AI CHALLENGES FOR SOCIETY AND ETHICS

CHAPTER FOR THE OXFORD HANDBOOK OF AI GOVERNANCE

Jess Whittlestone[1,2] and Sam Clarke[1]

[1]Leverhulme Centre for the Future of Intelligence, University of Cambridge
[2]Centre for the Study of Existential Risk, University of Cambridge

ABSTRACT

Artificial intelligence is already being applied in and impacting many important sectors in society, including healthcare, finance, and policing. These applications will increase as AI capabilities continue to progress, which has the potential to be highly beneficial for society, or to cause serious harm. The role of AI governance is ultimately to take practical steps to mitigate this risk of harm while enabling the benefits of innovation in AI. This requires answering challenging empirical questions about current and potential risks and benefits of AI: assessing impacts that are often widely distributed and indirect, and making predictions about a highly uncertain future. It also requires thinking through the normative question of what beneficial use of AI in society looks like, which is equally challenging. Though different groups may agree on high-level principles that uses of AI should respect (e.g., privacy, fairness, and autonomy), challenges arise when putting these principles into practice. For example, it is straightforward to say that AI systems must protect individual privacy, but there is presumably some amount or type of privacy that most people would be willing to give up to develop life-saving medical treatments. Despite these challenges, research can and has made progress on these questions. The aim of this chapter will be to give readers an understanding of this progress, and of the challenges that remain.

Overview

1   Introduction

AI is already being applied in and impacting many important sectors in society, including healthcare, finance, and policing. As investment into AI research continues, we are likely to see substantial progress in AI capabilities and their potential applications, precipitating even greater societal impacts. The use of AI promises real benefits by helping us to better understand the world around us and develop new solutions to important problems, from disease to climate change. However, the power of AI systems also means that they risk causing serious harm if misused or deployed without careful consideration for their immediate and wider impacts.[1]

The role of AI governance is ultimately to take practical steps to mitigate this risk of harm while enabling the benefits of innovation in AI. To do this requires answering challenging empirical questions about the possible risks and benefits of AI, as well as challenging normative questions about what beneficial use of AI in society looks like.

To properly assess risks and benefits, we need a thorough understanding of how AI is already impacting society, and how those impacts are likely to evolve in future—which is far from straightforward. Assessing even *current* impacts of a technology like AI is challenging since these are likely to be widely and variably distributed across society. Furthermore, it is difficult to determine the extent to which impacts are caused by AI systems, as opposed to other technologies or societal changes. Assessing *potential* impacts of AI in the future—which is necessary if we are to intervene while impacts can still be shaped and harms have not yet occurred—is even more difficult, since it requires making predictions about an uncertain future.

The normative question of what beneficial use of AI in society looks like is also complex. A number of different groups and initiatives have attempted to articulate and agree on high-level principles that uses of AI should respect, such as privacy, fairness, and autonomy (Jobin et al., 2019). Though this is a useful first step, many challenges arise when putting these principles into practice. For example, it seems straightforward to say that the use of AI systems must protect individual privacy, but there is presumably some amount or type of privacy that most people would be willing to give up to develop life-saving medical treatments. Different groups and cultures will inevitably have different views on what trade-offs we should make, and there may be no obvious answer or clear way of adjudicating between views. We must therefore also find politically feasible ways to balance different perspectives and values in practice, and ways of making decisions about AI that will be viewed as legitimate by all.

Despite these challenges, research can and has made progress on understanding the impacts of AI, and on illuminating the challenging normative questions that these impacts raise. The aim of this chapter will be to give the reader an understanding of this progress, and the challenges that remain. We begin by outlining some of the benefits and opportunities AI promises for society, before turning to some of the most concerning sources of harm and risk AI might pose. We then discuss the kinds of ethical and political challenges that arise in trying to balance these benefits and risks, before concluding with some recommendations for AI governance today.

2   Benefits and opportunities

The promise of AI ultimately lies in its potential to help us understand the world and solve problems more effectively than humans could do alone. We discuss potential benefits of AI in three related categories: (1) improving the quality and length of people's lives; (2) improving our ability to tackle problems as a society; (3) enabling moral progress and cooperation.

---

[1]When we talk about AI systems in this chapter, we mean software systems which use machine learning (ML) techniques. ML involves learning from data to build mathematical models which can help us with a variety of real-world tasks, including predicting the likelihood a loan will be repaid based on someone's financial history, translating text between languages, or deciding what moves to take to win at a board game.

2.1 Improving the quality and length of people's lives

AI can help improve the quality and efficiency of public services and products by tailoring them to a given person or context. For example, several companies have begun to use AI to deliver personalised education resources (Hao, 2019), collecting data on students' learning and performance and using this to better understand learning patterns and specific learning needs (Luan & Tsai, 2021). Similarly, the use of AI to personalise healthcare through precision medicine—i.e. tailoring treatment based on specific features of an individual patient—is in early stages but shows real promise (K. B. Johnson et al., 2021; Xu et al., 2019), with startups beginning to emerge in this space (Toews, 2020).

AI is also showing promise to drastically improve our understanding of disease and medical treatments. AI systems can now outperform human specialists on a number of specific healthcare-related tasks: for example, Google Health trained a model to predict risk of breast cancer from mammograms, which outperformed human radiologists (McKinney et al., 2020). The use of AI to advance drug discovery, for instance by searching through and testing chemical compounds more quickly and effectively, is receiving increasing attention (Paul et al., 2021): the first clinical trial of an AI-designed drug began in Japan (Burki, 2020) and a number of startups in this space raised substantial funds in 2020 (Hogarth & Benaich, 2020). DeepMind's AI system AlphaFold has led to substantial progress on the "protein folding" problem,[2] with potential to drastically improve our ability to treat disease (Jumper et al., 2021). Continued progress in AI for healthcare might even contribute to better understanding and slowing processes of ageing (Zhavoronkov et al., 2019), resulting in much longer lifespans than we enjoy today.

2.2 Improving our ability to tackle problems as a society

AI could help tackle many of the big challenges we face as a society, such as climate change and threats to global health, by helping model the complex systems underpinning these problems, advancing the science behind potential solutions, and improving the effectiveness of policy interventions.

For instance, AI can support early warning systems for threats such as disease outbreaks: machine learning algorithms were used to characterise and predict the transmission patterns of both Zika (Jiang et al., 2018) and SARS-CoV-2 (Liu, 2020; Wu et al., 2020) outbreaks, supporting more timely planning and policymaking. With better data and more sophisticated systems in future it may be possible to identify and mitigate such outbreaks much earlier (Schwalbe & Wahl, 2020). There is also some early discussion of how AI could also be used to identify early signs of inequality and conflict: Musumba et al. (2021), for instance, use machine learning to predict the occurrence of civil conflict in Sub-Saharan Africa. This could make it much easier to intervene early to prevent conflict.

