments can be debated
c. Commitments often limited to formal procedural issues around a regular Conference of the Parties, with few if any substantive commitments.

More substantive commitments are typically concentrated in Protocols, where it is possible that there might not be 100% sign-up

A Framework Convention is particularly suitable[241] where:

a. Political consensus to take strong substantive measures is lacking
b. Scientific understanding is still evolving
c. The problem itself is changeable

In the case of AI, all three reasons would appear to apply. There has been a reluctance from key Governments to embark on regulating AI and there is certainly no consensus on how to handle the control problem. Scientific understanding of AI and particularly AGI continues to evolve, with one of the leaders[242] (Stuart Russell) suggesting that the whole current construct is dangerous and needs to be transformed. There is no universal agreement as to the impact of AI on employment, with many seeing the likely impact as profound.

And AI itself is evolving as researchers around the world seek ways towards Artificial General Intelligence and beyond.

Furthermore, a Framework Convention would seem particularly suitable for a technology that could have such a profound impact on the future of humanity. At some point, it is likely that there will need to be complete agreement with regard to the measures needed to protect mankind. The protocols might not have 100% signature rate in the initial stages, though over time complete global commitment is likely to become necessary.

Scope

The scope of the Framework Convention would need serious discussion at the outset. It could be limited to Artificial Intelligence, or it could be expanded to embrace other similar issues.

"Other Disruptive Technologies" (for example Nanotechnology) typically offer a combination of benefits and risks and the management of some of the risks is likely to be similar to the management of some AI risks. There is, as such, a clear logic in combining other Disruptive Technologies within the same Framework Convention, allowing Protocols to be developed for each new Disruptive Technology. There is a currently a move to consolidate Environmental Conventions into one super-convention[243], so as to reduce the administrative burden, particularly on developing countries. Including all Disruptive Technologies within the same Framework Convention from the start would provide the administrative efficiency that is being sought in the Environmental field.

AI represents longer term an existential risk for humanity if the control problem is not resolved. There are other existential risks however, as discussed in Section 8.1. The Framework Convention could be conceived as addressing all existential risk. Whilst there is no doubt that there should be a UN body responsible for reviewing catastrophic and existential risk, there are very wide variations in terms of the extent to which the global population would be involved. The development of a system that could deflect large asteroids on a dangerous path towards earth would be costly, but would otherwise have a limited impact on the world's population, in contrast to the steps necessary to avoid the loss of control of AI. The scope of the Framework Convention could be expanded to embrace man-made catastrophic risks as well as natural disasters.

A key element of the longer-term issue of Artificial Intelligence is its potential impact on the future of humanity. There are other technologies such as genetic engineering, which could also affect the future of humanity, albeit in a somewhat different manner. But this example could be seen as being classified as a Disruptive Technology.

Finally, another option would be to combine AI with other digital technologies. In any event, his requires further discussion and debate: for now, it is suggested that the scope of the Framework Convention should be "AI and other Disruptive Technologies".

Name
The full title may well be too long for everyday use and require an abbreviation. For this paper, the abbreviation used is UN Framework Convention on AI (UNFCAI).

Objective
The objective itself will be key and require serious discussion. It needs to be something that carries real weight and which has widespread support.

A potential objective would be "to develop AI and other Disruptive Technologies for the benefit of all, whilst ensuring that a flourishing future for humanity is not put at serious risk."

Initial Commitments and subsequent Protocols
The amount of AI governance substance that is included in the Framework Convention itself, as distinct from a subsequent protocol, would depend upon several factors such as the timing of the establishment of the Framework Convention, the progress made by the existing AI Governance Bodies (See Section 3.2) and the urgency of the issues. More controversial proposals would be best left to a Protocol so as to maximise the number of countries prepared to sign up to the Framework Convention from the start.

In due course one could imagine a number of protocols such as:

a. The loss of control issue – about development and early deployment: needs very close surveillance (links to UNICRI / IAEA?))
b. The economic and social issue: this is about Government policies, and so is totally different from loss of control in terms of how to approach the problem. In conjunction with some other body (such as UNESCO / OECD / ILO)
c. Short term / immediate issues (UNESCO / CoE)
d. Engineered pandemics (Links to WHO)
e. Genomics: (Links to WHO) There could be a specific Protocol about Genomics – but also this could link in with longevity, and the whole issue of extension of life of the body.

10.5 Transition to Framework Convention

The formation of the UN Secretary General's Advisory Group on AI (See Section 10.1) is a thoroughly appropriate way forward, bringing the global community together to discuss common issues in an open and loose framework. This group should not be seen as an end in itself, but rather as an important step in a necessary transition.

A useful comparison for an AI related Framework Convention is the UN Framework Convention on Climate Change (UNFCCC). The subject matter of the two Framework Conventions is quite distinct, one addressing a widespread historic and growing pollution of greenhouse gases whilst the other addresses the development and application of new technologies. Nevertheless, there are strong similarities in the significance of the issues and their widespread ramifications.

Origins of the UNFCCC
A key step in the development of a climate change regime was the First World Conference in 1979. This eventually led to the establishment of an Intergovernmental Panel on Climate Change (IPCC) in 1988, which duly reported

two years later in time for a Second World Conference in 1990. There it was agreed to establish an International Negotiating Committee (INC) which got under way the following year and produced a text for signature in time for the World Summit in Rio de Janeiro in 1992.

Table 3 shows this time table for Climate Change in the first column of dates. If the AI Framework Convention were to be established over a similar time period, the next column shows that it would not be ready for signature before 2036. Whilst some of the immediate issues are being addressed by the current set of AI-related governance initiatives, the lack of a single co-ordinating AI related regime before that date would reflect a disastrous dereliction of duty. The right-hand column suggests a more ambitious timescale. An AI regime is not facing quite the same Machiavellian resistance from "AI deniers" or vested interests as the Climate Change regime experienced. It should be possible to dispense with one World Conference and aim to set up an International Negotiating Committee soon after the first World Conference. If the World Conference on AI were to take place in 2023, that could enable the Framework Convention text to be open for signature by 2026.

