Philosophy & Technology (2020) 33:571–593
https://doi.org/10.1007/s13347-020-00402-x

RESEARCH ARTICLE

Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance

Seán S. ÓhÉigeartaigh[1,2] · Jess Whittlestone[1] · Yang Liu[1,2] · Yi Zeng[2,3] · Zhe Liu[2,4]

Received: 3 January 2020 / Accepted: 22 April 2020 / Published online: 15 May 2020

Abstract

Achieving the global benefits of artificial intelligence (AI) will require international cooperation on many areas of governance and ethical standards, while allowing for diverse cultural perspectives and priorities. There are many barriers to achieving this at present, including mistrust between cultures, and more practical challenges of coordinating across different locations. This paper focuses particularly on barriers to cooperation between Europe and North America on the one hand and East Asia on the other, as regions which currently have an outsized impact on the development of AI ethics and governance. We suggest that there is reason to be optimistic about achieving greater cross-cultural cooperation on AI ethics and governance. We argue that misunderstandings between cultures and regions play a more important role in undermining cross-cultural trust, relative to fundamental disagreements, than is often supposed. Even where fundamental differences exist, these may not necessarily prevent productive cross-cultural cooperation, for two reasons: (1) cooperation does not require achieving agreement on principles and standards for all areas of AI; and (2) it is sometimes possible to reach agreement on practical issues despite disagreement on more abstract values or principles. We believe that academia has a key role to play in promoting cross-cultural cooperation on AI ethics and governance, by building greater mutual understanding, and clarifying where different forms of agreement will be both necessary and possible. We make a number of recommendations for practical steps and initiatives, including translation and multilingual publication of key documents, researcher exchange programmes, and development of research agendas on cross-cultural topics.

Keywords Artificial intelligence · AI ethics · AI governance · Cross-cultural cooperation

Seán S. ÓhÉigeartaigh and Jess Whittlestone contributed equally to this work.

✉ Seán S. ÓhÉigeartaigh
so348@cam.ac.uk

Extended author information available on the last page of the article

1 Introduction

Artificial intelligence has been identified as a key suite of technologies for many countries worldwide, in large part motivated by its general-purpose nature (Brynjolfsson and McAfee 2014). AI technologies, and machine learning techniques in particular, are being fruitfully applied to a vast range of domains, including language translation, scientific research, education, logistics, transport and many others. It is clear that AI will affect economies, societies and cultures profoundly at a national, international and global level. This has resulted in increasing attention being paid to both AI ethics: questions about how we should develop and deploy AI systems, given their potential impact on wellbeing and other deeply held values such as autonomy or dignity; and to AI governance: the more practical challenge of ensuring the ethical use of AI in society, be that through regulation, governance frameworks, or 'softer' approaches such as standards and ethical guidelines.[1]

Cross-cultural cooperation will be essential for the success of these ethics and governance initiatives. By 'cross-cultural cooperation', we mean groups from different cultures and nations working together on ensuring that AI is developed, deployed, and governed in societally beneficial ways. In this paper, we focus in particular on cooperation extending across national boundaries. Examples include (but are not limited to) the following: AI researchers from different countries collaborating on projects to develop systems in safe and responsible ways; establishing networks to ensure that diverse global perspectives can feed equally into international discussions about the ethical issues raised by AI; and involving a range of global stakeholders in the development of practical principles, standards, and regulation. By encouraging cross-cultural cooperation, we do not necessarily mean that all parts of the world should be subject to the same norms, standards, and regulation relating to AI, or that international agreement is always needed. Identifying which issues will need global standards or agreements, and where more cultural variation is needed, is itself a key challenge that will require cooperation to address.

Cross-cultural cooperation is important for several reasons. First, cooperation will be essential if AI is to bring about broad benefits across societies globally, enabling advances in one part of the world to be shared with other countries, and ensuring that no part of society is neglected or disproportionately negatively impacted by AI. Second, cooperation enables researchers around the world to share expertise, resources, and best practices. This enables faster progress both on beneficial AI applications, and on managing the ethical and safety-critical issues that may arise. Third, in the absence of cooperation, there is a risk that competitive pressures between states or commercial ecosystems may lead to underinvestment in safe, ethical, and socially beneficial AI development (Askell et al. 2019; Ying 2019). Finally, international cooperation is also important for more practical reasons, to ensure that applications of AI that are set to cross-national and regional boundaries (such as those used in major search engines or

[1] AI ethics and governance are closely related: governance proposals are often a practical response to recognized ethical issues, and ethical frameworks can be an important starting point for developing policy and regulation. Especially as AI ethics becomes more practical—developing principles and guidelines around the ethical use of AI in society—it starts to blend into some of the 'softer' governance approaches. Throughout this chapter, we will therefore often use 'AI ethics and governance' as a broad term to refer to the whole process of identifying ethical issues, codifying them, and implementing them through governance measures.

autonomous vehicles) can interact successfully with a sufficient range of different regulatory environments and other technologies in different regions (Cihon 2019).

Drawing on the insights of a group of leading scholars from East Asia and Europe,[2] we analyse current barriers to cross-cultural cooperation on AI ethics and governance, and how they might be overcome. We focus on cooperation between Europe and North America on the one hand and East Asia on the other. These regions are currently playing an outsized role in influencing the global conversation on AI ethics and governance (Jobin et al. 2019), and much has been written recently about competition and tensions between nations in these regions in the domains of AI development and governance, especially in the case of China and the USA. However, our discussion and recommendations have implications for broader international cooperation around AI, and we hope they will spur more attention to promoting cooperation across a wider range of regions.

As AI systems become more capable and their applications more impactful and ubiquitous, the stakes will only get higher. Establishing cooperation over time may become more difficult, especially if cross-cultural misunderstandings and mistrust become entrenched in intellectual and public discussions. If this is the case, then the earlier a global culture of cooperation can be established, the better. Cultivating a shared understanding and deep cooperative relationships around guiding the impacts of AI should therefore be seen as an immediate and pressing challenge for global society.

2 The Role of North America, Europe, and East Asia in Shaping the Global AI Conversation

North America, Europe, and East Asia in particular are investing heavily in both fundamental and applied AI research and development (Benaich and Hogarth 2019; Haynes and Gbedemah 2019; Perrault et al. 2019), supported by corporate and government investment. Various analyses have framed progress on the development and deployment of AI between the USA and China in particular through a competitive lens (Simonite 2017; Allen and Husain 2017; Stewart 2017), though this framing has received criticism both on normative and descriptive grounds (Cave and ÓhÉigeartaigh 2018).

Scholars and policy communities in these regions are also taking deliberate and active steps to shape the development of ethical principles and governance recommendations for AI, both at a regional and global level. This is reflected in government-linked initiatives, such as the activities of the European Union's High-Level Expert Group on Artificial Intelligence, which has produced ethics guidelines and policy and

[2] We convened a workshop on cross-cultural trust in July 2019 (https://www.eastwest.ai/), with representatives from UK universities (Cambridge, Bath) as well as from Chinese and Japanese universities and initiatives (Universities of Hong Kong, Peking, Fudan, Keio, and the Berggruen Institute China Center and the Chinese Academy of Sciences). These representatives have been heavily involved in AI ethics and governance conversations and collaborative projects across Europe, North America, and Asia. This paper draws on insights from this workshop, which focused particularly on the role of academia in building cross-cultural trust in AI. We do not suggest that the regions represented are the only ones that matter in this conversation, nor that the members present can be considered to represent the diversity of views and expertise in the regions they are based in. However, we feel that this small step in establishing a network was of value, and generated useful insights and ground to build on. The workshop was held under Chatham House Rule.

investment recommendations as its first two publications[3]; the UK Government's commitments to 'work closely with international partners to build a common understanding of how to ensure the safe, ethical and innovative deployment of Artificial Intelligence' (May 2018); and the Chinese Government's similar commitment to 'actively participate in global governance of AI, strengthen the study of major international common problems such as robot alienation and safety supervision, deepen international cooperation on AI laws and regulations, international rules' (China State Council 2017). North America, Europe, and East Asia are each also contributing disproportionately towards international AI standards work ongoing in fora such as the International Organization for Standardization (ISO),[4] the Institute of Electrical and Electronics Engineers (IEEE),[5] and the Organization for Economic Co-operation and Development (OECD).[6]

The prominence of North America, Europe, and East Asia is further seen in the leadership and composition of multi-stakeholder and nongovernmental initiatives such as the Partnership on AI,[7] the Future Society,[8] the International Congress for the Governance of AI,[9] and the Global Partnership on AI (Hudson 2019).[10]A large majority of the most prominent conferences on AI ethics and governance have taken place in these regions, including the US-based Beneficial AI conference series,[11] the Beijing Academy of AI Conference series,[12] the Beijing Forum,[13] the US-based Artificial Intelligence, Ethics, and Society conferences,[14] governance and ethics workshops attached to the leading machine learning conferences, and many more.