AI-based modelling of complex systems can improve resource management, which may be particularly important in mitigating the effects of climate change. For instance, AI is beginning to see application in predicting day-ahead electricity demand in the grid, improving efficiency, and in learning how to optimally allocate resources such as fleets of vehicles to address constantly changing demand (Hogarth & Benaich, 2019). Similarly, a better understanding of supply and demand in electricity grids can also help reduce reliance on high-polluting plants, and make it easier to proactively manage an increasing number of variable energy sources (Rolnick et al., 2019). Similar kinds of analysis could help with a range of other problems, including disaster response: for example, machine learning can be used to create maps from aerial imagery and retrieve information from social media to inform relief efforts (Rolnick et al., 2019).

AI also has potential to advance science in critical areas. There are many ways that AI could improve different aspects of the scientific process: by helping us to understand and visualise patterns in data of enormous volume and dimensionality (Mjolsness & DeCoste, 2001; Ourmazd, 2020); or by conducting more 'routine' aspects of scientific research such as literature search and summarisation, hypothesis generation, and experimental design and analysis (Gil et al., 2014). DeepMind's work on protein folding mentioned earlier is a good example of AI already advancing science in an important area. In the future, we could see AI accelerating progress in areas like materials science, by automating the time-consuming processes in the discovery of new materials, which could help develop better materials for storing or harnessing energy, for example (Rolnick et al., 2019).

As well as improving our understanding of problems and advancing the science needed to solve them, AI can help identify the most effective solutions that currently exist. There is evidence that ML tools can be used to improve policymaking by clarifying uncertainties in data, and improving existing tools for designing and assessing interventions (Rolnick et al., 2019). For instance, Andini et al. (2018) show that a simple ML algorithm could have been used to increase the effectiveness of a tax rebate program. It may even be possible to use AI to design more competent institutions which would help tackle many problems. One idea here is that (human) participants could determine

---

[2]This is the problem of predicting the 3D structure of a protein from its 2D genetic sequence.

desiderata that some institution should achieve, and leave the design of the institution to an AI system (Dafoe et al., 2020). This could allow novel approaches to old problems that humans cannot spot.

2.3 Enabling moral progress and cooperation

Most would agree that the world we live in today is a better place for most people than the world of centuries ago. This is partly due to economic and technological progress improving standards of living across the globe. But moral progress also plays an important role. Fewer people and animals experience suffering today, for example, because most people view an increasing proportion of sentient beings as worthy of care and moral concern. It has been suggested that AI could help accelerate moral progress (Boddington, 2021), for example by playing a "Socratic" role in helping us to reach better (moral) decisions ourselves (inspired by the role of deliberative exchange in Socratic philosophy as an aid to develop better moral judgements) (Lara & Deckers, 2020). Specifically, such systems could help with providing empirical support for different positions, improving conceptual clarity, understanding argumentative logic, and raising awareness of personal limitations.

AI might similarly help improve cooperation between groups, which arguably underlies humans' success in the world so far. Dafoe et al. (2020) outline a number of ways AI might support human cooperation: AI tools could help groups jointly learn about the world in ways that make it easier to find cooperative strategies, and more advanced machine translation could enable us to overcome practical barriers to increased international cooperation, including increased trade and possibly leading to a more borderless world. AI could also play an important role in building mechanisms to incentivise truthful information sharing, and explore the space of distributed institutions that promote desirable cooperative behaviours.

3 Harms and risks

Despite these many real and potential benefits, we are already beginning to see harms arise from the use of AI systems, which could become much more severe with more widespread application of increasingly capable systems.

In this section we'll discuss five different forms of harm AI might pose for individuals and society, in each case outlining current trends and impacts of AI pointing in this direction, and what we might be especially concerned about as AI systems increase in their capabilities and ubiquity across society.

3.1 Increasing the likelihood or severity of conflict

AI could impact the severity of conflict by enabling the development of new and more lethal weapons. Of particular concern are lethal autonomous weapons (LAWs): systems that can select and engage targets without further intervention by a human operator, which may recently have been used in combat for the first time (UN Security Council, 2021).[3] There is a strong case that "armed fully autonomous drone swarms", one type of lethal autonomous weapon, qualify as a weapon of mass destruction (WMD) (Kallenborn, 2020). This means they would pose all the threats that other WMDs do: geopolitical destabilisation, and use in acts of terror or catastrophic conflict between major powers. They would also be safer to transport and harder to detect than most other WMDs (Aguirre, 2020). Beyond LAWs, AI applied to scientific research or engineering could enable the development of other extremely powerful weapons. For example, it could be used to calculate the most dangerous genome sequences in order to create especially virulent biological viruses (O'Brien & Nelson, 2020; Turchin & Denkenberger, 2020).

Furthermore, we are seeing more integration of AI into defense and conflict domains, which could increase the likelihood of unintentional or rapid escalation in conflict: if more military decisions are automated, this makes it harder to intervene to prevent escalation (Deeks et al., 2018; J. Johnson, 2020). This is analogous to how algorithmic decision-making in financial systems led to the 2010 'flash crash': automated trading algorithms, operating without sufficient oversight, caused a trillion-dollar stock market crash over a period of approximately 36 minutes. The consequences could be even worse in a conflict scenario than in finance, because there is no overarching authority to enforce failsafe mechanisms (J. Johnson, 2020).

AI could also alter incentives in a way that makes conflict more likely to occur or to escalate (Zwetsloot & Dafoe, 2019). For example, AI could undermine second strike capabilities which are central to nuclear strategic stability, by

---

[3]According to the UN Security Council (2021) report, "Logistics convoys and retreating HAF [in Libya] were subsequently hunted down and remotely engaged by the unmanned combat aerial vehicles or the lethal autonomous weapons systems such as the STM Kargu-2 . . . programmed to attack targets without requiring data connectivity between the operator and the munition: in effect, a true "fire, forget and find" capability."

improving data collection and processing capabilities which would make it easier to discover and destroy previously secure nuclear launch facilities (Geist & Lohn, 2018; Lieber & Press, 2017).