Event	Climate Change	AI and Disruptive Technologies	
		In pace with CC	Target
UNSG Advisory Group established			2020
First World Conference	1979	2023	2023
Intergovernmental Panel set up	1988	2032	2024
First report	1990	2034	
Second World Conference	1990	2034	
International Negotiating Committee (INC) established	1991	2035	2024
INC agrees text; Text open for signature	1992	2036	2026

Table 3: Framework Convention timelines

This will be all be in parallel in the short-term with the functioning of organisations such as the Council of Europe and UNESCO. They would hopefully address any immediate issues that arise, whilst the new Framework Convention gets under way and starts to develop one or more Protocols, working with the current AI governance initiatives to avoid overlap and ensure a seamless overall AI governance and a global legal mechanism for the regulation of AI.

10.6 Complementing the Framework Convention

There are other logical institutional steps that need to be made to complement the Framework Convention, namely

Protocol on AI

Once the Framework Convention on AI has been negotiated and come into force, negotiation can start on the Protocol on AI. This would seek to deliver the first set of global AI regulations. It is likely to prove more difficult to negotiate than the Framework Convention. The meeting would hopefully have all the parties in the room for the negotiation, though the concluding protocol might not have all parties as willing signatories, at least not initially.

Given current momentum towards regulation of AI within the Council of Europe and elsewhere, one can fully expect that legislation (binding on those states that ratify) will exist several years before the UNFCAI Protocol has been negotiated. The protocol can be edited in the light of what exists at the time so as to minimise any transitional confusion.

IPCC equivalent
The Intergovernmental Panel on Climate Change (IPCC) has played a key role in supporting the UNFCCC, providing it with objective science advice and policy options. The UNFCAI would need a comparable objective source of information in relation to AI.

The Global Partnership on AI (GPAI) was reportedly designed to play a similar role in relation to AI that the IPCC plays in relation to Climate Change.[244] At present it remains to be seen how GPAI (and its membership) develops but it is possible that GPAI could evolve into an institution providing the UNFCAI with the authoritative information and support that it needs. If this were not possible, a new organisation (IPAI) would need to be created but that would seem a wasteful duplication.

AI Global Authority (AIGA)
In addition, a body will be needed to provide an inspection. Here the IAEA and the OPCW provide models and the United Nations Interregional Crime and Justice Research Institute (UNICRI) also has relevant capability. The AIGA could address both military and civilian requirements, and should be designed to do so whether first set up for military purposes or vice versa. Such an institution would be capable of inspection and enforcement.

An associated Parliamentary Assembly
Many international treaties have an associated Parliamentary Assembly associated with them such as the Council of Europe, the WTO and NATO. The establishment of an associated supervisory body might, as with other treaty-based institutions, provide a democratic input and a constructive monitoring role.

10.7 Recommendations

The establishment by the UN Secretary General of an Advisory Group on AI co-operation should be followed by a decision to hold a multi-stakeholder World Conference on the Governance of AI. These two steps should be seen as preparatory to the negotiation of

a. **a UN Framework Convention on AI**
b. **a Protocol on AI.**

To support the negotiation and implementation of these agreements providing a global legal framework for the regulation of AI, new bodies should be established:

a. A **Global Panel on AI**, possibly building upon the Global Partnership on AI, providing a technical support analogous to that provided by the IPCC to the UNFCCC.
b. An **AI Global Authority**, empowered to provide monitoring and inspection to support the work of the UN Framework Convention on AI (cf Section 9.7)
c. A **supervisory body** might also be associated with a democratic input as with other treaty-based institutions.

11 PHASED SUMMARY

The recommendations set out in the body of the paper are summarised at the beginning on pages 7 and 8. These recommendations relate both to short term action within the next few years and to some longer-term issues with a profound significance. The reason for combining the two types of issue is that the longer-term issues are of such significance that action needs to be take now in order to prepare for them. This action spans both AI research, debate as to the future of humanity that we want and bringing about the changes in society that will be necessary if we are to be able to experience that future. The table below is a high-level summary of the phasing of these different areas of activity over the near term and the longer term.

<table>
<tr><td>S</td><td></td><td>Short term</td><td>Medium term</td><td>Longer term</td></tr>
<tr><td></td><td></td><td>Now to 2025</td><td>2025-2030</td><td>2030-2070</td></tr>
<tr><td>2</td><td>Categorisation</td><td>Develop categories for Good and Bad AI</td><td colspan="2">Review, update and apply</td></tr>
<tr><td>4</td><td>Values and principles</td><td>Negotiate global set of principles</td><td colspan="2">Review, update and apply</td></tr>
<tr><td>5</td><td>Social media</td><td>Immediate regulation required</td><td colspan="2">Update regulation as required</td></tr>
<tr><td>6</td><td>How to respond to the impact of AI on work</td><td>Start the global debate, with global citizens assemblies</td><td>Agreement on a common approach</td><td>Coordination of implementation of common approach</td></tr>
<tr><td>7</td><td>Deskilling (including moral)</td><td>Research and plan response</td><td colspan="2">Educational and cultural campaign</td></tr>
<tr><td>8</td><td>Beneficial AI and other Safe AI initiatives</td><td colspan="2">Support</td><td>Deliver</td></tr>
<tr><td>8</td><td>Before transition beyond current human existence</td><td>Develop teaching / communication material</td><td>Deploy the material. Launch debate</td><td rowspan="2">Is humanity ready to transition?
If not deploy Pause capability, if available, to retain desirable options.</td></tr>
<tr><td>8</td><td>Pause capability</td><td colspan="2">Support research</td></tr>
<tr><td>9</td><td>Military</td><td>Moratorium, Protocol banning LARs and Autonomous WMD.</td><td colspan="2">New CCW protocols as necessary</td></tr>
<tr><td>10</td><td>Governance / Institutional</td><td>Negotiate Frame-work Convention; Establish institutions</td><td colspan="2">Further treaties and institutional evolution, as the challenges of AI governance evolve</td></tr>
</table>