This combination of:

a. technological leadership in North America, Europe, and East Asia;
b. the outsized role of these regions in shaping the global ethics and governance conversation; and
c. the underlying tension introduced by progress being framed through a competitive lens, and a perception of disagreements on fundamental ethical and governance issues

leads us to focus on the barriers that exist to productive intellectual exchange and cooperation between these regions and cultures in particular. A full analysis of cross-

[3] Available here: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
[4] North American, European, and East Asian countries account for 22 out of 28 participating countries in the AI standards committee (https://www.iso.org/committee/6794475.html)
[5] E.g. see committee membership of the IEEE's *Ethically Aligned Design* AI principles document: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf
[6] E.g. see participant membership of the OECD Expert Group on AI (AIGO) https://www.oecd.org/going-digital/ai/oecd-aigo-membership-list.pdf
[7] https://www.partnershiponai.org/partners/
[8] https://thefuturesociety.org/our-team/
[9] https://icgai.org/icgai-members/
[10] Formerly the International Panel on AI, initiated by Canada and France (not to be confused with the Partnership on AI) (https://www.canada.ca/en/innovation-science-economic-development/news/2019/05/declaration-of-the-international-panel-on-artificial-intelligence.html)
[11] https://futureoflife.org/beneficial-agi-2019/
[12] https://mp.weixin.qq.com/s/tAGOoqqA6ods9uaigWE7uA
[13] http://newsen.pku.edu.cn/news_events/news/global/9133.htm
[14] https://www.aies-conference.com/2020/

cultural cooperation on AI ethics and governance, which is outside the scope of this paper, must consider the roles of all nations and cultures, given that the domain of impact of AI technologies is truly global. It will be particularly important for further work to address the inequalities in power and influence that are emerging between technology-leading nations and those to which these technologies are being exported (Lee 2017), and the responsibility of technology-leading nations to include and empower those nations in global governance and ethics conversations.

3 Barriers to Cross-cultural Cooperation on AI

Despite the emergence of several international alliances in AI ethics and governance, many barriers remain to achieving real cross-cultural cooperation on the norms, principles, and governance frameworks that should guide how AI is developed and used.

Mistrust between different regions and cultures is one of the biggest barriers to international cooperation in AI ethics and governance. At present, there is a particular environment of mistrust between scholars, technologists, and policymakers in the USA and China.[15] This culture of mistrust is underpinned by both:

a. a history of political tensions between these two powerful regions that has increased significantly in recent years, and is currently contributing to the competitive framing of AI development as a 'race' between 'Eastern' and 'Western' nations.[16]

and

b. the divergent philosophical traditions upon which these regions are founded, leading to a perception of significant and irresolvable value differences between 'Western' and 'Eastern' cultures on key issues such as data privacy (Larson 2018; Horowitz et al. 2018; Houser 2018).

A range of recent technological and political developments may also be contributing to this mistrust, including concerns about the public and political influence of technology giants in the USA (Ochigame 2019); perceptions of and reactions to the Chinese Social Credit score system (Chorzempa et al. 2018; Song 2019); and concerns about contentious uses of AI technologies, with notably controversial examples including the use of AI in immigration control (Whittaker et al. 2018), criminal risk assessment in the USA (Campolo et al. 2017), and in tracking communities such as the Uighur Muslim

[15] This perception of mistrust was highlighted throughout the July workshop. Several participants reported being aware of or present at workshops focused on geopolitical implications of Chinese AI progress in which Chinese participants were excluded and events at which 'what should be done about China' (from a Western perspective) was raised as a key concern.

[16] As Ess (2005) notes, the terms 'Eastern' and 'Western' are not unproblematic, and are more products of colonialization than accurate terms for diverse nations and cultures. However, the terms continue to be used widely in the literature as shorthand for many of the broad cultural differences that are relevant to this paper; we therefore use the terms while being cognisant of their limitations.

minority in China (Mozur 2019). Adversarial rhetoric from political and defence leaders in the USA also contributes to this tension. Recent examples reported in the media include stating intentions to 'be the threat' of AI[17]; comments focused on the 'otherness' of China as an adversary,[18] amid broader concerns regarding Chinese technological progress as a threat to US global leadership (Jun 2018). A continued exacerbation of this culture of mistrust could severely undermine possibilities for global cooperation on AI development and governance.

In addition, it is unclear how far existing cross-cultural collaborations and alliances can go to shape the behaviour of actors that are as globally dominant as the USA, China, and the large multinational corporations based in these countries. Even if AI ethics frameworks can be agreed on in principle by multi-stakeholder groups, for example, it will be far from straightforward to implement them in practice to constrain the behaviour of those with disproportionate power to shape AI development and governance.

Another challenge for effective cooperation is balancing the need for global cooperation with the need for culturally and geographically sensitive differentiated approaches (Hagerty and Rubinov 2019). It is crucial we avoid a situation where one or two nations simply try to impose their values on the rest of the world (Acharya 2019). In certain specific domains, for example where AI is being used to support the delivery of healthcare, different cultures may perceive tradeoffs very differently (Feldman et al. 1999), and it may be not just possible but necessary to implement region-specific standards and governance. AI systems will also have different impacts as they are deployed in different cultural regions, which may also require different governance approaches (Hagerty and Rubinov 2019).

For some aspects of AI development and governance, however, cooperation will be much more crucial. For example, some potential uses of AI technologies in military contexts, such as in automated targeting and attack, could impinge upon human rights and international humanitarian law (Asaro 2012). Another concern is that by automating aspects of information gathering, decision-making, and response in military arenas, the potential for unwanted escalation in conflict situations may increase due to events occurring and requiring responses faster than is compatible with effective human judgement and oversight (Altmann 2019). In both these cases, there may be individual military advantages to nations pursuing the technology; but in the absence of international agreements and standards, the overall effect may be destabilizing.[19] International agreement will also be of particular importance for all cases in which AI technologies are developed in one region, but used or deployed in a different region. A key challenge for cross-cultural cooperation will therefore be to identify the areas where international agreement is most important, and distinguish these from areas where it is more appropriate to respect a plurality of approaches.

[17] 'Plenty of people talk about the threat from AI; we want to be the threat.' US Deputy Secretary of Defence Patrick Shanahan in an email to Department of Defence employees (Houser 2018)

[18] At a security forum in Washington D.C. in 2019, Kiron Skinner, the director of policy planning at the State Department, was quoted as saying about China 'This is a fight with a really different civilization and a different ideology and the United States has not had that before' and 'It's the first time that we will have a great power competitor that is not Caucasian' (Gehrke 2019).

[19] Similarly, a range of challenges in digital security, a domain likely to be affected by AI (Brundage et al. 2018), are likely to be greatly exacerbated in the absence of agreed upon international cybersecurity practices and standards (UN General Assembly 2015).

There are also more practical barriers to cooperation between nations: language barriers, lack of physical proximity, and immigration restrictions put limits on the ability of different cultures and research communities to communicate and collaborate. Furthermore, despite science being a global enterprise, the language of scientific publication remains predominantly English.

4 Overcoming These Barriers to Cooperation

While cross-cultural cooperation in AI ethics and governance will be genuinely challenging, we suggest that there are steps that can be taken today to make progress, without first needing to tackle the greater problems of finding consensus between cultures on all fundamental ethical and philosophical issues, or resolving decades of political tension between nations.