3.2 Making society more vulnerable to attack or accident

As AI systems become more integral to the running of society this may create new vulnerabilities which can be exploited by bad actors. For instance, researchers managed to fool an ML model trained to recognise traffic signs into classifying a 'stop' sign as a 'yield' sign, simply by adding a small, imperceptible perturbation to the image (Papernot et al., 2017). An autonomous vehicle using this model could therefore be targeted by bad actors using stickers or paint to alter traffic signs. As AI systems become more widely deployed, these kinds of attacks could have more catastrophic consequences. For example, as AI is more widely integrated into diagnostic tools in hospitals or into our transport systems, adversarial attacks could put many lives at risk (Brundage et al., 2018; Finlayson et al., 2019).

Similarly, more widespread deployment of increasingly capable AI systems could also increase the severity of accidents. In particular, although the integration of AI into critical infrastructure has potential to bring efficiency benefits, it would also introduce the possibility of accidents on a far more consequential scale than is possible today. For example, as driverless cars become more ubiquitous, computer vision systems failing in extreme weather or road conditions could cause many cars to crash simultaneously. The direct casualties and second-order effects on road networks and supply chains could be severe. If and when AI systems become sufficiently capable to run large parts of society, these kinds of failures could possibly result in the malfunction of several critical systems at once, which at the extreme could put our very civilisation at risk of collapse.

One might think that these accidents could be avoided by making sure that a human either approves or makes the final decision. However, progress in AI capabilities such as deep reinforcement learning (DRL) could lead us to develop more autonomous systems, and there will likely be commercial pressure to deploy them. For such systems, especially when their decisions are too fast-moving or incomprehensible to humans, it is not clear how human oversight would work (Whittlestone et al., 2021).

These risks may be exacerbated by competitive dynamics in AI development. AI development is often framed in terms of a 'race' for strategic advantage and technological superiority between nations (Cave & ÓhÉigeartaigh, 2018). This framing is prominent in news sources, the tech sector and reports from governmental departments such as the U.S. Senate and Department of Defense (Imbrie et al., 2020). The more AI development is underpinned by these competitive dynamics, there may be a greater incentive for actors developing in AI to underinvest in the safety and security of their systems in order to stay ahead.

3.3 Increasing power concentration

Several related trends suggest AI may change the distribution of power across society, perhaps drastically. Absent major institutional reform, it seems plausible that the harms and benefits of AI will be very unequally distributed across society. AI systems are already having discriminatory impacts on marginalised groups: for example, facial recognition software has been shown to perform many times worse for darker faces (Raji & Buolamwini, 2019), and an AI system developed by Amazon to rank job candidates downgraded applications whose CVs included evidence they were female (West et al., 2019). Marginalised groups are less technologically literate on average, so are also more likely to be impacted by harms of AI such as the scaling up of misinformation and manipulative advertising (Lutz, 2019). These groups are also less likely to be in a financial position to benefit from advances in AI such as personalised healthcare (West et al., 2019).

At the same time, AI development is making already wealthy and powerful actors more so. The companies who already have the greatest market share have access to the most data, computing power, and research talent, enabling them to build the most effective products and services—increasing their market share further and making it easier for them to continue amassing data, compute, and talent (Dafoe, 2018; Kalluri, 2020; Lee, 2018). This creates a positive feedback loop cementing the powerful position these technology companies are already in. Similarly, wealthier countries able to invest more in AI development are likely to reap economic benefits more quickly than developing economies, potentially widening the gap between them. Especially if AI development leads to more rapid economic growth than previous technologies (Aghion et al., 2019), this might cause more extreme concentration of power than we have ever seen before.

In addition, AI-based automation has the potential to drastically increase income inequality. Progress in AI systems will inevitably make it possible to automate an increasing range of tasks. Progress in reinforcement learning specifically could improve the dexterity and flexibility of robotic systems (Ibarz et al., 2021), leading to increased automation of manual labour jobs with lower wages. The automation of these jobs will force those people to retrain; even in the

best case, they will face temporary disruptions to income (Lee, 2018). However, it is not just low-wage or manual labour jobs that are at risk. Advances in language modelling could spur rapid automation of a wide range of knowledge work, including aspects of journalism, creative writing, and programming (Tamkin et al., 2021). Many of these knowledge workers will flood the highly social and dextrous job market (which is hard to automate, but already has low wages), further increasing income inequality (Lee, 2018). There is also reason to think that changes in the availability of jobs due to AI may happen more quickly than previous waves of automation, due to the fact that algorithms are infinitely replicable and instantly distributable (unlike, for example, steam engines and even computers), and the emergence of highly effect venture capital funding driving innovation (Lee, 2018). This gives us less time to prepare, for example by retraining those whose jobs are most likely to be lost, and makes it more likely that the impacts on inequality will be more extreme than anything seen previously.

Developments in AI are also likely to give companies and governments more control over individuals' lives than ever before. The fact that current AI systems require large amounts of data to learn from creates incentives for companies to collect increasing amounts of personal data from users (though only certain applications such as medicine and advertising require highly personal data). Citizens are increasingly unable to consent to—or even be aware of—how their data is being used, while the collection of this data may increasingly be used by powerful actors to surveil, influence, and even manipulate and control populations. For example, the company ClearView AI scraped billions of images from Facebook, YouTube, Venmo and millions of other websites, using them to develop a "search engine for faces", which they then licensed, without public scrutiny, to over 600 law enforcement agencies (Hill, 2020). We are already seeing harmful uses of facial recognition, such as in their use to surveil Uighur and other minority populations in China (Hogarth & Benaich, 2019).[4] The simultaneous trends of apparently eroding privacy norms, and increased use of AI to monitor and influence populations, are seriously concerning.

Relatedly, AI has the potential to scale up the production of convincing yet false or misleading information online (e.g. via image, audio and text synthesis models like BigGAN and GPT-3), and to target that content at individuals and communities most likely to be receptive to it (e.g via automated A/B testing) (Seger et al., 2020). Whilst the negative impact of such techniques has so far been fairly contained, more advanced versions would make it easier for groups to seek and retain influence, for instance by influencing elections or enabling highly effective propaganda. For example, further advances in language modelling could be applied to design tools that "coach" their users to persuade other people of certain claims (Kokotajlo, 2020). Whilst these tools could be used for social good—e.g. New York Times' chatbot that helps users to persuade people to get vaccinated against Covid-19 (Gagneur & Tamerius, 2021)—they could equally be used by self-interested groups to gain or retain influence.