Table 4: Phased summary

12 CONCLUSIONS

Artificial Intelligence is a fast-growing technology with an immense amount of value to offer humanity. There is an awareness of the need to strengthen the governance and regulation surrounding AI, though the extent and timing of that governance and regulation is the subject of debate. What is clear is that

a. The nature of AI and its expected impact on humanity are such that the governance needs to be global.
b. It is important that that governance addresses not only current issues, but also the need to be able to evolve as rapidly as the technology.
c. That governance needs to ensure that due priority is given to the preparation for some of the major future impacts identified in this paper.

BIBLIOGRAPHY

- **Ad Hoc Expert Group (AHEG) for the Preparation of a Draft text of a Recommendation on the Ethics of Artificial Intelligence, (2020).** *First Draft Of The Recommendation On The Ethics Of Artificial Intelligence.* UNESCO. Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000373434?utm_content=buffer89e47&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer>
- **AIHLEG: European Commission's High-Level Expert Group on Artificial Intelligence, (2018)** *Ethics Guidelines for Trustworthy AI*. European Commission. Available at: https://ec.europa.eu/digital-single-market/en/ news/ethics-guidelines-trustworthy-ai
- **AIHLEG: European Commission's High-Level Expert Group on Artificial Intelligence, (2020)** Artificial Intelligence – A European Approach
- **Algorithm Watch (2020).** *AI Ethics Guidelines Global Inventory: About*. [online] Available at: <https://inventory.algorithmwatch.org/about> [Accessed 4 November 2020].
- **Ali, I (2018)** 'U.S. Military Puts 'Great Power Competition' at Heart of Strategy: Mattis', *Reuters* 19 Jan. Available at: <https://www.reuters.com/article/us-usa-military-china-russia/u-s-military-puts-great-power-competition-at-heart-of-strategy-mattis-idUSKBN1F81TR>
- **Arkin et al. (2018)** 'Lethal Autonomous Systems and the Plight of the Non-combatant' *AISB Quarterly, July 2013, Vol. 137,*
- **Armstrong, S. (2014)** *Smarter Than Us – the rise of machine intelligence*, US: MIRI Berkeley
- **Batin, M., Turchin, A., Sergey, M., Zhila, A., Denkenberge, D., (2017).** 'Artificial Intelligence in Life Extension: From Deep Learning to Superintelligence' *Informatica* 41 p.401-417. Available at: <http://www.informatica.si/index.php/informatica/article/view/1797>
- **Bergstrom, L., (2019). '**The United States should drop its opposition to a killer robot treaty'. *Bulletin of the Atomic Scientists*, 7 November [online] Available at: <https://thebulletin.org/2019/11/the-united-states-should-drop-its-opposition-to-a-killer-robot-treaty/>

- **Bloomberg (2018)** *The Rise of AI.* Bloomberg Businessweek
- **Bodansky, D. (1999)** *The Framework Convention / Protocol Approach.* WHO. Technical briefing series (Framework Convention on Tobacco Control) (1). Available at: < https://apps.who.int/iris/handle/10665/65355>
- **Bostrom, N. (2002)** 'Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards', *Journal of Evolution and Technology*, 9(1).
- **Bostrom, N. (2014)** *Superintelligence.* Oxford: Oxford University Press
- **Burton-Hill, C., (2016.) '**The superhero of artificial intelligence: can this genius keep it in check?' *The Guardian*, 16 February [online] Available at: <https://www.theguardian.com/technology/2016/feb/16/demis-hassabis-artificial-intelligence-deepmind-alphago>
- **CAHAI (2020)** 'Item 31: Synergy and complementarity of CAHAI's work with that of other international organisations' *1384th meeting, 23 September 2020.* Ad Hoc Committee on Artificial Intelligence (CAHAI). Available at: < https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016809ed062>
- **The Campaign to Stop Killer Robots. (2020a)**. *All Action and Achievements*. [online] Available at: <https://www.stopkillerrobots.org/action-and-achievements/>
- **The Campaign to Stop Killer Robots (2020b)**. **'**Diplomatic talks re-convene**'** *www.stopkillerrobots.org* 25 September. Available at: < https://www.stopkillerrobots.org/2020/09/diplomatic2020/>
- **Carr, N. (2010)** *The Shallows: How the Internet is Changing the Way We Think, Read and Remember.* Atlantic Books
- **Carr, N. (2013)** *All Can Be Lost: The Risk of Putting Our Knowledge in the Hands of Machines.* The Atlantic, November
- **Carr, N. (2016)** *The Glass Cage: Who Needs Humans Anyway?* Vintage
- **Cath, C. (2018).** 'Governing artificial intelligence: Ethical, legal and technical opportunities and challenges'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133). https://doi.org/10.1098/rsta.2018.0080
- **Cevasco, L., Corvalan, J., Le Fevre Cervini, E. (2019)** *Artificial Intelligence and Work – Building a New Employment Paradigm*, Buenos Aires: Editorial Astra
- **Chase, Callum (2018)** *The Economic Singularity* Oxford: Three C's Publishing
- **Chui, M., Manyika, J. and Miremadi, M., (2015)** 'Four Fundamentals of Workplace Automation'. *Mckinsey Quarterly*,2016 (number 1) [online] Available at: <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/four-fundamentals-of-workplace-automation#> [Accessed 9 November 2020].
- **Clark, G. (2007)** *A Farewell to Arms,* New Jersey: Princeton University Press,
- **Coeckelbergh, M. (2012).** 'Technology as skill and activity: revisiting the problem of alienation'. *Techne, 16*(3), 208–230.
- **Coeckelbergh, M. (2020).** *AI Ethics.* Cambridge MA: MIT Press.
- **Council of Europe (2020)** *CAHAI – Ad hoc Committee on Artificial Intelligence.* Available at*:* < https://www.coe.int/en/web/artificial-intelligence/cahai>
- **Csikszentmihalyi, M, (2002)** *Flow: The Psychology of Happiness.* UK*:* Rider
- **Czarnecki, T (2020),** *Becoming a butterfly – Extinction or Evolution? Will Humans Survive beyond 2050?* UK: Sustensis