4.1 Building Greater Mutual Understanding, Including Around Disagreements

Mistrust between nations is a serious concern for the future of AI ethics and governance. However, we suggest that this mistrust is at least partly fuelled by misunderstandings and misperceptions, and that a first step towards building greater cross-cultural trust must therefore be to identify and correct important misperceptions, and enhance greater mutual understanding between cultures and nations.

It would be easy to assume that the main barrier to building trust between East and West around AI is that these regions of the world have very different fundamental values, leading them to different—perhaps conflicting—views of how AI should be developed, used, and governed from an ethical perspective. While value differences between cultures certainly exist, claims about how those differences manifest often depend on unexamined concepts and entrenched assumptions, and lack empirical evidence (Whittlestone et al. 2019). The idea that 'Eastern' and 'Western' ethical traditions are fundamentally in conflict also oversimplifies the relationship between the two. There are many different philosophical traditions that might be referred to under either heading: there are, for example, many important differences between relevant philosophical perspectives across China, Japan, and Korea (Gal 2019), and the values and perspectives of 'Western' philosophy have changed a great deal over time (Russell 1945). More broadly, ethical and cultural values in both regions are in constant evolution, as captured by projects such as the World Values Survey[20] and the Asian Barometer.[21]

Differences in ethical and cultural traditions and norms across regions are often assumed to underpin contrasting governance approaches. For example, privacy is often seen as an issue for which significant value differences exist between East and West, leading to a perception that laxer regulation and controls exist on data privacy in China compared with the USA and Europe. However, such claims are often made in very broad terms, without substantive evidence or analysis of how significant these differences are or how they manifest in practice (Ess 2005; Lü Yao-Huai 2005). This leads to

[20] http://www.worldvaluessurvey.org/wvs.jsp
[21] http://www.asianbarometer.org/

misunderstandings in both directions. First, there are significant differences between the USA and Europe on both conceptions of privacy (Szeghalmi 2015) and regulations that relate to privacy (McCallister et al. 2018). These are often missed in Chinese perceptions of Western societies, which tend to focus just on the USA.[22] Second, Western perceptions of data privacy in China may be outdated: Lü Yao-Huai (2005) highlighted as early as 2005 that the relevant literature on information ethics was much younger in China than in the USA, but was evolving quickly and in a manner significantly informed by Western scholarship. A range of papers and reports from Chinese scholars and policymakers have highlighted the importance of data privacy in AI ethics and governance (Beijing Academy of Artificial Intelligence 2019; Ying 2019; Zeng et al. 2018; Ding 2018b). Principles around protecting individuals' data privacy are also beginning to be borne out in regulatory action in China; over 100 apps have been banned by the government for user data privacy infringements, with dozens more being required to make changes relating to data collection and storage.[23] This is not to suggest that there are not meaningful differences in values, norms, and regulations relating to data privacy between these countries, but that such differences have often been oversimplified and are not well understood.

Another example of differing perceptions are those surrounding China's social credit score (SCS) system. The SCS has been discussed with great concern in Western media, policy circles, and scholarship, and presented as an example of Orwellian social control by the Chinese government (Botsman 2017; Pence 2018), representative of a culture and government with profoundly different values to the West (Clover 2016). However, both Chinese and Western sources have argued that there are significant misunderstandings surrounding the SCS. Multiple scholars have pointed out that the SCS is not designed to be a single unified platform that rates all 1.4 billion Chinese citizens (as is often supposed), but rather a web of individual platforms with latitude for different interpretations, with social credit scores mostly given by financial institutions (as opposed to a big data-driven comprehensive rating) (Mistreanu 2019; Sithigh and Siems 2019). Song (2019) notes that many of the measures in the SCS are designed to tackle issues such as fraud and corruption in local government. Chorzempa et al. (2018) also highlight that 'many of the key components of social credit, from blacklists to widespread surveillance… already exist in democracies like the United States.' China's SCS is likely to evolve significantly over time, and there are likely to be genuine reasons for concern both in terms of present and future implementation. However, a much clearer cross-cultural understanding of how the SCS works, is being used, and is impacting Chinese citizens, would allow dialogue on relevant ethics and governance issues to progress more constructively.

Given the lack of shared knowledge and discourse that has existed historically between regions such as the USA, Europe, and China, it is not surprising that many

[22] This was a misperception noted by multiple Chinese scholars at the aforementioned July workshop.

[23] In November 2019, the Chinese Ministry of Public Security banned 100 apps that failed to meet standards on individuals' data privacy; the body has investigated 683 apps in 2019 (National Cyber Security Advisory Centre 2019). In addition, in December 2019, the Chinese Ministry of Industry and Information Technology released a list of 41 apps that would have to make changes to comply with data regulations by the end of 2019 (Ministry of Industry and Information Technology of the People's Republic of China 2019). In July 2018, China's Shandong Province brought a major case relating to infringement of personal information against 11 companies (Ding 2018c).

misperceptions exist between them. We should therefore be wary of jumping too quickly to assume intractable and fundamental disagreements. Misunderstandings clearly exist in both directions: analyses of public opinion survey data suggest, for example, that both American and Chinese populations hold various misperceptions about the other nation's traits and characteristics (Johnston and Shen 2015). As mentioned above, in China, the diversity of Western societies is also often oversimplified down to a single pattern of American life. At the same time, the USA and Europe have historically struggled to understand China (Chen and Hu 2019), evidenced for example by repeated failures to predict China's liberalization (or lack thereof) or periods of economic growth (The Economist 2018; Cowen 2019; Liu 2019). Language barriers present a particular difficulty for Western nations in gleaning what is happening in China in terms of AI development, ethics, and governance (Zhang 2017; Ding 2019). As Andrew Ng points out in a 2017 interview in the Atlantic: 'The language issue creates a kind of asymmetry: Chinese researchers usually speak English so they have the benefit of access to all the work disseminated in English. The English-speaking community, on the other hand, is much less likely to have access to work within the Chinese AI community' (Zhang 2017). For example, Tencent released a book on AI strategy (Tencent Research Institute et al. 2017) which includes deep analysis of ethics, governance, and societal impacts, but has received relatively little English-language coverage (Ding 2018a). Even on empirical matters such as level of Chinese public investment in AI research and development, widely reported figures in the USA may be inaccurate by an order of magnitude (Acharya and Arnold 2019).

The recently published *Beijing AI Principles* (Beijing Academy of Artificial Intelligence 2019) and similar principles developed around the world (Cowls and Floridi 2018) in fact show substantial overlap on key challenges (Zeng et al. 2018; Jobin et al. 2019). The Beijing Principles make clear reference to the key concepts and values which have been prominent in other documents, including that AI should 'benefit all humankind'; respect 'human privacy, dignity, freedom, autonomy and rights'; and be 'as fair as possible, reducing possible discrimination and biases, improving its transparency, explainability and predictability.' In addition, both the *Beijing AI Principle*s, and the *National Governance Principles of the New AI, China*, call for openness and collaboration, with the latter encouraging 'cooperation across disciplines, domains, regions, and borders' (Laskai and Webster 2019). However, nations with different cultures may interpret and prioritize the same principles differently in practice (Whittlestone et al. 2019), which may be a further source of misunderstanding. We cannot simply assume, for example, that 'Western' cultures value privacy more highly than 'Eastern' ones; instead, we need a more nuanced understanding of how privacy may be prioritized differently when it comes into conflict with other important values, such as security (Capurro 2005). Similarly, although it is important to recognize that many cultures value autonomy, it is equally important to understand the different connotations and philosophical assumptions underpinning this value in different contexts (Yunping 2002).

Given the rich history of misunderstanding between nations, to build greater cross-cultural cooperation, we should start by focusing on identifying those misperceptions most relevant to AI ethics and building greater mutual understanding of where more substantive disagreements exist and are likely to impact governance approaches. In doing so, it is worth distinguishing explicitly between disagreements pertaining to

ethics as opposed to governance issues, since it may sometimes be possible for groups to agree on the same governance principles despite justifying them with different ethical assumptions, as we will discuss later. It may also be helpful to distinguish between misunderstandings that pertain directly to AI (such as misperceptions of other countries' investment in technology, or misinterpretation of data protection laws) and those that pertain to broader societal, political, or philosophical matters that are more indirectly relevant to AI, as they may require different approaches to resolve.