3.4 Undermining society's ability to solve problems

The use of AI in the production and dissemination of information online may also have broader negative impacts. In particular, it has been suggested that the use of AI to improve content recommendation engines by social media companies is contributing to worsened polarisation online (Faddoul et al., 2020; Ribeiro et al., 2019).[5]

Looking to the future, the use of AI in production or targeting of information could have substantial impacts on our information ecosystem. If advanced persuasion tools are used by many different groups to advance many different ideas, we could see the world splintering into isolated 'epistemic communities', with little room for dialogue or transfer between them. A similar scenario could emerge via the increasing personalisation of people's online experiences: we may see a continuation of the trend towards "filter bubbles" and "echo chambers", driven by content selection algorithms, that some argue is already happening (Barberá et al., 2015; Flaxman et al., 2016; Nguyen et al., 2014). In addition, increased awareness of these trends in information production and distribution could make it harder for anyone to evaluate the trustworthiness of any information source, reducing overall trust in information.

In all of these scenarios, it would be much harder for humanity to make good decisions on important issues, particularly due decreasing trust in credible multipartisan sources, which could hamper attempts at cooperation and collective action. The vaccine and mask hesitancy which exacerbated the negative impacts of Covid-19, for example, were likely the result of insufficient trust in public health advice (Seger, 2021). We could imagine an even more virulent pandemic, where actors exploit the opportunity to spread misinformation and disinformation to further their own ends. This could lead to dangerous practices, a significantly increased burden on health services, and much more catastrophic outcomes (Seger et al., 2020).

---

[4]CloudWalk Technology, a key supplier to the Chinese government, markets its "fire eye" facial recognition service to pick out "Uighurs, Tibetans and other sensitive groups"

[5]Note that this suggestion has been disputed (e.g. Boxell et al. 2017; Ledwich and Zaitsev 2019). The underlying methodological problem is that social media companies have sole access to the data required to perform a thorough analysis, and lack incentive to publicise this data or perform the analysis themselves.

3.5 Losing control of the future to AI systems

If AI systems continue to become more capable and begin running significant parts of the economy, we might also worry about humans losing control of important decisions. Currently, humans' attempts to shape the world are the only goal-directed process influencing the future. However, more automated decision-making would change this, and could result in some (or all) human control over the future being lost (Christiano, 2019; Critch, 2021; Ngo, 2020; Russell, 2019).

This concern relies on two assumptions. First, that AI systems will become capable enough that it will be not only possible but desirable to automate a majority of tasks making up the economy, from managing critical infrastructure to running corporations. Second, that despite our best efforts, we may not understand these systems well enough to be sure they are fully aligned with what their operators want.

How plausible are these assumptions? Considering the first, there is increasing reason to believe we might build AI systems as capable as humans across a broad range of economically useful tasks this century. Enormous amounts of resources are going into AI progress, and developing human-level AI is the stated goal of two very well-resourced organisations (DeepMind and OpenAI), as well as a decent proportion of AI researchers. In recent years, we have seen advances in AI defy expectations, especially in terms of their ability to solve tasks they weren't explicitly trained for, and the improvements in performance that can be derived from simply increasing the size of models, the datasets they are trained on, and the computational resources used for training them (Branwen, 2021).[6] For example, GPT-3 (the latest language model from OpenAI at the time of writing), shows remarkable performance on a range of tasks it was not explicitly trained on, such as generating working code from natural language descriptions, functioning as a chatbot in limited contexts, and being used as a creative prompt (Tamkin et al., 2021). These capabilities are quickly spurring a range of commercial applications, including GitHub Copilot, a tool that helps programmers work faster by suggesting lines of code or entire functions (Chen et al., 2021). This progress was achieved simply by scaling up previous language models to larger sizes and training them with more data and computational resources. There is good evidence that this trend will continue to result in more powerful systems without needing 'fundamental' breakthroughs in machine learning (Kaplan et al., 2020).

The second assumption, that advanced AI systems might not be fully aligned with or understandable to humans, is perhaps on even stronger ground. We currently train AI systems by "trial and error", in the sense that we search for a model that does well on some objective, without necessarily knowing how a given model produces the behaviour it does. This leaves us with limited assurance about how the system might behave in new contexts or environments. A particular concern is that AI systems might help us to optimise for what we can *measure* in society, but not what we actually value (Christiano, 2019). For example, we might deploy AI systems in law enforcement to help increase security and safety in communities, but later find that these systems are in fact increasing *reported* sense of safety by driving down complaints and hiding information about failures. If we don't notice these kinds of failures until AI systems are integral to the running of society, it may be very costly or even impossible to correct them. As mentioned earlier, competitive pressures to use AI for economic gain may make this more likely, driving actors to deploy AI systems without sufficient assurances that they are optimising for what we want.

This could happen gradually or suddenly, depending on the pace and shape of AI progress. The most high-profile versions of these concerns have focused on the possibility of a single misaligned AI system rapidly increasing in intelligence (Bostrom, 2014), but a much more gradual 'takeover' of society by AI systems may be more plausible, where humans don't quite realise they are losing control until society is almost entirely dependent on AI systems and it is difficult or impossible for humans to regain control over decision-making.

4 Ethical and political challenges

It is fairly uncontroversial to suggest that improving the length and quality of people's lives is something we should strive for, and that catastrophic accidents which claim thousands of lives should be avoided.

However, enabling the benefits of AI while mitigating the harms is not necessarily so straightforward. Sometimes what is needed to enable some area of benefit may also be the exact thing that carries risk.

For example, using AI to automate increasing amounts of the economy has potential to improve the quality of services and rapidly boost economic growth, which could result in drastic improvements to quality of life across the globe. However, the economic gains of this kind of progress, as well as the harms of job displacement, may be drastically

---

[6]A similar point is made by Sutton (2019): using general methods like search and learning (rather than specific methods than involve building human knowledge into AI systems) and applying a lot of computation to them, has and will continue to yield the biggest breakthroughs in AI.

unequal, leading to a concentration of power and rise in inequality across society never seen before. There are empirical questions here, about what processes are most likely to exacerbate inequality, that research could make progress on. There are also practical interventions that could be implemented in order to increase the likelihood that the economic gains of AI can be redistributed. However, there are still fundamental value judgements that must be made when envisioning what we want from the future of AI: how should we balance the potential for societal progress, and the possibility of huge gains in average quality of life, against the risk of radically increased inequality? If applying AI to science has potential to increase human health and lifespans, but also risks the creation of dangerous new technologies if not approached with care and wisdom, how much risk should we be willing to take? If outsourcing decisions to AI systems has potential to help us solve previously intractable societal problems, but at the cost of reduced human autonomy and understanding of the world, what should we choose?

Because these questions are normatively complex, there will be plenty of room for reasonable disagreement. Those who prioritise aggregate wellbeing will want to make different choices today to those who prioritise equality. Younger people may be happier to sacrifice privacy than older generations; those from countries which already have a strong welfare state will likely be more concerned about threats to equality; and values such as human autonomy may be perceived very differently in different cultures.