- **Deep Genomics (2020)** *Creating a New Universe of Genetic Medecine. Available at* <https://www.deepgenomics.com/>
- **Defense Innovation Board (2019)** *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense.* USA: Department of Defense
- **Diamandis, P., Kotler, S. (2015)** *Abundance: The Future is Better Than You Think.* Free Press
- **Eager, J., Whittle, M., Smit, J., Cacciaguerra, G., & Lale-demoz, E. (2020).** *Opportunities of Artificial Intelligence*. European Parliament: Policy Department for Economic, Scientific and Quality of Life Policies.
- **The Economist (2020)** 'An understanding of AI's limitations is starting to sink in' *The Economist.* 11 June. Available at: <https://www.economist.com/technology-quarterly/2020/06/11/an-understanding-of-ais-limitations-is-starting-to-sink-in>
- **European Commission (2018)** *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.* Brussels: European Commission. COM (2018) 237 final Available at: < https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN>
- **European Commision (2020)** *Policy: Artificial Intelligence*. Available at: <https://ec.europa.eu/digital-single-market/en/artificial-intelligence>
- **Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020).** *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI.* Berkman Klein Center for Internet & Society. Available at: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420>
- **Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018).** 'AI4People— an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations.' *Minds and Machines*, 28(4), 689–707. doi: <https ://doi.org/10.1007/s11023-018-9482-5.>
- **Ford, M. (2016)** *Rise of the Robots – Technology and the Threat of Mass Unemployment.* One World Publications
- **Frey, C., Osborne, M. (2013)** *The Future of Employment: How Susceptible are Jobs to Computerisation* Oxford Martin School
- **Friedman, M. (1962)** *Capitalism and Freedom*. Chicago: University of Chicago Press
- **Future of Life Institute (2015)** *Autonomous Weapons: an Open Letter from AI & Robotics Researchers* [online]. Available at: < https://futureoflife.org/open-letter-autonomous-weapons/>
- **Gartner (2020) '**Gartner Identifies Five Areas Where AI Can Improve Decision Making for Government and Healthcare CIOs During the Coronavirus Pandemic' *Gartner* 4 May. Available at: <https://www.gartner.com/en/newsroom/press-releases/2020-05-04-gartner-identifies-five-areas-where-ai-can-improve-decision-making-during-the-coronavirus-pandemic>
- **Goertzel, B (2012)** 'Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood?', *Journal of Consciousness Studies.* 19(1-2). pp.96-111
- **Good, I. J. (1965).** 'Speculations concerning the first ultra-intelligent machine'. *Advanced in Computers,* 6: pp.31-88. doi: <https://doi.org/10.1016/S0065-2458(08)60418-0>

- **GOV.UK. (2019).** *A Guide to Using Artificial Intelligence In The Public Sector*. [online] Available at: <https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector> [Accessed 4 November 2020].
- **Green, B. (2018) '**Artificial Intelligence, Decision-Making and Moral Deskilling' *Markkula Center for Applied Ethics.* Available at: < https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/artificial-intelligence-decision-making-and-moral-deskilling/>
- **Guterres, A., (2018).** *Remarks At "Web Summit"*. [online] United Nations Secretary-General. Available at: <https://www.un.org/sg/en/content/sg/speeches/2018-11-05/remarks-web-summit>
- **G20 Trade Ministers and Digital Economy Ministers, (2019)** *G20 Ministerial Statement on Trade and Digital Economy*. Available at: https://www.mofa.go.jp/files/000486596.pdf
- **Hack (2019)** 'Your phone isn't spying on you – it's listening to your "voodoo doll"' *Hack* 2 May. Available at: <https://www.abc.net.au/triplej/programs/hack/your-phone-is-not-spying-its-listening-to-your-voodoo-doll/11073686>
- **Haenlein, M., and Kaplan, A. (2019).** 'A brief history of artificial intelligence: On the past, present, and future of artificial intelligence'. *California Management Review*, 61(4), 5-14. doi: <https://doi.org/10.1177/0008125619864925>
- **Harari, Y. (2017)** *Homo Deus***,** London: Vintage
- **HC Deb (2013)** Vol. 564, col. 733-738. Available at:<https://publications.parliament.uk/pa/cm201314/cmhansrd/cm130617/debtext/130617-0004.htm#130617-0004.htm_spnew1>
- **The He Lab (2018)** *About Lulu and Nana: Twin Girls Born Healthy After Gene Surgery as Single-Cell Embryos.* Available at: < https://www.youtube.com/watch?v=th0vnOmFltc&app=desktop&t=10s>
- **Huxley, A (1932),** *Brave New World* Vintage Classics
- **Hughes, C., (2019).** 'Opinion: It's time to break up Facebook'. *The New York Times,* 9 May [online] Available at: <https://www.nytimes.com/2019/05/09/opinion/sunday/chris-hughes-facebook-zuckerberg.html>
- **International Telecommunication Union (2019)** *United Nations Activities on Artificial Intelligence (AI)* Available at: <http://handle.itu.int/11.1002/pub/813bb49e-en>
- **Jahoda, M., Lazarsfeld P., Zeisel, H. (2009)** *Marienthal: the Sociography of an Unemployed Community.* New Jersey:Transaction Publishers,
- **Jepson, M., and Ryan, J. (2019**) 'Artificial Intelligence is helping us talk to animals (yes, really)' *Wired* 29 December Available at : <https://www.wired.co.uk/article/ai-talk-animals>
- **Jobin, A., Ienca, M., & Vayena, E. (2019).** 'Artificial Intelligence: The global landscape of ethics guidelines. arXiv :1906.11668 [Cs].
- **Kahneman, D. (2011)** *Thinking Fast and slow* Penguin Random House
- **Kaplan, J. (2015)** *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence.* Yale University Press
- **Kearney, M., Mogstad, M. (2019),** *Universal Basic Income (UBI) as a Policy Response to Current Challenges,* Aspen Institute Economic Strategy Group