Acknowledging the role of misunderstandings does not, of course, imply that all matters of intercultural tension in AI ethics and governance are fundamentally based on misunderstandings. Deep and fundamental disagreements across regions will remain on a range of issues, including those relating to the relationship between the individual, society, and the state; the level and nature of integration between civil, private, and military sectors; and various specific matters of social policy. However, focusing initially on reducing misunderstandings will aid in establishing more clearly where these fundamental differences exist, while at the same time identifying contexts in which sufficient agreement exists for fruitful cooperation. Doing so is a crucial first step towards addressing the broader challenges of cross-cultural cooperation on AI ethics and governance.

4.2 Finding Ways to Cooperate Despite Disagreements

Even where important differences of view on AI ethics, governance, and broader societal issues exist, forms of agreement and cooperation can still be possible.

As mentioned earlier, a key outstanding challenge for AI ethics and governance is identifying those areas where cross-cultural agreement on norms, standards, or regulation is crucial, and where different interpretations and approaches are acceptable or even desirable. This is precisely the kind of challenge which itself requires cross-cultural cooperation: the delineations must be informed by diverse cultural perspectives on the impacts of AI in different contexts, and the needs and desires of different populations. Indeed, this approach is reflected in the *National Governance Principles of the New AI, China*, which includes the recommendation to 'Launch international dialogue and cooperation; with full respect for each country's principles and practices for AI governance, promote the formation of a broad consensus on an international AI governance framework, standards, and norms' (Laskai and Webster 2019).

Regional and cultural differences on the abstract level of ethical assumptions and high-level principles are also not necessarily a barrier to agreement on more concrete norms and governance. If it were impossible to reach any practical agreement without consensus on fundamental ethical issues, many important international agreements—such as the Nuclear Weapons Ban Treaty—would not have been possible. The notion of an 'incompletely theorized agreement' in legal scholarship (Sunstein 1995), describes how it is often possible for people who disagree on fundamental or abstract matters to nonetheless agree on specific cases—and that this is central to the functioning of law as well as of a pluralistic society more broadly. Several authors in the literature on intercultural information ethics have promoted the related idea of aiming to arrive at an 'overlapping consensus' (Rawls 1993), where different groups and cultures may have different reasons for supporting the same norms and practical guidelines (Taylor 1996; Søraker 2006; Hongladarom 2016). For example, Taylor (1996)

discusses how we have managed to ground internationally shared norms of human rights in different cultural traditions. While Western philosophies differ substantially from others such as Buddhism in how much importance they give to the human agent and its unique place in the cosmos, both seem to end up grounding the same norms of human rights.

Wong (2009) criticizes this idea that intercultural information ethics can arrive at shared norms with different justifications, suggesting that this risks making norms too 'thin', devoid of all normative content. Søraker (2006) acknowledges a similar objection to this 'pragmatic' approach to information ethics: that it may result in agreements that are fragile due to not being sufficiently grounded in substantive normative content. However, in line with Søraker's own response to these objections, we believe that the aim of 'overlapping consensus' should be to arrive at shared norms and practical guidelines which are in fact more robust by virtue of being endorsed and justified from a range of different philosophical or normative perspectives. This should be distinguished from a situation where one culture uses pragmatic arguments to attempt to force their own values upon others, or where several cultures reach agreement but for reasons with little normative content, which, we agree with Wong, would be concerning. Taylor's example of human rights being supported by multiple philosophical perspectives appears to demonstrate the plausibility of this kind of well-substantiated overlapping consensus.

Indeed, we suggest that finding areas of overlapping consensus on norms and practical guidelines may be much more important for ensuring the benefits of AI than aiming for global consensus on a shared set of fundamental values—an aim which underpins many recent proposals.[24] Consensus on high-level ethical principles does not necessarily mean they are well-justified (Benjamin 1995), and the best way to arrive at more robustly justified norms, standards, and regulation for AI will be to find those that can be supported by a plurality of different value systems.

4.3 Putting Principles into Practice

Even where it is possible to improve mutual understanding and arrive at shared governance approaches in theory, some might object that it will still be difficult to influence the development and use of AI in practice, especially since to do so requires influencing the behaviour of powerful states and companies that have little incentive to cooperate.

Although a full discussion of the complex power dynamics between states, companies, and other actors on issues relevant to AI ethics and governance is beyond the scope of this paper (and worthy of further research), we briefly explain why we do not think this barrier undermines our proposals. While challenging, historical precedent suggests that it is possible for public and academic alliances to influence the behaviour of powerful actors on issues of global importance. There is evidence to suggest that broad, cross-cultural 'epistemic communities'—i.e. networks of experts in a particular domain—can be particularly effective at supporting international policy coordination

[24] For example, Floridi et al., 2018 present a 'unified framework'; Awad et al. 2018 talk about 'developing global, socially acceptable principles for machine ethics'; Jobin et al. 2019 set out to investigate 'global agreement' on what constitutes ethical AI.

(Haas 1992). For example, a community of arms control experts helped shape cooperation between the USA and Russia during the Cold War by creating an internationally shared understanding of the problem of nuclear arms control (Adler 1992), and the ecological epistemic community managed to successfully coordinate national policies to protect the stratospheric ozone layer (Haas 1992).

The commitments of large companies and even nations around AI have already been influenced by combinations of employee activism, international academic research, and campaigning. A notable example is in the application of AI in military contexts. Concerns over the use of AI in warfare have been the subject of high-profile campaigns by experts across academia and civil society internationally, such as those involved in the International Committee for Robot Arms Control (ICRAC) and the Campaign to stop Killer Robots. These campaigns played a leading role in establishing discussion on lethal autonomous weapons (LAWs) at the United Nations Convention on Certain Conventional Weapons (CCW; the body that hosted negotiations over the banning of cluster munitions, blinding laser weapons, and landmines) (Belfield 2020). Ninety countries have put forward statements on LAWs, with most doing so at the CCW; 28 countries support a ban (Campaign to Stop Killer Robots 2018).[25] In 2018, over 4000 Google employees signed a letter protesting Google's involvement in the Pentagon's Project Maven, a military project exploring the use of AI in footage analysis (Shane and Wakabayashi 2018), and a number resigned (Conger 2018). Several other campaigning groups, comprised of scholars and researchers from the USA, Europe, Japan, China, Korea and elsewhere, released public articles and letters supporting the concerns of the Google practitioners (ICRAC 2018).[26] Google subsequently announced it would not renew its contract on Project Maven, and would not bid on a $10 billion Department of Defence cloud computing contract (Belfield 2020).

More broadly, international academic and civil society input has played a significant role in shaping principles that are likely to form the basis for binding regulation in years to come. For example, the European Commission's white paper *On Artificial Intelligence - A European Approach to Excellence and Trust* (European Commission 2020) lays out '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'(European Commission 2020b).This document was strongly influenced by the work of the European Union's High-Level Expert group on Artificial Intelligence, comprising 52 European experts from across academia, industry, and civil society.[27] Similarly, the US Department of Defence has formally adopted ethical principles on AI (US Department of Defense 2020), after 15 months of consultation with US-based academic, industry, and government stakeholders, and has hired staff to implement these principles (Barnett 2020). While in both cases the groups consulted were region-specific, the degree of alignment and overlap between principles

[25] China supports a ban on use of fully autonomous weapons on the battlefield, but not their production and development. The USA, Russia, UK, Israel, and France oppose a ban. (Kania 2018)

[26] An open letter was released by the International Committee for Robot Arms control and signed by over 1000 scholars and researchers, and members of the Campaign to Stop Killer Robots wrote public articles and letters to company leaders supporting the concerns of the Google petitioners: https://www.stopkillerrobots.org/2019/01/rise-of-the-tech-workers/

[27] Information on process and composition available at: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

developed in different regions suggests the insights and recommendations are substantially informed by interaction with broader epistemic communities from other regions. This suggests that insights gained from cross-cultural cooperation and consensus can meaningfully feed into regulatory frameworks at a regional and national level.