How do we deal with these disagreements? In part, this is the domain of AI ethics research, which can help to illuminate important considerations and clearly outline arguments for different perspectives. However, we should not necessarily expect ethics research to provide all the answers, especially on the timeframe in which we need to make decisions about how AI is developed and used. We can also provide opportunities for debate and resolution, but in most cases it will be impossible to resolve disagreements entirely and use AI in ways everyone agrees with.[7] We must therefore find ways to make choices about AI despite the existence of complex normative issues and disagreement on them.

Some political scientists and philosophers have suggested that where agreement on final decisions is impossible, we should instead focus our attention on ensuring the *process* by which a decision is made is legitimate (Patty & Penn, 2014). This focus on decision-making procedures as opposed to outcomes has also arisen in debates around public health ethics; Daniels and Sabin (2008) suggest that in order to be seen as legitimate, decision-making processes must be, among other things, open to public scrutiny, revision and appeal.[8]

We do not currently have legitimate procedures for making decisions about how we develop and use AI in society. Many important decisions are being made in technology companies whose decisions are not open to public or even government scrutiny, meaning they have little accountability for the impacts of their decisions on society. For instance, despite being among the world's most influential algorithms, Facebook's and YouTube's content selection algorithms are mostly opaque to those most impacted by them. The values and perspectives of individuals making important decisions have disproportionate influence over how "beneficial AI" is conceived of, while the perspectives of minority groups and less powerful nations have little influence.

5   Implications for governance

What should we be doing to try and ensure that AI is developed and used in beneficial ways, today and in the future? We suggest that AI governance today should have three broad aims.

Ultimately, AI governance should be focused on **identifying and implementing mechanisms which enable benefits and mitigate harms of AI**. However, as we've discussed throughout this chapter, in some cases doing this may not be straightforward, for two reasons. First, there are many actual and potential impacts of AI which we do not yet understand well enough to identify likely harms and benefits. Second, even where impacts are well understood, tensions may arise, raising challenging ethical questions on which people with different values may disagree. AI governance therefore also needs to develop methods and processes to address these barriers: **to improve our ability to assess and anticipate the impacts of AI**; and **to make decisions even in the face of normative uncertainty and disagreement**. We conclude this chapter by making some concrete recommendations for AI governance work in each of these three categories.

---

[7]A number of formal results from social choice theory demonstrate that when there are numerous different preferences and criteria relevant to a decision, only under strong assumptions can an unambiguously "best" option be found - i.e. in many real-life cases, no such resolution will be possible (Patty & Penn, 2014).

[8]Of course, there will also be room for reasonable disagreement about decision-making procedures, but we think there is likely to be less disagreement on this level, than on the level of object level decisions

5.1 Enabling benefits and mitigating harms

In some cases, **we might need to consider outright bans on specific applications of AI**, if the application is likely to cause a level or type of harm deemed unacceptable. For example, there has been substantial momentum behind campaigns to ban lethal autonomous weapons (LAWs),[3] and the European Commission's proposal for the first AI regulation includes a prohibition on the use of AI systems which engage in certain forms of manipulation, exploitation, indiscriminate surveillance, and social scoring (European Commission, 2021). Another area where prohibitions may be appropriate is in the integration of AI systems into nuclear command and control, which could increase the risk of accidental launch with catastrophic consequences, without proportional benefits (Ord et al., 2021).

However, effective bans on capabilities or applications can be challenging to enforce in practice. It can be difficult to achieve the widespread international agreement needed—for example, the US government have cited the fact that China is unlikely to prohibit LAWs as justification for not making the ban themselves (NSCAI, 2021). In other cases it may be difficult to delineate harmful applications clearly enough. In the case of the EU regulation, it is likely to be very difficult to clearly determine whether an AI system should be deemed as "manipulative" or "exploitative" in the ways stated, for example.

Where outright bans are infeasible, **it may be possible to limit access to powerful capabilities to reduce risk of misuse**. For example, companies might choose not to publish the full code behind specific capabilities to prevent malicious actors from being able to reproduce them, or limit access to commercial products with potential for misuse (Radford et al., 2019). However, this introduces a tension between the need for caution and the benefits of open sharing in promoting beneficial innovation (Whittlestone & Ovadya, 2020), which has prompted substantial debate and analysis around the role of publication norms in AI (Gupta et al., 2020). Governments might also consider monitoring and regulating access to large amounts of computing power, which would allow them oversight and control over which actors have access to more powerful AI systems (Brundage et al., 2018).

To go beyond preventing harms and realise the full benefits of AI, it will be crucial to **invest in both socially beneficial applications, and in AI safety and responsible AI research**. Many of the potential benefits we discussed early in this chapter seem relatively underexplored: the potential uses of AI to enhance cooperation between groups, to combat climate change, or improve moral reasoning, for example, could receive a great deal more attention. Part of the barrier to working on these topics is that they may not be well-incentivised by either academia (which often rewards theoretical progress over applications) or industry (where economic incentives are not always aligned with broad societal benefit). Similarly, AI safety and responsible AI research will be crucial for ensuring even the most beneficial applications of AI do not come with unintended harms. One concrete idea would be for governments to create a fund of computational resources which is available free of charge for projects in these areas (Brundage et al., 2020).

5.2 Improving our ability to assess and anticipate impacts

In many cases we may first need to better understand the potential impacts of AI systems before determining what kinds of governance are needed. Better and more standardised processes for impact assessment would be valuable on multiple levels.

First, we need to **establish clearer standards and methods for assuring AI systems** (also sometimes called test, evaluation, validation and verification—TEVV—methods) before they go to market, particularly in safety-critical contexts. There are currently no proven effective methods for assuring the behaviour of most AI systems, so much more work is needed (Flournoy et al., 2020). It is likely that rather than a single approach to assuring AI systems, an ecosystem of approaches will be needed, depending on the type of AI system, and the decision to be made (Ahamat et al., 2021). Better assurance processes would make it easier to decide where the use of AI systems should be restricted, by requiring uses of AI to pass certain established standards. It would also make it possible to identify and mitigate potential harms from unintended behaviour in advance, and to incentivise technical progress to make systems more robust and predictable.

Continual monitoring and stress-testing of systems will also be important, given it may not be possible to anticipate all possible failure modes or sources of attack in advance of deployment. Here it may be useful to build on approaches to 'read-teaming' in other fields including information and cyber security (Brundage et al., 2018).