- **Khatchadourian, R. (2015).** 'The Doomsday Invention' *The New Yorker*. The Tech Issue (November 23). Available at: < https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom> (Retrieved 15 January 2020).
- **Kooistra, P. (1993)** *Het Ideale Eigenbelang - Een UNO Marshall Plan voor alle your period*, Uitgeverij Kok Agora, Kampen, ISBN 978-90-391-0574-0. 201
- **Kurzweil Network (2020) '**AI tool detects Alzheimer's disease with 95% accuracy.' *Kurzweil accelerating intelligence*. 1 September. Available at: <https://www.kurzweilai.net/digest-ai-tool-detects-alzheimers-disease-with-95-percent-accuracy>
- ***LAIP - Linking Artificial Intelligence Principles*. (2020)** [online] Available at: <http://www.linking-ai-principles.org/> [Accessed 9 November 2020].
- **Lanier, J. (2018)** *Ten Arguments for Deleting Your Social Media Accounts Now*, UK: The Bodley Head
- **Leslie, D. (2019).** *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.* The Alan Turing Institute. doi: <https://doi.org/10.5281/zenodo.3240529>
- **Li, J. R., Walker, S., Nie, J. B., & Zhang, X. Q. (2019).** 'Experiments that led to the first gene-edited babies: the ethical failings and the urgent need for better governance'. *Journal of Zhejiang University. Science. B*, *20*(1), pp.32–38. doi: <https://doi.org/10.1631/jzus.B1800624>
- **Lomas, N., (2017) '**Lyrebird is a voice mimic for the fake news era' Tech Crunch 25 April. Available at: <https://techcrunch.com/2017/04/25/lyrebird-is-a-voice-mimic-for-the-fake-news-era/>
- **Lowrey, A. (2018)** *Give People Money,* USA: Penguin Random House.
- **Marr B. (2019)** *Artificial Intelligence in Practice - how 50 successful companies used AI and machine-learning to solve problems.* UK: Wiley.
- **McAfee, A., Brynjolfsson E. (2014)** *The Second Machine Age,* New York: Norton.
- **Miailhe, N. (2018) '**AI & Global Governance: Why We Need an Intergovernmental Panel for Artificial Intelligence' *United Nations University Centre for Policy Research*. 20 December. Available at: <https://cpr.unu.edu/ai-global-governance-why-we-need-an-intergovernmental-panel-for-artificial-intelligence.html>
- **Miailhe, N. and Bing Song (2020)** *Cross-Looks – AI, Lost in Translation?* Politico AI Summit 2020, 30 September-1 October, Online
- **Minsky, M. (1972)** *Computation: Finite and Infinite Machines*, Prentice Hall
- **MIRI (2020)** *About Miri.* Available at: <https://intelligence.org/about/>
- **Mondal, M., Bertranpetit, J. & Lao, O. (2019)** 'Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania'. *Nat Commun* 10, (246) doi: <https://doi.org/10.1038/s41467-018-08089-7>
- **Moriarty, D. (2020)** 'Information Gerrymandering – What is it and why does it matter?' *The Startup* 26 January. Available at:<https://medium.com/swlh/information-gerrymandering-what-is-it-and-why-does-it-matter-b6f07ac9370>
- **Musk, E., and Neuralink (2019) '**An integrated brain-machine interface platform with thousands of channels'. bioRxiv 703801; doi: https://doi.org/10.1101/703801

- **Natural Institute of Natural Sciences (2020)** 'Artificial Intelligence Identifies 80,000 Spiral Galaxies – Promises More Astronomical Discoveries in the Future'. [online] *SciTechDaily* 25 August. Available at: <https://scitechdaily.com/artificial-intelligence-identifies-80000-spiral-galaxies-promises-more-astronomical-discoveries-in-the-future/>
- **Nesta (2020)** *AI Governance Database.* Available at: <https://www.nesta.org.uk/data-visualisation-and-interactive/ai-governance-database/> (Accessed 12/11/2020)
- **Oberthur, S. (2002)** 'Clustering of Multilateral Environmental Agreements: Potentials and Limitations'. *International Environmental Agreements: Politics, Law and Economics.* 2, pp.317–340. doi: <https://doi.org/10.1023/A:1021364902607>
- **OECD. (2019a).** *Recommendation of the Council on Artificial Intelligence*. OECD Available at: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.>
- **OECD. (2019b).** *Forty-two countries adopt new OECD principles on Artificial Intelligence.* Available at: <https ://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm>
- **O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., Dafoe, A. (2020)** *The Windfall Clause – Distributing the Benefits of AI for the Common Good* Oxford: Future of Humanity Institute. Available at: <https://www.fhi.ox.ac.uk/windfallclause/>
- **Omohundro, S. M. (2008).** The basic AI drives. *Frontiers in Artificial Intelligence and Applications*, *171*(1), 483–492.
- **O'Neill, C (2016)** *Weapons of Math Destruction* US: Penguin Random House.
- **Ord, T. (2020)** *The Precipice Existential Risks and the Future of Humanity*, London: Bloomsbury.
- **Pasquale, F. (2020) ''**Machines set loose to slaughter': the dangerous rise of military AI'. *The Guardian.* 15 October [online]. Available at: < https://www.theguardian.com/news/2020/oct/15/dangerous-rise-of-military-ai-drone-swarm-autonomous-weapons>
- **Paul, C., and Posard, M., (2020)** 'Artificial Intelligence and the Manufacturing of Reality' *The Rand Blog.* 20 January. Available at: <https://www.rand.org/blog/2020/01/artificial-intelligence-and-the-manufacturing-of-reality.html>
- **People's Republic of China State Council (2017)** *New Generation Artificial Intelligence Development Plan.* Translated by Webster, G., Creemers, R., Triolo, P., Kania, E. Available at: <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/>
- **Pistono**, **F. (2012)** *Robots will Steal Your Jobs but That's OK* CreateSpace Independent Publishing Platform
- **Rideout, V. (2015)** *The Common Sense Census: Media Use by Tweens and Teens.* Common Sense media. Available at: <https://www.commonsensemedia.org/sites/default/files/uploads/research/census_researchreport.pdf>
- **Rieger, S., and Sinders, C. (2020)** *Dark Patterns: Regulating Digital Design* Berlin: Stiftung Neue Verantwortung Available at: < https://www.stiftung-nv.de/sites/default/files/dark.patterns.english.pdf>