5 Recommendations

Academia has an important role to play in supporting cross-cultural cooperation on AI ethics and governance: both through research into where and what kinds of cooperation are most needed, and by establishing initiatives to overcome more practical barriers to cooperation. Our discussion in this paper raises many questions that will require diverse academic expertise to answer, including questions about what important misperceptions most hinder cooperation across regions; where international agreement is needed on AI ethics and governance; and how agreement might be reached on specific governance standards despite differences on ethical issues. Academic communities are also particularly well-suited to building greater mutual understanding between regions and cultures in practice, due to their tradition of free-flowing, international, and intercultural exchange of ideas. Academics can have open conversations with international colleagues in a way that is often challenging for those working in industry or government, and two academics from different parts of the world can productively collaborate even if each has strong criticisms of the other nation's government and/or companies.

The following recommendations indicate a number of specific steps that academic centres, research institutions, and individual researchers can take to promote cross-cultural understanding and cooperation on AI ethics and governance. Some excellent work is already ongoing in each of these areas. However, we believe that the pace at which AI is being deployed in new domains and regions calls for a greater focus within the academic community on building cross-cultural bridges and incorporating cross-cultural expertise within a wider range of ethics and governance research projects.

Develop AI Ethics and Governance Research Agendas Requiring Cross-cultural Cooperation Greater cross-cultural collaboration on research projects will play a crucial role in building an international research community that can support international policy cooperation (Haas 1992).

An example of a research project that might be well-suited to such collaboration is to conduct comparative foresight exercises exploring differences in how both positive visions and prominent concerns about AI's impact on society vary across cultures. This could help with developing a more global vision for what we hope to both achieve and avoid in AI development, which could guide more practical discussions around ethics and governance frameworks. There may be particularly valuable opportunities for consensus generation around avoidance of particular negative outcomes; safety and security are fundamental to human cultures worldwide, and so developing agreements to avoid threats to these may be an easier starting point. However, the authors feel that it is important not to neglect positive visions, as the opportunity for scholars across cultures to co-create shared positive futures may represent an excellent way to delve into nuances within shared values.

It would also be particularly valuable to explore the ongoing and expected impact of AI on developing countries, in collaboration with experts from these countries. Such research should aim to ensure that decisions to deploy technology in developing nations are made with the guidance of local expertise in such a way as to empower local communities (Hagerty and Rubinov 2019). On a more practical level, international research groups could productively work together to develop frameworks for international sharing of research, expertise, and datasets on AI safety, security, and avoidance of societal harm.

Collaboration between researchers from multiple regions and cultures will also be essential to further research on the topic of cross-cultural cooperation itself. Our discussion in this paper, especially in section 4, has pointed to many research areas in need of further exploration, including the following:

- Exploring, identifying, and challenging perceived cross-cultural differences in values, assumptions, and priorities relevant to AI ethics and governance, on multiple different levels, including[28] the following:

  - Analysing similarities and differences in technology ethics across different philosophical traditions, and exploring how these may affect AI development, deployment, impacts, and governance in practice;
  - Exploring the empirical evidence behind claims of key value differences between cultures. For example, a project might identify and explore perceived value differences between Eastern and Western cultures relevant to AI governance, such as those relating to data privacy, the role of the state vs. the individual, and attitudes towards technological progress;
  - Understanding regional differences in practical priorities and constraints relating to the use of AI in society, and the implications of these differences for AI research and development.

- Further analysis to identify aspects of AI governance where global agreement is needed, and differentiating these from areas in which cross-cultural variation is either acceptable or desirable;
- Cross-cultural contribution to the development of international and global AI standards in key domains for which these are needed; exploration of flexible governance models that allow for a combination of global standards and regional adaptability where appropriate;
- Exploring models and approaches for reaching agreement on concrete cases, decisions, or governance standards despite disagreement on more fundamental or abstract ethical issues, and identifying cases of this being done successfully in other domains that can be translated to AI ethics and governance.

Some excellent work is already ongoing in each of these areas. However, we believe that the pace at which AI is being deployed in new domains and regions calls for a

[28] Some excellent work on this question and related topics already exists in the field of intercultural information ethics—see for example Capurro (2005, 2008), Ess (2006), Hongladarom et al. (2009)—but we would be particularly keen to see cross-cultural research collaborations around these topics, and greater attention paid to this work in more practical AI ethics discussions and initiatives.

greater focus within the academic community on building cross-cultural bridges, and incorporating cross-cultural expertise, within a wider range of ethics and governance research projects.[29]

Translate Key Papers and Reports Language is a major practical barrier to greater cross-cultural understanding around AI development, governance and ethics, as it has been in many other areas of science (Amano et al. 2016). It would therefore be extremely valuable for the burgeoning literature on AI ethics and governance, as well as the literature on AI research, to be available in multiple languages. While many leading Asian scholars in AI are fluent in English, many are not; and the fraction of Western researchers fluent in Mandarin or Japanese is far lower.

Furthermore, some sources of the misunderstandings we have discussed may link to the ways in which key documents from one region are understood and presented in other regions. In Western media, China's 2017 *New Generation Artificial Intelligence Development Plan* has been presented as promoting the aim of global dominance in AI economically and strategically (Knight 2017; Demchak 2019). However, from the Chinese perspective, national AI development goals appear to be primarily motivated by the needs of the Chinese economy and society (China State Council 2017), rather than necessarily international competitive superiority (Ying 2019). It appears that some translations may have led to misinterpretations of key terms and points. For example, the original Chinese text referred to China becoming 'a primary AI innovation center of the world by 2030' (Ying 2019).[30] However, some English translations of the report translated this phrase as China becoming 'the world's primary AI innovation center' (e.g. Webster et al. 2017). This was then interpreted and presented by Eric Schmidt, former executive chairman of Google parent Alphabet, as 'By 2030 they will dominate the industries of AI. Just stop for a sec. The [Chinese] government said that.' (Shead 2017). While this evolution of wording is not substantial in one sense, it carries important connotations; the language in the original context carries much softer connotations of leadership and progress, as opposed to global dominance. Having multiple high-quality translations of the most important documents would allow scholars to explore nuances of language and context that may be lost in the reporting of these documents.

Producing high-quality versions of papers and reports in multiple languages also conveys respect and an appetite to engage cross-culturally, which is likely to encourage cooperation. High-quality translation of academic and policy materials is a challenging and time-consuming task that we would encourage being supported and celebrated more strongly. There is a growing body of work to be celebrated in this vein; for example, Jeff Ding's translation of a range of key Chinese AI documents (Ding 2019), the work of Intellisia in China on international relations, technology, and other topics, which publishes in 5 languages (http://www.intellisia.org/); Brian Tse's translation to Chinese of documents including OpenAI's Charter[31]; and New America's translation of the Chinese Ministry of Industry and Information Technology's *Three Year Action Plan* (Triolo et al. 2018).

[29] Several authors of this paper are part of an initiative that aims to support cross-cultural research of this nature between the UK and China: https://ai-ethics-and-governance.institute/

[30] This specific translation was also provided to us by participants in the July 2019 workshop.

[31] https://openai.com/charter/

Alternate Continents for Major AI Research Conferences and Ethics and Governance Conferences To encourage more global participation in AI development, ethics, and governance, we recommend that many of the leading conferences and fora on these topics alternate between multiple continents. This has several advantages. It reduces the cost and time commitment for scholars from parts of the world in which these conferences do not frequently take place to participate. It avoids restrictive visa limitations differentially affecting certain parts of the global research community. It encourages the involvement of local organizers, who can play an effective role in engaging local research communities who might not consider travelling far overseas for an event. It also encourages organizers to run events multilingually rather than monolingually.