We also need **broader ways to assess and anticipate the structural impacts of AI systems**. Assurance and stress-testing can help to identify where unintended behaviours or attacks on AI systems might cause harm, but cannot identify where a system behaving as intended might nonetheless cause broader structural harms (for example, polarising online discourse or changing incentives to make conflict more likely) (Zwetsloot & Dafoe, 2019). This will likely require looking beyond existing impact assessment frameworks and drawing on broader perspectives and methodologies, including: social science and history, fields which study how large societal impacts may come about without

anyone intending them (Zwetsloot & Dafoe, 2019); foresight processes for considering the future evolution of impacts (Government Office for Science, 2017); and participatory processes to enable a wider range of people to communicate harms and concerns (Smith et al., 2019).

More systematic monitoring of AI progress would improve our ability to anticipate and prepare for new challenges before they arise (Whittlestone et al., 2021). As technologies advance and more AI systems are introduced into the market, they will raise increasingly high-stakes policy challenges, making it increasingly important that governments have the capacity to react quickly. AI as a sector is naturally producing a wide range of data, metrics and measures that could be integrated into an 'early warning system' for new capabilities and applications which may have substantial impacts on society. Monitoring progress on widely studied benchmarks and assessment regimes in AI could enable AI governance communities to identify areas where new or more advanced applications of AI may be forthcoming. Monitoring *inputs* into AI progress, such as computational costs, data, and funding, may also help to give a fuller picture of where societally-relevant progress is most likely to emerge (Martínez-Plumed et al., 2018). For example, early warning signs of recent progress in language models could have been identified via a combination of monitoring progress on key benchmarks in language modelling, and monitoring the large jumps in computational resources being used to train these models.

5.3 Making decisions under uncertainty and disagreement

Even with better methods for assessing and anticipating the impacts of AI systems, challenges will remain: there will be uncertainties about the future impacts of AI that cannot be reduced, and conflicting perspectives on how we *should* be using AI for global benefit that cannot be easily resolved. AI governance will therefore need to grapple with what *processes* for making decisions about AI should look like, given this uncertainty and disagreement.

Greater use of participatory processes in decision-making around AI governance could help with ensuring the legitimacy and public acceptability of decisions, and may also improve the quality of the decisions themselves. There is evidence that participatory approaches used in the domain of climate policy lead to both increased engagement and understanding of decisions, and to better decisions (Hügel & Davies, 2020). Various projects have begun to engage a wider variety of perspectives in thinking through governance and societal issues related to AI (Balaram et al., 2018; Ipsos MORI, 2017), but much more could be done, especially in terms of integrating these processes into policymaking. We would also like to see participatory studies focused on concerns and hopes about the *future* of AI rather than just current AI systems, since these are more likely to be timely and relevant enough to influence decision-making. Public engagement is of course only one kind of input into decision-making processes, and must be combined with relevant expert analysis. However, participatory processes can be especially useful for understanding the wider impacts of policies which might be neglected by decision-makers, and for highlighting additional considerations or priorities, and policymaking around AI would benefit from giving them greater attention.

More generally, we need to think about how **processes for making important decisions about AI can be sufficiently open to scrutiny and challenge**. This is particularly difficult given that some of the most important decisions about the future of AI are being made within technology companies, which are not subject to the same forms of accountability or transparency requirements as governments. Some greater scrutiny may be achieved through regulation requiring greater transparency from companies. It may also be possible to improve transparency and accountability through shifts in norms—if there is enough public pressure, companies may have an incentive to be more transparent—or by improving the capacity of government to monitor company behaviour, such as by increasing technical expertise in government and establishing stronger measurement and monitoring infrastructure.

6 Conclusion

In this chapter, we have outlined some of the possible ways AI could impact society into the future, both beneficial and harmful. Our aim has not been to predict the future, but to demonstrate that the possible impacts are wide-ranging, and that there are things we can do today to shape them. As well as intervening to enable specific benefits and mitigate harms, AI governance must develop more robust methods to assess and anticipate the impacts of AI, and better processes for making decisions about AI under uncertainty and disagreement.

Bibliography

Aghion, P., Jones, B. F., & Jones, C. I. (2019). Artificial intelligence and economic growth. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 237–282). The University of Chicago Press. Retrieved August 16, 2021, from http://www.nber.org/chapters/c14015

Aguirre, A. (2020). *Why those who care about catastrophic and existential risk should care about autonomous weapons* [EA forum]. Retrieved July 22, 2021, from https : / / forum . effectivealtruism . org / posts / oR9tLNRSAep293rr5/why-those-who-care-about-catastrophic-and-existential-risk-2

Ahamat, G., Chang, M., & Thomas, C. (2021). *Types of assurance in AI and the role of standards* [Centre for data ethics and innovation blog]. Retrieved July 22, 2021, from https://cdei.blog.gov.uk/2021/04/17/134/

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., & Salvestrini, V. (2018). Targeting with machine learning: An application to a tax rebate program in italy. *Journal of Economic Behavior & Organization*, *156*, 86–102. https://doi.org/10.1016/j.jebo.2018.09.010

Balaram, B., Greenham, T., & Leonard, J. (2018). *Artificial intelligence: Real public engagement*. Retrieved July 30, 2021, from https://www.thersa.org/reports/artificial-intelligence-real-public-engagement

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Boddington, P. (2021). AI and moral thinking: How can we live well with machines to enhance our moral agency? *AI and Ethics*, *1*(2), 109–111. https://doi.org/10.1007/s43681-020-00017-0

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies* (First edition) [OCLC: ocn881706835]. Oxford University Press.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences*, *114*(40), 10612–10617.

Branwen, G. (2021). The scaling hypothesis. Retrieved July 22, 2021, from https://www.gwern.net/Scaling-hypothesis

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., . . . Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv:1802.07228 [cs]*. Retrieved April 13, 2021, from http://arxiv.org/abs/1802.07228

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., . . . Anderljung, M. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv:2004.07213 [cs]*. Retrieved July 27, 2021, from http://arxiv.org/abs/2004.07213

Burki, T. (2020). A new paradigm for drug development. *The Lancet Digital Health*, *2*(5), e226–e227. https://doi.org/10.1016/S2589-7500(20)30088-1

Cave, S., & ÓhÉigeartaigh, S. S. (2018). An AI race for strategic advantage: Rhetoric and risks. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 36–40. https://doi.org/10.1145/3278721.3278780

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., . . . Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv:2107.03374 [cs]*. Retrieved July 28, 2021, from http://arxiv.org/abs/2107.03374

Christiano, P. (2019). *What failure looks like* [AI alignment forum]. Retrieved April 15, 2021, from https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like

Critch, A. (2021). *What multipolar failure looks like, and robust agent-agnostic processes (RAAPs)* [AI alignment forum]. Retrieved April 15, 2021, from https://www.alignmentforum.org/posts/LpM3EAakwYdS6aRKf/what-multipolar-failure-looks-like-and-robust-agent-agnostic