- **Rifkin, J. (1995)** *The End of Work – The Decline of the Global Labor Force and the Dawn of the Post-Market Era,* New York: Penguin.
- **Rolnick, D., Donti, P., Kaack, L., Kochanski, K., Lacoste, A., Sankaran, K., Slavin Ross, A., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E., Mukkavilli, S. K., Kording, K., Gomes, C., Ng, A., Hassabis, D., Platt, J., Creutzig, F., Chayes, J., Bengio, J. (2019)** 'Tackling Climate Change with Machine Learning' arXiv:1906.05433v2
- **Royakkers, L., Timmer, J., Kool, L., & van Est, R. (2018).** 'Societal and ethical issues of digitization.' *Ethics and Information Technology*, 20(2), pp.127–142. doi: <https://doi.org/10.1007/s10676-018-9452-x.>
- **Russell, S., Norvig, P. (2003)** *Artificial Intelligence – a Modern Approach* (2nd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-790395-2
- **Russell, S. (2015)** *Long Term Future of (Artificial) Intelligence,* presentation at Centre for the Study of Existential Research, Cambridge
- **Russell, S. (2019)** *Human Compatible – AI and the Problem of* Control London: Allen Lane.
- **Salinas, S. (2017)** 'Facebook co-founder Sean Parker bashes company, saying it was built to exploit human vulnerability'. *CNBC.* 9 November. Available at: <https://www.cnbc.com/2017/11/09/facebooks-sean-parker-on-social-media.html>
- **Schick, N. (2020)** *Deep Fakes – and the Infocalypse.* UK: Octopus Publishing Group,
- **Seger, E. et al (2020)** *Tackling threats to informed decision-making in democratic countries: Promoting epistemic security in a technologically-advanced world* Centre for the Study of Existential Risk, Cambridge
- **Seldon, A., Metcalf, T. & Oladimeji, A. (2020)** *The Fourth Education Revolution Reconsidered – Will Artificial Intelligence Enrich or Diminish Humanity?* UK: University of Buckingham Press
- **Seligman, M. (2011)** *Flourish: A New Understanding of Happiness and Well-Being - and How to Achieve Them.* Nicholas Brealey Publishing
- **Sharkey, N., (2007).** 'Robot wars are a reality'. *The Guardian*, 18 August [online] Available at: <https://www.theguardian.com/commentisfree/2007/aug/18/comment.military>
- **Silberg, J., and Manyika, J. (2019) '**Tackling bias in artificial intelligence (and in humans)' *Mckinsey Global Institute*. 6 June. Available at: < https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>
- **Stiglitz, J. (2013)** *The Price of Inequality.* Penguin Books
- **Susskind, D. (2020)** *A World Without Work,* London: Allen Lane.
- **Susskind, D. and R (2015)** *The Future of the Professions: How Technology will Transform the Work of Human Experts.* Oxford university Press
- **Tegmark, M. (2017)** *Life 3.0: Being Human in the Age of Artificial Intelligence*. UK: Allen Lane
- **UNESCO, (2020).** *Major progress in UNESCO's development of a global normative instrument on the ethics of AI*. [online] Available at: <https://en.unesco.org/news/major-progress-unescos-development-global-normative-instrument-ethics-ai> [Accessed 24 October 2020].