Again, there are encouraging steps. Of AI research conferences, IJCAI took place in Macau in 2019 and Beijing in 2013, the first two times the conference had been held in China (although it has been held in Japan twice). ICML took place in Beijing in 2014, and will be in Seoul in 2021, and ICLR 2020 will be held in Ethiopia, making it the first of the top tier major machine learning conferences to be held in Africa. There are fewer established conferences explicitly focused on AI ethics and governance since the field's growth is relatively recent, but it may be particularly important for these conferences to ensure a global presence by alternating the continent on which they are held if possible. AI Ethics and Society, for example, is currently held in the USA due to ties to AAAI; the importance of building an international academic community around these issues may justify finding some way to change this. There is a burgeoning set of AI ethics and governance conferences in China, including the Beijing Academy of AI Conference series. There are also several existing conferences which cover topics relevant to AI ethics and governance (even if not so explicitly centred around them), which do enable more international participation, such as the World Summit on the Information Society Forum (held most years in Geneva), the Internet Governance Forum, and RightsCon (which have both been held in a range of locations historically including in South America, India, and Africa, though neither in East Asia).[32]

Establish Joint and/or Exchange Programmes for PhD Students and Postdocs Encouraging cross-cultural collaboration between researchers from different cultures early on in their careers will help support greater cooperation and mutual understanding as research advances. Many international fellowships and exchange programmes exist, especially between the USA and China (e.g. the Zhi-Xing China Fellowship and the Schwarzman Scholars programme) as well as between the UK and Singapore (King's College London offers a joint PhD programme in Philosophy or English with the National University of Singapore). To our knowledge, no such initiatives exist explicitly focused on AI; the only initiative focused on AI ethics and governance that we are currently aware of is the international fellowship programme recently established by the Berggruen China Institute (Bauch 2019).[33] Establishing more such programmes could be valuable for the future of international cooperation

[32] There is also a role for industry groups to play in encouraging cross-cultural collaboration through international workshops and conferences. We highlight the work of the AI Industry Alliance as an example of a recently established initiative in this space http://www.aiiaorg.cn/

[33] The Tianxia Fellowship, established by the Center for Long-term Priorities (http://www.longtermpriorities.org/), also includes AI safety and governance among other topics.

around AI, and there are many existing models and initiatives from which to build and learn.

More broadly, the authors endorse the Partnership on AI's recommendations to governments on establishing visa pathways, simplifying and expediting visa processes, and ensuring just standards to support international exchange and collaboration of AI/ML multidisciplinary experts. We emphasize that these recommendations include experts working or seeking to work on AI ethics and governance (which sometimes fall outside of what is classed as 'skilled technology work') (PAI Staff 2019).

6 Limitations and Future Directions

We believe that academia has an important role to play in supporting cross-cultural cooperation in AI ethics and governance: that it is possible to establish effective communities of mutual understanding and cooperation without needing to resolve all fundamental value differences, and that reducing misunderstandings and misperceptions between cultures may be of particular importance.

However, we recognize that the suggestions in this paper cannot go all the way to overcoming the many barriers to cross-cultural cooperation, and that much more work needs to be done to ensure AI will be globally beneficial. We briefly highlight two broad avenues of further research in support of this goal:

More Detailed Analysis of Barriers to Cross-cultural Cooperation, Especially Those Relating to Power Dynamics and Political Tensions While analysis of historical successes suggest that it is possible for cross-cultural initiatives around AI ethics and governance to considerably shape how norms, standards, and regulation evolve in practice, there are still many barriers to implementation and enforcement that we were unable to consider in this analysis. Further research into when and how attempts to influence globally relevant norms and regulation have been successful in the past would be of considerable value.

We acknowledged earlier in this paper that various issues related to power relations and political tensions likely pose significant barriers to cross-cultural cooperation, beyond problems of value differences and misunderstandings between cultures. More research on how these issues present barriers to cross-cultural cooperation in AI ethics and governance would therefore be particularly valuable, helping us to understand the limits of academic initiatives in promoting cooperation, and in what ways these approaches need to be embedded within a broader analysis of power and political dynamics.

Considering the Unique Challenges of Cross-cultural Cooperation Around More Powerful Future AI Systems Future advances in AI, which some scholars have theorized could have impacts as transformative as the industrial or agricultural revolutions (Karnofsky 2016; Zhang and Dafoe 2019), may raise new challenges for global cooperation of a greater scale than we already face. Without careful global stewardship, such advances could lead to unprecedented inequalities in wealth and power between technology-leading and lagging nations. Others have gone further, theorizing about the possibility of developing systems exhibiting superintelligence (i.e. greater-than-human

general intelligence; Bostrom 2014). Such systems, due to their tremendous capability, might pose catastrophic risks to human civilisation if developed without careful forethought and attention to safety. It has been proposed that a key challenge for avoiding catastrophic outcomes will be value alignment—designing systems that are aligned with humanity's values (Russell 2019). This would greatly increase the importance and urgency of reaching global consensus on shared values and principles, as well as finding ways to design systems to respect values that are not shared.

Expert views differ widely on how far in the future such advances might lie, with most predicting decades. However, developing the collaborations and agreements necessary for an effective and coordinated response may also require decades of work. This suggests that cooperative initiatives today must address not just the ethics and governance challenges of current AI systems, but should also lay the groundwork for anticipating and engaging with future challenges.

7 Conclusion

The full benefits of AI cannot be realized across global societies without a deep level of cooperation—across domains, disciplines, nations, and cultures. The current unease and mistrust between the USA and Europe on the one hand, and China on the other hand, places a particular strain on this. Misunderstandings may play an important role in fuelling this mistrust, and differences in broader societal and political priorities frequently appear to be overemphasized or misunderstood.

At the same time, it would be naive to assume that all major ethical principles relating to AI can be shared in full between these regions, and can be enshrined in rules and standards. Even if this were the case, it would not be desirable for these regions to be overly dominant in shaping the ethics and governance of AI globally; all global communities to be affected by AI must be included and empowered. However, efforts to achieve greater understanding between these 'AI superpowers' may help in two ways: Firstly, by reducing key tensions within the global AI governance sphere. Secondly, by providing lessons that can contribute to ethics and governance frameworks capable of supporting a greater diversity of values while allowing consensus to be achieved where needed. For a well-functioning system of global cooperation in AI, the challenge will be to develop models that combine both principles and standards shaped and supported by global consensus, and the variation that allows research and policy communities to best serve the needs of their societies.

On a more practical level, the international AI research community, and AI ethics and governance communities, must think carefully about how their own activities can support global cooperation, and a better understanding of different societal perspectives and needs across regions. Greater cross-cultural research collaboration and exchange, conferences taking place in different regions, and more publication across languages can lower the barriers to cooperation and to understanding different perspectives and shared goals. With political winds increasingly favouring isolationism, it has never been more important for the research community to work across national and cultural divides towards globally beneficial AI.

Acknowledgments The authors would like to thank the participants of the July 11–12 Cross-cultural trust for beneficial AI workshop for valuable discussions relevant to the themes of the paper, as well as Emma Bates, Haydn Belfield, Martina Kunz, Amritha Jayanti, Luke Kemp, Onora O'Neill, and two anonymous manuscript reviewers for helpful comments on previous drafts of this paper.

References

Acharya, A. (2019). Why international ethics will survive the crisis of the liberal international order. *SAIS Review of International Affairs, 39*(1), 5–20.

Acharya, A., & Arnold, Z. (2019). Chinese Public AI R&D Spending: provisional findings. Centre for Security and Emerging Technologies Issue Brief. Available at: https://cset.georgetown.edu/wp-content/uploads/Chinese-Public-AI-RD-Spending-Provisional-Findings-2.pdf. Accessed 23 Dec 2019.

Adler, E. (1992). The emergence of cooperation: national epistemic communities and the international evolution of the idea of nuclear arms control. *International Organization, 46*(1), 101–145.

Allen, J. R., Husain, A. (2017). The next space race is artificial intelligence. *Foreign Policy*. Available at: https://foreignpolicy.com/2017/11/03/the-next-space-race-is-artificial-intelligence-and-america-is-losing-to-china/. Accessed 21 Dec 2019.

Altmann, J. (2019). Autonomous weapon systems–dangers and need for an international prohibition. Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz) (pp. 1–17). Springer, Cham.

Amano, T., González-Varo, J. P., & Sutherland, W. J. (2016). Languages are still a major barrier to global science. *PLoS Biology, 14*(12), e2000933.

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross, 94*(886), 687–709.