Dafoe, A. (2018). *AI governance: A research agenda*. Centre for the Governance of AI. Retrieved August 16, 2021, from http://www.fhi.ox.ac.uk/govaiagenda

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. (2020). Open problems in cooperative AI. *arXiv:2012.08630 [cs]*. Retrieved April 13, 2021, from http://arxiv.org/abs/2012.08630

Daniels, N., & Sabin, J. E. (2008). Accountability for reasonableness: An update. *BMJ (Clinical research ed.)*, *337*, a1850. https://doi.org/10.1136/bmj.a1850

Deeks, A., Lubell, N., & Murray, D. (2018, November 16). *Machine learning, artificial intelligence, and the use of force by states* (SSRN Scholarly Paper ID 3285879). Social Science Research Network. Rochester, NY. Retrieved April 13, 2021, from https://papers.ssrn.com/abstract=3285879

European Commission. (2021). *Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative*

acts. Retrieved July 23, 2021, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Faddoul, M., Chaslot, G., & Farid, H. (2020). A longitudinal analysis of YouTube's promotion of conspiracy videos. *arXiv:2003.03318 [cs]*. Retrieved July 22, 2021, from http://arxiv.org/abs/2003.03318

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*, 298–320. https://doi.org/10.1093/poq/nfw006

Flournoy, M. A., Haines, A., & Chefitz, G. (2020). *Building trust through testing*. Retrieved August 16, 2021, from Building%20Trust%20through%20Testing

Gagneur, A., & Tamerius, K. (2021). Opinion — your friend doesn't want the vaccine. what do you say? *The New York Times*. Retrieved July 29, 2021, from https://www.nytimes.com/interactive/2021/05/20/opinion/covid-19-vaccine-chatbot.html

Geist, E., & Lohn, A. J. (2018). How might artificial intelligence affect the risk of nuclear war? Retrieved April 13, 2021, from https://www.rand.org/pubs/perspectives/PE296.html

Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. *Science*, *346*(6206), 171–172. https://doi.org/10.1126/science.1259439

Government Office for Science. (2017). *Government office for science annual report: 2016 to 2017*. Retrieved July 30, 2021, from https://www.gov.uk/government/publications/government-office-for-science-annual-report-2016-to-2017

Gupta, A., Lanteigne, C., & Heath, V. (2020). Report prepared by the montreal ai ethics institute (maiei) on publication norms for responsible ai. *arXiv:2009.07262 [cs]*. Retrieved April 13, 2021, from http://arxiv.org/abs/2009.07262

Hao, K. (2019). *China has started a grand experiment in AI education. it could reshape how the world learns.* [MIT technology review]. Retrieved April 13, 2021, from https://www.technologyreview.com/2019/08/02/131198/china-squirrel-has-started-a-grand-experiment-in-ai-education-it-could-reshape-how-the/

Hill, K. (2020). The secretive company that might end privacy as we know it. *The New York Times*. Retrieved April 13, 2021, from https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

Hogarth, I., & Benaich, N. (2019). *State of AI report 2019*. Retrieved April 13, 2021, from https://www.stateof.ai/2019

Hogarth, I., & Benaich, N. (2020). *State of AI report 2020*. Retrieved April 13, 2021, from https://www.stateof.ai/2020

Hügel, S., & Davies, A. R. (2020). Public participation, engagement, and climate change adaptation: A review of the research literature. *WIREs Climate Change*, *11*(4). https://doi.org/10.1002/wcc.645

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S. (2021). How to train your robot with deep reinforcement learning; lessons we've learned. *arXiv:2102.02915 [cs]*. https://doi.org/10.1177/0278364920987859

Imbrie, A., Dunham, J., Gelles, R., & Aiken, C. (2020). *Mainframes: A provisional analysis of rhetorical frames in AI*. Center for Security and Emerging Technology. Retrieved April 13, 2021, from https://cset.georgetown.edu/research/mainframes-a-provisional-analysis-of-rhetorical-frames-in-ai/

Ipsos MORI. (2017). *Public views of machine learning: Findings from public research and engagement conducted on behalf of the royal society*. Royal Society. London, United Kingdom. Retrieved August 16, 2021, from https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

Jiang, D., Hao, M., Ding, F., Fu, J., & Li, M. (2018). Mapping the transmission risk of zika virus using machine learning models. *Acta Tropica*, *185*, 391–399. https://doi.org/10.1016/j.actatropica.2018.06.021

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, J. (2020). Artificial intelligence in nuclear warfare: A perfect storm of instability? *The Washington Quarterly*, *43*(2), 197–211. https://doi.org/10.1080/0163660X.2020.1770968

Johnson, K. B., Wei, W.-Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., & Snowdon, J. L. (2021). Precision medicine, AI, and the future of personalized health care. *Clinical and Translational Science*, *14*(1), 86–93. https://doi.org/https://doi.org/10.1111/cts.12884

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 1–11. https://doi.org/10.1038/s41586-021-03819-2

Kallenborn, Z. (2020). *Are drone swarms weapons of mass destruction?* U.S. Air Force Center for Strategic Deterrence Studies. Retrieved August 16, 2021, from https://media.defense.gov/2020/Jun/29/2002331131/-1/-1/0/60DRONESWARMS-MONOGRAPH.PDF

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, *583*(7815), 169–169. https://doi.org/10.1038/d41586-020-02003-2

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv:2001.08361 [cs, stat]*. Retrieved July 29, 2021, from http://arxiv.org/abs/2001.08361

Kokotajlo, D. (2020). *Persuasion tools: AI takeover without AGI or agency?* [AI alignment forum]. Retrieved July 29, 2021, from https://www.alignmentforum.org/posts/qKvn7rxP2mzJbKfcA/persuasion-tools-ai-takeover-without-agi-or-agency

Lara, F., & Deckers, J. (2020). Artificial intelligence as a socratic assistant for moral enhancement. *Neuroethics*, *13*(3), 275–287. https://doi.org/10.1007/s12152-019-09401-y

Ledwich, M., & Zaitsev, A. (2019). Algorithmic extremism: Examining youtube's rabbit hole of radicalization. *arXiv:1912.11211 [cs]*. Retrieved July 22, 2021, from http://arxiv.org/abs/1912.11211

Lee, K.-F. (2018). *AI superpowers: China, silicon valley, and the new world order*. Houghton Mifflin Harcourt.