- **UNICRI (2020)** *Timeline of AI strategic documents, effective as of April 2020.* Image. United Nations Interregional Crime and Justice Research Institute, Available at: <http://www.unicri.it/topics/ai_robotics>
- **UN Secretary-General (2020)** *Roadmap for Digital Cooperation*. United Nations. Available at: <https://www.un.org/en/content/digital-cooperation-roadmap/>
- **US Department of Defense Directive (2012),** *Autonomy in Weapons Systems*, US: Department of Defense (Number 3000.09)
- **United States, Executive Office of the President [Donald Trump] (2019)** Executive Order 13859: Maintaining American Leadership in Artificial Intelligence, Federal Register 84 pp.3967-3972
- **Vallor, S. (2015)** 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the ambiguous Future of Character' *Philos. Technol*. 28, pp.107-124. doi: <https://doi.org/10.1007/s13347-014-0156-9>
- **Vollset, S. et al (2020)** *Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the Global Burden of Disease Study.* The Lancet
- **Vosoughi, Roy and Aral (2018**),'The spread of true and false news online' *Science* 359(6380), pp. 1146-1151 doi: 10.1126/science.aap955
- **Wade, N. (2010)** *The Faith Instinct – How Religion Evolved and Why It Endures* Penguin Books
- **Waikar, S. (2020)** 'Treating COVID-19: How Researchers Are Using AI to Scale Care, Find Cures, and Crowdsource Solutions'. *Standford Institute for Human-Centred Artificial Intelligence*. 5 April. Available at: <https://hai.stanford.edu/blog/treating-covid-19-how-researchers-are-using-ai-scale-care-find-cures-and-crowdsource-solutions>
- **Wareham, M. (2020)** *Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control.* USA: Human Rights Watch. Available at: < https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and>
- **Weber, E. U. (2006)** 'Experience-Based and Description-Based Perceptions of Long-Term Risk: Why Global Warming does not Scare us (Yet)'. *Climatic Change*, 77, pp. 103–120 (21 July).
- **Wiener, J. (2016),** The Tragedy of the Uncommons: On the Politics of Apocolypse, *Global Policy.* 7(1). pp.67-80 doi: < https://doi.org/10.1111/1758-5899.12319>
- **Wilson, W. (1997)** *When Work Disappears: The World of the New Urban Poor*. N.Y: Knoff
- **World Economic Forum (2016)** *What if: Robots Go to War?* [online] Available at: <https://www.weforum.org/events/world-economic-forum-annual-meeting-2016/sessions/what-if-robots-go-to-war>
- **Wright, T. (2018)** 'The Return to Great-Power Rivalry Was Inevitable', *The Atlantic* 12 September. Available at: <https://www.theatlantic.com/international/archive/2018/09/liberal-international-order-free-world-trump-authoritarianism/569881[https://perma.cc/7W42-VUB6>
- **Yudkowsky, E. (2001)** *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures* San Francisco: The Singularity Institute.
- **Zeng, Y., Lu, E., & Huangfu, C. (2019**). 'Linking Artificial Intelligence Principles.' ArXiv, abs/1812.04814

- **Zhandry, M. (2016)** *Lecture 1*, lecture notes, Recent Developments in Program Obfuscation COS 597C, Princeton University, delivered 15 September 2016. Available at: < https://www.cs.princeton.edu/~mzhandry/2016-Fall-COS597C/ln/ln1.pdf>
- **Zuboff, S (2019),** *The Age of Surveillance Capitalism*, UK: Profile Books
- **Zuckerberg, M. (2019)** 'Mark Zuckerberg: The Internet needs new rules. Let's start in these four areas' *The Washington Post*. 30 March. Available at: <https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html>

NOTES

[1] Algorithm Watch (2020)
[2] Ord (2020)
[3] Ord (2020)
[4] Russell and Norvig, (2003)
[5] Mondal, Bertranpetit and Lao (2019)
[6] Deep Genomics (2020)
[7] Kurzweil Network (2020)
[8] Waikar (2020)
[9] Eager et al (2020)
[10] Rolnick et al (2019)
[11] Seldon et al (2020), p.32
[12] Seldon et al (2020), p.33
[13] Marr (2019).
[14] European Commission (2020)
[15] International Telecommunication Union (2019)
[16] Haenlein & Kaplan, (2019), p. 10-11
[17] Cath (2018), p. 2
[18] Marr (2019), p. 6
[19] Eager et al (2020), p. 9
[20] Natural Institute of Natural Sciences (2020)
[21] Batin et al (2017)
[22] The Economist (2020)
[23] UNICRI (2020)
[24] People's Republic of China State Council (2017)
[25]United States, Executive Office of the President [Donald Trump] (2019)
[26] European Commission (2018) p.1
[27] Council of Europe (2020)
[28] AIHLEG (2020)
[29] AHEG (2020)
[30] G20 (2019)
[31] Jobin, Ienca and Vayena, (2019).
[32] OECD (2019a)
[33] AHEG (2020)
[34] AHEG (2020) p.6
[35] AIHLEG (2018) p.11
[36] GOV.UK (2019)
[37] Leslie (2019) p.11
[38] Fjeld et al (2020)
[39] Zeng, Lu and Huangfu (2019)
[40] Floridi et al (2018) p.696
[41] Fjeld et al (2020) p.62-3
[42] Jobin, Ienca and Vayena, (2019) p.15
[43] AHEG (2020) p.8
[44] Fjeld et al (2020) p.60
[45] AIHLEG (2018) p.6-7
[46] AHEG (2020) p.6
[47] Leslie (2019) p.7
[48] Floridi et al (2018) p.699-700
[49] Leslie (2019) p.9
[50] Jobin, Ienca and Vayena (2019) p.14
[51] OECD (2019a)
[52] AHEG (2020)
[53] OECD, (2019b).
[54] G20 (2019)
[55] UNESCO (2020)
[56] Fjeld et al (2020) p.5
[57] OECD, (2019a).
[58] AHEG (2020)
[59] Zeng, Lu and Huangfu, (2019)
[60] Jobin, Ienca and Vayena (2019)
[61] Fjeld et al (2020)
[62] Floridi et al (2018)
[63] Royakkers et al (2018)
[64] Algorithm Watch (2020)
[65] Zeng, Lu and Huangfu (2019)
[66] Jobin, Ienca and Vayena (2019)
[67] Fjeld et al, (2020)
[68] Jobin, Ienca and Vayena (2019),
[69] Zeng, Lu and Huangfu (2019)
[70] Floridi et al (2018)
[71] Jobin, Ienca and Vayena (2019)
[72] Royakkers et al, (2018)
[73] Fjeld et al, (2020)
[74] Floridi et al (2018)
[75] Jobin, Ienca and Vayena (2019),
[76] Fjeld et al (2020)
[77] Jobin, Ienca and Vayena (2019)
[78] Jobin, Ienca and Vayena (2019),
[79] Royakkers et al, (2018)
[80] Zeng, Lu and Huangfu (2019)
[81] Jobin, Ienca and Vayena (2019),
[82] Fjeld et al, (2020)

[83] Miailhe, N. and Bing Song (2020)
[84] Schick (2020)
[85] Rideout (2015)
[86] Vosoughi, Roy and Aral (2018)
[87] Lanier (2018)
[88] Zuboff (2019)
[89] Salinas (2017)
[90] Lomas (2017)
[91] Lanier (2018); Schick (2020)
[92] Seger, E (2020)
[93] Rieger & Sinders (2020)
[94] Zhandry (2016)
[95] Moriarty, D (2020)
[96] Jepsen & Ryan (2019)
[97] Hack (2019)
[98] Paul & Posard (2020)
[99] Ali (2018); Wright (2018)
[100] Hughes (2019)
[101] Zuckerberg (2020)
[102] Hughes (2019)
[103] Hughes (2019)
[104] Zuboff 2019
[105] Silberg and Manyika,(2019)
[106] Cevasco et al (2019)
[107] Eager et al (2020)
[108] Gartner (2020)
[109] McAfee and Brynjolfsson (2014)
[110] Susskind and Susskind (2015)
[111] Susskind (2020)
[112] Kaplan (2015)
[113] Pistono (2012)
[114] Frey and Osborne (2013)
[115] Chui, Manyika and Miremadi (2015)
[116] Ford (2016)
[117] Rifkin (1995)
[118] Ford (2016)
[119] Susskind (2020)
[120] Chase (2018)
[121] Vallor (2015)
[122] Ford (2016)
[123] Ford (2016)
[124] Carr (2010)
[125] Ford (2016)
[126] Chace (2018)
[127] Chace (2018)
[128] Clark (2007)
[129] Chace (2018)
[130] Susskind (2020)
[131] Chace (2018)
[132] Chace (2018)
[133] Wade (2010)
[134] Seligman (2011)
[135] Csikszentmihalyi (2002)
[136] Ford (2016)
[137] Chace (2018)
[138] Vollset et al (2020)
[139] Ford (2016)
[140] Chace (2018)
[141] Lowrey (2018)
[142] Ford (2016)
[143] Friedman (1962)
[144] Rifkin (1995)
[145] Susskind (2020)
[146] Kooistra (1993)
[147] Kearney and Mogstad (2019)
[148] Rifkin (1995)
[149] Rifkin (1995)
[150] Susskind (2020)
[151] Kooistra (1993)
[152] O'Keefe et al (2020)
[153] O'Keefe et al (2020)
[154] Diamandis and Kotler (2015)
[155] Stiglitz (2013)
[156] Harari (2017)
[157] Huxley (1932)
[158] Chace (2018)
[159] Chace (2018)
[160] Chace (2018)
[161] Carr (2013)
[162] Vallor (2015)
[163] Green (2018)
[164] Vallor (2015)
[165] Green (2018)
[166] Green (2018)
[167] As quoted in an interview for the Guardian, Burton-Hill (2016)
[168] Russel (2019)
[169] Russel (2019)
[170] Russel (2019)
[171] Russel (2019)
[172] Ord (2020)
[173] Wiener (2016)
[174] Kahneman (2011)
[175] Weber (2006)
[176] Kahneman, (2011); Wiener, (2016) as cited in Ord, (2020)
[177] Ord (2020)
[178] Bostrom (2014)
[179] Minsky (1972)
[180] Armstrong, S. (2014)
[181] Omohundro (2008)
[182] Said during a lecture in 1951
[183] Wiener (2016)
[184] Good (1965)
[185] Yudkowsky (2001)
[186] Bostrom (2014)
[187] Russel (2019)
[188] Khatchadourian, (2015)
[189] Khatchadourian, (2015)
[190] Coeckelbergh (2020)
[191] Bostrom (2014)
[192] Armstrong (2014)
[193] A land-mark presentation to the Centre for the Study of Existential Research (CSER) at Cambridge University in 2015 entitled The Long-Term Future of (Artificial) Intelligence. Russel (2015)
[194] Russel (2015)
[195] Russel (2015)
[196] Yudkowsky (2001)

[197] Russel (2015)
[198] MIRI (2020)
[199] Omohundro (2008)
[200] Goertzel (2012)
[201] Russel (2015)
[202] The He Lab (2018)
[203] Ord (2020)
[204] Ord (2020)
[205] Ord (2020)
[206] Tegmark (2017)
[207] Musk, E. and Neuralink (2019)
[208] Ord (2020)
[209] Ord (2020)
[210] Ord (2020)
[211] Ord (2020)
[212] Goertzel (2012)
[213] Ord (2020)
[214] Bostrom, N (2002)
[215] Ord (2020), p.96
[216] Sharkey (2007)
[217]The Campaign to Stop Killer Robots (2020a).
[218] The Campaign to Stop Killer Robots, (2020a)
[219] The Campaign to Stop Killer Robots, (2020a)
[220] The Campaign to Stop Killer Robots (2020a)
[221] Future of Life Institute (2015)
[222] Guterres (2018)
[223] The Campaign to Stop Killer Robots (2020a)
[224] The Campaign to Stop Killer Robots (2020a)
[225] Wareham (2020)
[226] HC Deb (2013)
[227] World Economic Forum (2016)
[228] The Campaign to Stop Killer Robots (2020b)
[229] Bergstrom (2019)
[230] US Department of Defense (2012)
[231] Defense Innovation Board (2019)
[232] Defense Innovation Board (2019) p.2
[233] Pasquale (2020)
[234] Arkin et al (2018)
[235] CAHAI (2020)
[236] UN Secretary-General (2020)
[237] Miailhe (2018)
[238] Czarnecki (2020)
[239] Algorithm Watch (2020).
[240] Nesta (2020)
[241] Bodansky (1999)
[242] Russel (2019)
[243] Oberthur, (2002)
[244] Miaihle (2018)