Askell, A., Brundage, M., & Hadfield, G. (2019). The role of cooperation in responsible AI development. arXiv preprint arXiv:1907.04534.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59.2018.

Barnett, J. (2020). DOD hires policy team to implement AI principles. Available at: https://www.fedscoop.com/dod-hires-new-ai-policy-team/. Accessed 12 Mar 2020.

Bauch, R. (2019). Berggruen institute announces 2019-2020 class of fellows in U.S. and China as international cohort of Berggruen thinkers to study great transformations. Berggruen Institute. Available at: https://www.berggruen.org/news/berggruen-institute-announces-2019-2020-class-of-fellows-in-u-s-and-china-as-international-cohort-of-berggruen-thinkers-to-study-great-transformations/. Accessed 27 Dec 2019.

Beijing Academy of Artificial Intelligence. (2019). Beijing AI Principles. Available at: https://www.baai.ac.cn/blog/beijing-ai-principles. Accessed 24 Dec 2019.

Belfield, H. (2020). Activism by the AI community: analysing recent achievements and future prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 15-21).

Benaich, N., & Hogarth, I. (2019). State of AI Report 2019. Available at https://www.stateof.ai/. Accessed 19 Dec 2019.

Benjamin, M. (1995). The value of consensus. In *Society's choices: Social and ethical decision making in biomedicine*. National Academy Press.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford Univ. Press.

Botsman, R. (2017). Big data meets Big Brother as China moves to rate its citizens. *Wired UK*, 21.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Campaign to Stop Killer Robots (2018). Country views on killer robots. Available at: https://www.stopkillerrobots.org/wp-content/uploads/2018/11/KRC_CountryViews22Nov2018.pdf. Accessed 11 Mar 2020.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). AI Now 2017 report. AI Now Institute at New York University. Available at: https://ainowinstitute.org/AI_Now_2017_Report.pdf. Accessed 18 Dec 2019.

Capurro, R. (2005). Privacy. An intercultural perspective. *Ethics and Information Technology, 7*(1), 37–47.

Capurro, R. (2008). Intercultural information ethics. The handbook of information and computer ethics, 639.

Cave, S., & ÓhÉigeartaigh, S. (2018). An AI race for strategic advantage: rhetoric and risks. *Proceedings of the 2018 AAAI/ACM conference on artificial intelligence, Ethics and Society*.

Chen, D., & Hu, J. (2019) No, there is no US-China 'clash of civilizations' the diplomat. Available at: https://thediplomat.com/2019/05/no-there-is-no-us-china-clash-of-civilizations/. Accessed Dec 2016.

China State Council (2017). New generation artificial intelligence development plan. Available at: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm (translation: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/). Both accessed 20 Dec 2019.

Chorzempa, M., Triolo, P., & Sacks, S. (2018). China's social credit system: a mark of progress or a threat to privacy? *Policy Briefs PB18–14*, Peterson Institute for International Economics.

Cihon, P. (2019). Standards for AI governance: international standards to enable global coordination in AI research & development. *Future of humanity institute technical report*. Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf. Accessed 20 Dec 2019.

Clover, C. (2016). China: when big data meets big brother. *Financial times*. Available at: https://www.ft.com/content/b5b13a5e-b847-11e5-b151-8e15c9a029fb.

Conger, K. (2018). Google employees resign in protest against pentagon contract. Available at: https://gizmodo.com/google-employees-resign-in-protest-against-pentagon-con-1825729300. Accessed 11 Mar 2020.

Cowen, T. (2019). What if everyone's wrong about China? *Bloomberg*. Available at: https://www.bloomberg.com/opinion/articles/2019-08-19/china-s-liberalization-shouldn-t-be-ruled-out-just-yet

Cowls, J., & Floridi, L. (2018). Prolegomena to a white paper on an ethical framework for a good AI society. SSRN preprint.

Demchak, C. C. (2019). China: determined to dominate cyberspace and AI. *Bulletin of the Atomic Scientists, 75*(3), 99–104.

Ding, J. (2018a). ChinAI #1 available at: https://mailchi.mp/b945e27a35ff/chinai-newsletter-1-welcome. Accessed 30 Dec 2019.

Ding, J. (2018b). Deciphering China's AI dream. *Future of Humanity Institute Technical Report*. Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf. Accessed 19 Dec 2019.

Ding, J. (2018c). ChinaAI #19: is the wild east of big data coming to an end? A turning point case in personal information protection. ChinAI Newsletter. Available at: https://chinai.substack.com/p/chinai-newsletter-19-is-the-wild-east-of-big-data-coming-to-an-end-a-turning-point-case-in-personal-information-protection Accessed 28 December 2019.

Ding, J. (2019). ChinAI #48: year 1 of ChinAI. ChinAI Newsletter. Available at: https://chinai.substack.com/p/chinai-48-year-1-of-chinai. Accessed 26 Dec 2019.

Ess, C. (2005). Lost in translation?: intercultural dialogues on privacy and information ethics. *Ethics and Information Technology, 1*, 1–6.

Ess, C. (2006). Ethical pluralism and global information ethics. *Ethics and Information Technology, 8*(4), 215–226.

European Commission (2020). On artificial intelligence - A European approach to excellence and trust. White Paper. Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf. Accessed 11 Mar 2020.

European Commission (2020b). https://ec.europa.eu/digital-single-market/en/artificial-intelligence. Accessed 11 Mar 2020.

Feldman, M. D., Zhang, J., & Cummings, S. R. (1999). Chinese and US internists adhere to different ethical standards. *Journal of General Internal Medicine, 14*(8), 469–473.

Gal, D. (2019). Perspectives and approaches in AI ethics: East Asia. *Oxford Handbook of Ethics of Artificial Intelligence,* Oxford University Press, Forthcoming.

Gehrke, J. (2019). State Department preparing for clash of civilizations with China. *The Washington Examiner*. Available at: https://www.washingtonexaminer.com/policy/defense-national-security/state-department-preparing-for-clash-of-civilizations-with-china. Accessed 22 Dec 2019.

Haas, P. M. (1992). Introduction: epistemic communities and international policy coordination. *International Organization, 46*(1), 1–35.

Hagerty, A., & Rubinov, I. (2019). Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv preprint arXiv:1907.07892.

Haynes, A. & Gbedemah, L. (2019). The global AI index: methodology. Available at: https://www.tortoisemedia.com/intelligence/ai. Accessed 21 Dec 2019.

Hongladarom, S. (2016). Intercultural information ethics: a pragmatic consideration. In *Information cultures in the digital age* (pp. 191–206). Wiesbaden: Springer VS.

Hongladarom, S., Britz, J., Capurro, R., Hausmanninger, T., & Nakada, M. (2009). Intercultural information ethics. *International Review of Information Ethics, 11*(10), 2–5.

Horowitz, M. C., Allen, G. C., Kania, E. B., & Scharre, P. (2018). *Strategic competition in an era of artificial intelligence*. Washington, DC: Center for New American Security.

Houser, K. (2018). US military declares mandate on AI. Futurism. Available at: https://futurism.com/the-byte/jaic-militarys-ai-center. Accessed 22 Dec 2019.

Hudson, R. (2019) France and Canada move forward with plans for global AI expert council. *Science Business*. Available at: https://sciencebusiness.net/news/france-and-canada-move-forward-plans-global-ai-expert-council. Accessed 27 Dec 2017.

International Committee for Robot Arms Control (2018). Open letter in support of Google employees and tech workers. Available at: https://www.icrac.net/open-letter-in-support-of-google-employees-and-tech-workers/. Accessed 11 Mar 2020.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Johnston, A. I., & Shen, M. (Eds.). (2015). *Perception and misperception in American and Chinese views of the other (p. 63)*. Washington, DC: Carnegie Endowment for International Peace.

Jun, Z. (2018). The West exaggerates China's technological progress. *Nikkei Asian Review*. Available at: https://asia.nikkei.com/Opinion/The-West-exaggerates-China-s-technological-progress. Accessed 30 Dec 2019.

Kania, E. (2018). China's strategic ambiguity and shifting approach to lethal autonomous weapons systems. *Lawfare, April*, *20*.

Karnofsky, H. 2016. Potential risks from advanced artificial intelligence: the philanthropic opportunity. Available at: https://www.openphilanthropy.org/blog/potential-risks-advanced-artificialintelligence-philanthropic-opportunity. Accessed 9 Mar 2020.

Knight, W. (2017). China plans to use artificial intelligence to gain global economic dominance by 2030. *MIT Technology Review*. Available at https://www.technologyreview.com/s/608324/china-plans-to-use-artificial-intelligence-to-gain-global-economic-dominance-by-2030/. Accessed 26 Dec 2019.

Larson, C. (2018). China's AI imperative. *Science, 359*(6376), 628–630.

Laskai, L. & Webster, G. (2019). Translation: Chinese expert group offers 'governance principles' for 'responsible ai'. New America, *DigiChina*. Available at: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/. Accessed 30 Dec 2019.

Lee, K. F. (2017). The real threat of artificial intelligence. *The New York Times*, 24. Available here: https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html. Accessed 30 Dec 2019.

Liu, M. (2019). 30 years after Tiananmen: how the West still gets China wrong. *Foreign Policy*. Available at: https://foreignpolicy.com/2019/06/04/30-years-after-tiananmen-how-the-west-still-gets-china-wrong/. Accessed 19 Dec 2019.

May, T. (2018). Transcript of keynote speech at 2018 World Economic Forum. Available at: https://www.weforum.org/agenda/2018/01/theresa-may-davos-address/. Accessed 27 Dec 2019.

McCallister, J., Zanfir-Fortuna, G., & Mitchell, J. (2018). Getting ready for the EU's stringent data privacy rule. *Journal of Accountancy, 225*(1), 36–41.

Ministry of Industry and Information Technology of People's Republic of China. (2019). APP (first batch) notification on infringement of user rights. Available at: http://www.miit.gov.cn/n1146290/n1146402/n1146440/c7575066/content.html. Accessed 29 Dec 2019.

Mistreanu, S. (2019). Fears about China's social-credit system are probably overblown, but it will still be chilling. *Washington Post*. Available at: https://www.washingtonpost.com/opinions/2019/03/08/fears-

about-chinas-social-credit-system-are-probably-overblown-it-will-still-be-chilling/. Accessed 20 Dec 2019.
Mozur, P. (2019). One month, 500,000 Face Scans: how China is using AI to profile a minority. *The New York Times*. Available at: https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html. Accessed 20 Dec 2019.
National Cyber Security Advisory Centre. (2019). Available at: https://mp.weixin.qq.com/s/smT4RbHsA_x0vIZjEKV_yg?. Accessed 29 Dec 2019.
Ochigame, R. (2019). The invention of "ethical AI": how big tech manipulates academia to avoid regulation. *The Intercept*. Available at: https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/. Accessed 22 Dec 2019.
PAI Staff. (2019). Partnership on AI calls for visa accessibility globally to accelerate responsible AI development. Available at: https://www.partnershiponai.org/the-partnership-on-ai-calls-for-visa-accessibility-globally-to-accelerate-responsible-ai-development/. Accessed 21 Dec 2019.
Pence, M. (2018) Remarks by Vice President Pence on the Administration's policy toward China. United States White House. Available at: https://www.whitehouse.gov/briefings-statements/remarks-vice-president-pence-administrations-policy-toward-china/. Accessed 31 Dec 2019.
Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S., & Niebles, J. C. (2019). *The AI Index 2019 Annual Report*. Stanford: AI Index Steering Committee, Human-Centered AI Institute, Stanford University.
Rawls, J. (1993). *Political liberalism* (pp. 134–149). Columbia University Press.
Russell, B. (1945). *A history of Western philosophy*. Allen & Unwin.
Russell, S. (2019). *Human compatible: artificial intelligence and the problem of control*. Penguin.
Shane, S., & Wakabayashi, D. (2018). 'The business of war': Google employees protest work for the pentagon. *The New York Times*, 4.
Shead, S. (2017). Eric Schmidt on AI: 'Trust me, these Chinese people are good'. *Business Insider*. Available at: https://www.businessinsider.my/eric-schmidt-on-artificial-intelligence-china-2017-11/. Accessed 30 Dec 2017.
Simonite, T. (2017). AI could revolutionise war as much as nukes. *Wired*. Available at: https://www.wired.com/story/ai-could-revolutionize-war-as-much-as-nukes/. Accessed 20 Dec 2019.
Sithigh, D. M., & Siems, M. (2019). The Chinese social credit system: A model for other countries?. EUI Department of Law Research Paper, (2019/01).
Song, B. (2019). The West may be wrong about China's social credit system. *New Perspectives Quarterly, 36*(1), 33–35.
Søraker, J. H. (2006). The role of pragmatic arguments in computer ethics. *Ethics and Information Technology, 8*(3), 121–130.
Stewart, P. (2017). U.S. weighs restricting Chinese investment in artificial intelligence. *Reuters*. Available at: https://www.reuters.com/article/us-usa-china-artificialintelligence/u-s-weighs-restricting-chinese-investment-in-artificial-intelligence-idUSKBN1942OX. Accessed 20 Dec 2019.
Sunstein, C. R. (1995). Incompletely theorized agreements. *Harvard Law Review, 108*(7), 1733–1772.
Szeghalmi, V. (2015). The definition of the right to privacy in the United States of America and Europe. *Hungarian Yearbook of International Law and European Law*, 397.
Taylor, C. (1996). Conditions of an unforced consensus on human rights. Available at: http://people.brandeis.edu/~teuber/Taylor,%20Conditions%20of%20an%20Unforced%20Consensus.pdf
Tencent Research Institute, China Academy of Information and Communications Technology, Tencent AI Lab, and Tencent Open Platform. (2017). *Artificial intelligence: A national strategic initiative for artificial intelligence (*人工智能:国家人工智能战略行动抓手*)*. China Renmin University Press.
The Economist. (2018). How the West got China wrong. Available at: https://www-economist-com.ezp.lib.cam.ac.uk/leaders/2018/03/01/how-the-west-got-china-wrong. Accessed 13 Dec 2019.
Triolo, P., Kania, E., & Webster, G. (2018). Translation: Chinese government outlines AI ambitions through 2020. *New America, DigiChina*, 26.
UN General Assembly (2015). Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the context of International Security. Seventieth Session, Item, 93.
US Department of Defense (2020). Release: DOD adopts ethical principles for artificial intelligence. Available at https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/. Accessed 11 Mar 2020.
Webster, G., Creemers, R., Triolo, P., & Kania, E. (2017). Full translation: China's 'new generation artificial intelligence development plan'. *New Ameria DigiChina*. Available at: https://www.newamerica.

org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/. Accessed 26 Dec 2019.
Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI now report 2018*. AI Now Institute at New York University.
Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research*. London: Nuffield Foundation.
Wong, P. H. (2009). What should we share? Understanding the aim of intercultural information ethics. *ACM SIGCAS Computers and Society, 39*(3), 50–58.
Yao-Huai, L. (2005). Privacy and data privacy issues in contemporary China. *Ethics and Information Technology, 7*(1), 7–15.
Ying, F. (2019). Understanding the AI challenge to humanity. China US focus. Available at: https://www.chinausfocus.com/foreign-policy/understanding-the-ai-challenge-to-humanity. Accessed 29 Dec 2019.
Yunping, W. (2002). Autonomy and the Confucian moral person. *Journal of Chinese Philosophy, 29*(2), 251–268.
Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking artificial intelligence principles. arXiv preprint arXiv: 1812.04814.
Zhang, S. (2017). China's artificial-intelligence boom. *The Atlantic*, 20170216, 20170924.
Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. Available at SSRN 3312874.

Affiliations

Seán S. ÓhÉigeartaigh[1,2] · Jess Whittlestone[1] · Yang Liu[1,2] · Yi Zeng[2,3] · Zhe Liu[2,4]

[1] Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

[2] China-UK Research Centre for AI Ethics And Governance, Beijing, China

[3] Research Center for AI Ethics and Safety, Beijing Academy of Artificial Intelligence, Beijing, China

[4] Centre for Philosophy and the Future of Humanity, Peking University, Beijing, China