Lieber, K. A., & Press, D. G. (2017). The new era of counterforce: Technological change and the future of nuclear deterrence. *International Security*, *41*(4), 9–49. https://doi.org/10.1162/ISEC_a_00273

Liu, J. (2020, April 2). *Deployment of health it in china's fight against the covid-19 pandemic* [Imaging technology news]. Retrieved April 13, 2021, from https://www.itnonline.com/article/deployment-health-it-china%E2%80%99s-fight-against-covid-19-pandemic

Luan, H., & Tsai, C.-C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society*, *24*(1), 250–266. Retrieved April 13, 2021, from https://www.jstor.org/stable/26977871

Lutz, C. (2019). Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies*, *1*(2), 141–148. https://doi.org/https://doi.org/10.1002/hbe2.140

Martínez-Plumed, F., Avin, S., Brundage, M., Dafoe, A., hÉigeartaigh, S. Ó., & Hernández-Orallo, J. (2018). Accounting for the neglected dimensions of AI progress. *arXiv:1806.00610 [cs]*. Retrieved July 23, 2021, from http://arxiv.org/abs/1806.00610

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., . . . Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mjolsness, E., & DeCoste, D. (2001). Machine learning for science: State of the art and future prospects. *Science*, *293*(5537), 2051–2055. https://doi.org/10.1126/science.293.5537.2051

Musumba, M., Fatema, N., & Kibriya, S. (2021). Prevention is better than cure: Machine learning approach to conflict prediction in sub-saharan africa. *Sustainability*, *13*(13), 7366. https://doi.org/10.3390/su13137366

Ngo, R. (2020). *AGI safety from first principles* [AI alignment forum]. Retrieved July 29, 2021, from https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. *Proceedings of the 23rd international conference on World wide web - WWW '14*, 677–686. https://doi.org/10.1145/2566486.2568012

NSCAI. (2021). *2021 final report*. Retrieved April 13, 2021, from https://www.nscai.gov/2021-final-report/

O'Brien, J. T., & Nelson, C. (2020). Assessing the risks posed by the convergence of artificial intelligence and biotechnology. *Health Security*, *18*(3), 219–227. https://doi.org/10.1089/hs.2019.0122

Ord, T., Mercer, A., & Dannreuther, S. (2021). *Future proof: The opportunity to transform the UK's resilience to extreme risks*. Retrieved August 16, 2021, from https://www.longtermresilience.org/futureproof

Ourmazd, A. (2020). Science in the age of machine learning. *Nature Reviews Physics*, *2*(7), 342–343. https://doi.org/10.1038/s42254-020-0191-7

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. *arXiv:1602.02697 [cs]*. Retrieved July 23, 2021, from http://arxiv.org/abs/1602.02697

Patty, J. W., & Penn, E. M. (2014). *Social choice and legitimacy: The possibilities of impossibility*. Cambridge University Press. https://doi.org/10.1017/CBO9781139030885

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, *26*(1), 80–93. https://doi.org/10.1016/j.drudis.2020.10.010

Radford, A., Wu, J., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019, February 14). *Better language models and their implications* [OpenAI]. Retrieved April 13, 2021, from https://openai.com/blog/better-language-models/

Raji, I. D., & Buolamwini, J. (2019). *Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products* [MIT media lab]. Retrieved April 13, 2021, from https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products/

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2019). Auditing radicalization pathways on YouTube. *arXiv:1908.08313 [cs]*. Retrieved July 22, 2021, from http://arxiv.org/abs/1908.08313

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., . . . Bengio, Y. (2019). Tackling climate change with machine learning. *arXiv:1906.05433 [cs, stat]*. Retrieved April 12, 2021, from http://arxiv.org/abs/1906.05433

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Schwalbe, N., & Wahl, B. (2020). Artificial intelligence and the future of global health. *The Lancet*, *395*(10236), 1579–1586. https://doi.org/10.1016/S0140-6736(20)30226-9

Seger, E. (2021). *The greatest security threat of the post-truth age* [BBC future]. Retrieved July 22, 2021, from https://www.bbc.com/future/article/20210209-the-greatest-security-threat-of-the-post-truth-age

Seger, E., Avin, S., Pearson, G., Briers, M., Ó hÉigeartaigh, S., & Helena, B. (2020). *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world*. Retrieved April 13, 2021, from https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf

Smith, L., Peach, K., Ramos, J., & Sweeney, J. A. (2019). *Our futures: By the people, for the people*. Retrieved April 13, 2021, from https://www.nesta.org.uk/report/our-futures-people-people/

Sutton, R. (2019). *The bitter lesson* [Incomplete ideas]. Retrieved July 22, 2021, from http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv:2102.02503 [cs]*. Retrieved April 13, 2021, from http://arxiv.org/abs/2102.02503

Toews, R. (2020). *These are the startups applying ai to transform healthcare* [Forbes]. Retrieved April 13, 2021, from https://www.forbes.com/sites/robtoews/2020/08/26/ai-will-revolutionize-healthcare-the-transformation-has-already-begun/

Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & SOCIETY*, *35*(1), 147–163. https://doi.org/10.1007/s00146-018-0845-5

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminating systems: Gender, race and power in AI*. AI Now Institute. Retrieved August 16, 2021, from https://ainowinstitute.org/discriminatingsystems.pdf

Whittlestone, J., Arulkumaran, K., & Crosby, M. (2021). The societal implications of deep reinforcement learning. *Journal of Artificial Intelligence Research*, *70*. https://doi.org/10.1613/jair.1.12360

Whittlestone, J., & Ovadya, A. (2020). The tension between openness and prudence in AI research. *arXiv:1910.01170 [cs]*. Retrieved April 13, 2021, from http://arxiv.org/abs/1910.01170

Wu, J., Leung, K., & Leung, G. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in wuhan, china: A modelling study. *The Lancet*, *395*(10225), 689–697. https://doi.org/10.1016/S0140-6736(20)30260-9

Xu, J., Yang, P., Xue, S., Sharma, B., Sanchez-Martin, M., Wang, F., Beaty, K. A., Dehan, E., & Parikh, B. (2019). Translating cancer genomics into precision medicine with artificial intelligence: Applications, challenges and future perspectives. *Human Genetics*, *138*(2), 109–124. https://doi.org/10.1007/s00439-019-01970-5

Zhavoronkov, A., Mamoshina, P., Vanhaelen, Q., Scheibye-Knudsen, M., Moskalev, A., & Aliper, A. (2019). Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Research Reviews*, *49*, 49–66. https://doi.org/10.1016/j.arr.2018.11.003

Zwetsloot, R., & Dafoe, A. (2019, February 11). *Thinking about risks from AI: Accidents, misuse and structure* [Lawfare]. Retrieved April 13, 2021, from https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure