AI and Ethics
https://doi.org/10.1007/s43681-020-00037-w

ORIGINAL RESEARCH

Bridging the gap: the case for an 'Incompletely Theorized Agreement' on AI policy

Charlotte Stix[1] · Matthijs M. Maas[2]

Received: 10 September 2020 / Accepted: 30 November 2020

Abstract

Recent progress in artificial intelligence (AI) raises a wide array of ethical and societal concerns. Accordingly, an appropriate policy approach is urgently needed. While there has been a wave of scholarship in this field, the research community at times appears divided amongst those who emphasize 'near-term' concerns and those focusing on 'long-term' concerns and corresponding policy measures. In this paper, we seek to examine this alleged 'gap', with a view to understanding the practical space for inter-community collaboration on AI policy. We propose to make use of the principle of an 'incompletely theorized agreement' to bridge some underlying disagreements, in the name of important cooperation on addressing AI's urgent challenges. We propose that on certain issue areas, scholars working with near-term and long-term perspectives can converge and cooperate on selected mutually beneficial AI policy projects, while maintaining their distinct perspectives.

Keywords Artificial intelligence · AI · Artificial intelligence policy · Long term · Short term · Artificial intelligence ethics · Cooperation models · Incompletely theorized agreement · Overlapping consensus

1 Introduction

The prevailing uncertainty around the trajectory and impact of artificial intelligence (AI) makes it clear that appropriate technology policy approaches are urgently needed. The possible negative ethical and societal impacts of AI are considerable: from algorithmic bias to AI-enabled surveillance, and from lethal autonomous weapons systems to widespread technology-induced unemployment. Moreover, some forecast that continuing progress in AI capabilities will eventually make AI systems a 'general-purpose technology' [1], or may even enable the development of 'high-level machine intelligence' (HLMI) [2] or other 'transformative' capabilities [3, 4]. Debate on these latter scenarios is diverse, and has at times focused on what some have referred to as 'Artificial General Intelligence' (AGI) [5]. On the surface, those concerned with AI's impacts can appear divided between those who focus on discernible problems in the near term, and those who focus on more uncertain problems in the longer term [6–9].

This paper wants to investigate the dynamics and debates between these two communities, with an eye to fostering policy effectiveness through greater cooperation. In doing so, this paper seeks to take up the recent call to 'bridge the near- and long-term challenges of AI' [9]. The focus therefore is not on the relative urgency of existing algorithmic threats (such as e.g. facial recognition or algorithmic bias), nor on the relative plausibility of various advanced AI scenarios (such as e.g. HLMI or AGI), nor do we mean to suggest that a long-term perspective is solely focused on or concerned with AGI [10–12]. Rather, the paper proposes that even if some community divergence exists, each group's overarching intention to contribute to responsible and ethical AI policy[1] would benefit from cooperation within key

✉ Matthijs M. Maas
mmm71@cam.ac.uk

Charlotte Stix
c.stix@tue.nl

[1] Philosophy and Ethics Group, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands

[2] Centre for the Study of Existential Risk & Centre for the Future of Intelligence, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, UK

[1] We define 'AI policy' broadly, as concrete soft or hard governance measures which may take a range of forms such as principles, codes of conduct, standards, innovation and economic policy or legislative approaches, along with underlying research agendas, to shape AI in a responsible, ethical and robust manner. Our paper works under the

Published online: 15 January 2021

domains to maximize policy effectiveness. The paper suggests that differences may be overstated, and proposes that even if one assume such differences, these are not practically insurmountable. Rather, it argues that the principle of an 'incompletely theorized agreement', originally derived from constitutional law, provides both philosophical foundations and historical precedent for a form of cooperation between divergent communities that enables progress on urgent shared issues, without compromising on their respective goals.

The paper proceeds as follows: in Sect. 2, we provide a short rationale for our proposed intervention. We briefly lay out the landscape of AI policy concerns and the structure of the associated AI ethics and policy community. This is followed by a discussion, drawing on historical cases as well as the contemporary challenges facing AI policy scholars, of how fragmentation within an expert community might hinder progress on key and urgent policies. In Sect. 3, we explore potential sources which could contribute to community divergence. We touch on epistemic and methodological disagreements and normative disagreements, and home in on pragmatic disagreements around the tractability of formulating AI policy actions today which maintain long-term relevance. We briefly review how serious these disagreements are, arguing that these trade-offs are often exaggerated, or do not need to preclude collaboration. Finally, in Sect. 4, we propose that one consolidating avenue to harness mutually beneficial cooperation for the purpose of effective AI policy could be anchored in the constitutional law principle of an 'incompletely theorized agreement'. This proposal works under the assumption that the influence of a community on policy making is significantly stronger if they act as a united front, rather than as scattered subgroups.

2 AI policy: a house divided?

Recent progress in AI has given rise to an array of ethical and societal concerns.[2] Accordingly, there have been calls for appropriate policy measures to address these. As an "omni-use technology" [13], AI has both potential for good [14–16] as well as for bad [17–19] applications. The latter include: various forms of pervasive algorithmic bias [20, 21], challenges around transparency and explainability [22, 23]; the safety of autonomous vehicles and other cyber-physical systems [24], or the potential of AI systems to be used in (or be susceptible to) malicious or criminal attacks [25–27]; the erosion of democracy through e.g. 'computational propaganda' or 'deep fakes' [28–30], and an array of threats to the full range of human rights [31, 32]. The latter may eventually cumulate in the possible erosion of the global legal order by the comparative empowerment of authoritarian states [33, 34]. Finally, some express concern that continued technological progress might eventually result in increasingly more 'transformative' AI capabilities [3], up to and including AGI. Indeed, a number of AI researchers expect some variation of 'high-level machine intelligence' to be achieved within the next five decades [2]. Some have suggested that if those transformative capabilities are not handled with responsibility and care, such developments could well result in new and potential catastrophic risks to the welfare, autonomy, or even long-term survival of societies [35, 36].

Looking at the current debate and scholarship involved in the aforementioned areas, we note, along with other scholars [6, 8, 37], that there appears to be a fuzzy split, along a temporal 'near-term'/'long-term' axis. This perception matters because, as in many other fields and contexts, a perceived or experienced distinction may eventually become a self-fulfilling prophecy [38]. This holds true even if the perceived differences are based on misperceptions or undue simplification by popular scientific media [39–41]. Of course, fragmentation between near- and longer-term considerations of AI's impact is only one way to explore the growing community, and it may not be the sole issue to overcome to maximize policy impact. However, for the purpose of this paper, our focus is on this specific gap.

2.1 The policy advantages of collaboration: lessons from history

The current AI ethics and policy community is a young one. Policy shifts, on the other hand, take time. As such, it is difficult to clearly outline what impact current dynamics have. We are, after all, still in the early stages of these developments. Nevertheless, historical examples of adjacent fields can help to demonstrate and forecast how fragmentation, or, conversely cooperation, on policy goals within the AI ethics and policy community could strengthen impact on technology policy.

Why should potential fragmentation along an axis such as near- and longer-term concerns worry us? History shows that the structure of a field or community affects the ability of its members to shape and influence policy downstream. Importantly, it shows that there is significant benefit derived from collaboration. We put forward three historic examples of adjacent technology policy fields to AI, meaning those that tackled equally new and emerging technologies. We briefly highlight one case where fragmentation may have contributed to a negative impact on the overall policy impact

Footnote 1 (continued)

assumption that policy making can positively influence the development and deployment of AI technology.

[2] This paper perceives AI's ethical and societal concerns to be closely intertwined, and as such refers to the broader set of these actual and potential concerns throughout.

of the community and two cases where a collaborative effort yielded a positive impact on policy formulation.

2.1.1 Nanotechnology

One community that arguably suffered from a public pursuit of fractious division was the nanotechnology community in the early 2000s [42, 43]. Internal disagreements came to a head in the 2003 'Drexler-Smalley' debate [44], which cemented an oversimplified public caricature of the field. Scholars reviewing this incident have argued that 'para-scientific' media created "polarizing controversy that attracted audiences and influenced policy and scientific research agendas. […] bounding nanotechnology as a field-in-tension by structuring irreconcilable dichotomies out of an ambiguous set of uncertainties." [38]. This showcases a missed opportunity within a fragmented community to come together to promote greater political engagement with the responsible development of the technology.

2.1.2 Recombinant DNA

In the 1970s, concerns arose over recombinant DNA (rDNA) technology. In particular, the ethical implications of the ability to reshape life, as well as fears over potential biohazards from new infectious diseases led the biotechnology community to come together at the 1975 Asilomar Conference on Recombinant DNA to set shared standards [45]. The conference is widely considered a landmark in the field [46]: the scientist's and lawyers' commitment to a forthright open and public discussion has been argued to have stimulated both public interest and grounded policymaker discussion about the social, political and environmental issues related to genetic biotechnology in medicine and agriculture [47].

2.1.3 Ballistic missile defense arms control

In the wake of the creation of the atom bomb, a number of scientists expressed dismay and sought to institutionalize global control of these weapons. Early efforts to this end, such as the 1946 Baruch Plan, proved unsuccessful [48, 49]. However, by the 1950s–1960s, a new 'epistemic community' emerged, bringing together both technical and social scientists, specifically in opposition to the development of anti-ballistic missile (ABM) systems. This community proved able to develop and disseminate this new understanding of nuclear deterrence dynamics to policymakers [50]. They achieved this by maintaining a high level of consensus on concrete policy goals, by framing public discourse on the ethical goals, and by fostering links to both policymakers as well as to Soviet scientists. This allowed them to persuade key administration figures and shift policymaker norms and perceptions at home and internationally. Ultimately, setting the stage for the 1972 ABM Treaty, the first arms control agreement of this kind [50, 51].

2.2 The pitfalls of fragmented efforts in AI Policy

While some of the historical context is surely different, a number of these historical dynamics may well transfer to the emerging AI ethics and policy community [52]. Those concerned with AI policy could benefit from exploring such historical lessons. This should be done with urgency, for two reasons.

First, there is a *closing window of opportunity*. The field of AI policy is a relatively new one, which offers a degree of flexibility in terms of problem framings, governance instrument choice and design, and community alignment. Going forwards, however, this field has a high likelihood of becoming progressively more rigid as framings, public perceptions, and stakeholder interests crystallize. Current dynamics could, therefore, have far-reaching impacts, given the potential to lock in a range of path-dependencies, for example through particular framings of the issues at hand [53]. In this context, a divided community which potentially treats policymakers or public attention as a zero-sum good for competing policy projects may compromise the legitimacy of its individual efforts in front of these. This could undercut the leverage of policy initiatives today and in the future. Worse, public quarrels or contestation may 'poison the well'. Policymakers may begin to perceive and treat a divided community as a series of interest groups rather than an 'epistemic community' with a multi-faceted but coherent agenda for beneficial societal impact of AI. Finally, from a policy perspective, it is important to note that while current regulatory initiatives are and should not always be directly transferable to future issues, neither are they categorically irrelevant. As such, they can often provide the second-best tools for rapidly confronting new AI challenges. This has its own pitfalls, but is often superior to waiting out the slow and reactive formulation of new policies.

Moreover, the *risks are concrete and timely*. It is plausible that political moods will shift within the coming years and decades, in ways that make policy progress much harder. Furthermore, it is possible that other epistemic communities may converge and mobilize faster to embed and institutionalize alternative, less broadly beneficial framings of AI. Indeed, public and global framings of AI in recent years have seemed to drift towards narratives of competition and 'arms races' [54–56, but see also 57]. An inflection point for how societies use and relate to AI may eventually be reached. Missing such a window of opportunity could mean that the relative influence of those concerned with making the impact of AI beneficial (whether in the near or longer term) will decline, right as their voices are needed most.

Conversely, many gains secured today could have lasting benefits down the road.

3 Examining potential grounds for divergence

There are a range of factors that could contribute to the clustering into fuzzy 'near-' and 'long-term' communities, and different scholars may hold distinct and overlapping sets of beliefs on them [cf. 8]. In the following paragraphs, we provide our first attempt at mapping some of these factors.[3]

Some part of the divergence may be due to varying *epistemic* or *methodological* commitments. These could reflect varying levels of tolerance regarding scientific uncertainty and distinct views on the threshold of probability required before far-reaching action or further investigation is warranted. This means that concerns surrounding AI may depend on qualitatively different conceptions of 'acceptable uncertainty' for each group of observers. This may well be hard to resolve. Moreover, epistemic differences over the implicit or explicit disagreements of the modal standards in these debates, for example, debates over what types of data or arguments are admissible in establishing or contesting the plausibility or probability of risk from AI may contribute to further divergence. This could even lead to differential interpretations of evidence that are available. For instance, do empirically observed failure modes of present-day architectures [58–61] provide small-scale proof-of-concepts of the type of difficulties we might one day encounter in AI 'value alignment', or are such extrapolations unwarranted?

For our purposes, however, the most salient factor may be essentially *pragmatic*. Different perceptions of the empirical dynamics and path-dependencies of governing AI can inform distinct theories-of-change. These are intertwined with one's expectations about the tractability and relevance of formulating useful and resilient policy action today. In this context, Prunkl and Whittlestone [8] have recently argued that a more accurate picture and more productive dialogue could be achieved if scholars differentiated amongst the four dimensions on which views vary, in terms of the capabilities, impacts, certainty or extremity of AI systems. They emphasize that views on each of these questions fall on a spectrum. Taking this point on board, there are additional ways to cash out possible divergences. One debate might concern the question, *how long-lasting are the consequences of near-term AI issues?* If those that care about the longer term are convinced that these issues will not have long-lasting consequences, or that they would eventually be swamped by the much larger trends and issues [3], then this could lead them to discount work on near-term AI problems. However, it is important to note that near-term issues are likely to considerably affect the degree to which society is vulnerable to longer-term dangers posed by future advanced AI systems. Short-term or medium-term issues [7, 37] can easily increase society's general turbulence [62], or lock in counterproductive framings of AI or our relation to it. In general, we might expect many nominally near-term effects of AI on society (such as in surveillance; job automation; military capabilities) to scale up and become more disruptive as AI capabilities gradually increase [18, 37]. Indeed, some longer-term scholars have argued that advanced AI capabilities considerably below the level of HLMI might already suffice to achieve a 'prepotence' which could pose catastrophic risks [10]. This would make mid-term impacts particularly important to handle, and collaboration between different groups on at least some generalizable projects crucial.

Another pragmatic question or concern is over *how much leverage we have today to meaningfully shape policies* that will be applicable or relevant in the long term, especially if AI architectures or the broader political and regulatory environment change a lot in the meantime [8]. Some scholars may hold that future AI systems will be technically so different from today's AI architectures that research into this question undertaken today will not be relevant, or they might hold that such advanced AI capabilities may be so remote that the regulatory environment will have changed too much for meaningful policy work to be conducted right now [63]. These people might argue that we had better wait until things are clearer and we are in a better position to understand whether and what research is needed or meaningful.

In practice, this critique does not appear to be a very common or deeply held position. Indeed, as a trade-off it may be overstated. It is plausible that there are diverse areas on which both communities can undertake valuable research today, because the shelf life of current policy and research efforts might be longer than is assumed. To be sure, there is still significant uncertainty over whether current AI approaches can at all be scaled up to very advanced performance [64–66]. Nonetheless, research could certainly depart from a range of areas of overlap [67] and shared areas of concern [68, 69], as we will discuss shortly.

Moreover, policy making is informed by a variety of aspects which range across different time spans. Starting with political agendas that often reflect the current status quo, policy making is equally shaped by shifting public discourse, societal dynamics and high-impact shocks. The latter factor has played a key role in AI policy, where high-profile incidents involving algorithmic discrimination, lack

[3] It should be emphasized that this mapping is only an indicative sketch, and would be much enriched by further examination, for example through structured interviews or comprehensive opinion surveys.

of transparency, or surveillance have driven policy shifts, as seen for example in the pushback on discriminatory algorithms used in the UK visa selection processes [70], the Artificial Intelligence Video Interview Act regulating the use of AI in employee interviews, or the California B.O.T. Law requiring AI systems to self-identify [71, 72].

In sum, it is plausible that many perceived 'barriers' to inter-community cooperation on policy are not all that strong, and that many 'tradeoffs' are likewise overemphasized. However, does that mean there are also positive, mutually productive opportunities for both communities to work on with regard to policy in spite of outstanding disagreements? What would such an agreement look like?

4 Towards 'incompletely theorized agreements' for AI policy

Above, we have reviewed potential sources for divergence within the community. We will now discuss how even in the context of apparent disagreement, pragmatic agreements on shared policy goals and norms could be reached.

We propose to adopt and adapt the legal principle of an 'incompletely theorized agreement' for this purpose. Legal scholarship in constitutional law and regulation has long theorized the legal, organizational and societal importance of such incompletely theorized agreements. Their key use is that they allow a given community to bypass or suspend [73, 74] any theoretical disagreement on matters where (1) the disagreement appears relatively intractable and (2) there is an urgent need to address certain shared practical issues. Disagreements are intractable in cases where either it simply does not appear as if the question *will be decisively resolved* one way or the other in the near term, or where there is *limited time and capacity to reason* through all underlying disagreements [73]. Incompletely theorized agreements can therefore apply to deep philosophical and ethical questions as much as to contexts of pervasive scientific uncertainty. The latter is especially the case on questions where it still remains unclear where and how we might procure the information that allows definitive resolution.

Incompletely theorized agreements are a fundamental component to well-functioning legal systems, societies, and organizations. They allow for stability and flexibility to get urgent things done [75]. These agreements have long played a key role in constitutional and administrative law, and have made possible numerous landmark achievements of global governance, such as the establishment of the Universal Declaration of Human Rights [75, 76]. The framework has also been extended to other domains, such as the collective development and analysis of health-care policies in the face of pluralism and conflicting views [77].

Incompletely theorized agreements have broad similarities with the notion of an 'overlapping consensus', developed by John Rawls, which refers to the way adherents of different (and apparently inconsistent) normative doctrines can nonetheless converge on particular principles of justice to underwrite the shared political community [78]. This concept has been read as a key mechanism in the field of bioethics, serving to enable agreement despite different fundamental outlooks [79]. It also already plays a role in the existing literature on computer ethics [80], as well as in the field of intercultural information ethics [81]. Indeed, overlapping consensus has been proposed as a mechanism on which to ground global cooperation on AI policy across cultural lines [82].

If overlapping consensus can ground inter-cultural cooperation, incompletely theorized agreements might serve as a similar foundation for practical cooperation between near- and long-term perspectives. In a related context, Baum has suggested that policy interventions aimed at securing long-term resilience to various catastrophes can often involve significant co-benefits in the near term, and so do not narrowly depend on all parties agreeing on the deep reasons for the policies proposed [83]. Could incompletely theorized agreements ground cooperation amongst AI policy communities? We suggest that they could.

4.1 Incompletely theorized agreements in AI policy: examples and sketches

There are a range of issue areas where both groups could likely locate joint questions they would want addressed, and shared goals for which particular AI policies should be implemented. This holds even if their underlying reasons for pursuing these are not fully aligned. Without intending to provide an exhaustive, in-depth or definitive overview, a brief survey might highlight various areas for cooperation.

For one, *gaining* insight *into- and leverage on the general levers of policy formation around AI* [52] is a key priority. What are the steps in the policymaking process which determine what issues get raised to political agendas and eventually acted upon, and which might be derailed by other coalitions [84]? Given the above, research into underlying social and societal developments is fruitful to advance all groups' ability to navigate mutually agreeable policy goals across this policy making cycle [85]. Likewise, research into when, where or why global AI governance institutions might become vulnerable to regulatory capture or institutional path dependency ought to be an important consideration, whatever one's AI concerns are [86, 87].

On a more operational level, this can feed into joint investigation into the relative efficacy of various policy levers for AI governance. For example, insights into when and how AI research labs or individual researchers adopt, or alternately cut corners on, responsible and accountable AI, or the incentivization of shifts in workplace culture or employee norms, could shape the policy proposals the community might make. The question of how to promote prosocial norms in AI research environments is of core interest to both communities with an eye to technology policy [88]. This might cover, e.g. whether to publicly name problematic performance (e.g. biased results; lack of safety) in commercial AI products results in tech companies actually correcting the systems [89]; or whether codes of ethics are effective at changing programmers' decision making on the working floor [90, 91]. All of these could be fruitful areas of collaboration on eventual policy proposals for either community.

More specifically, there are a range of particular AI policy programs that we expect could be the site of an incompletely theorized agreement.

1. Incompletely theorized agreements could shape norms and policy debates over the *appropriate scientific culture for considering the impact and dissemination of AI research* [92, 93], especially where it concerns AI applications with potentially salient misuses. The underlying reasons for such policies might differ. Some may be concerned over abuses of vulnerable populations, new vectors for criminal exploitation or the implications of new language models for misinformation [94, 95]; and others over the long-term risks from the eventual open development of advanced AI systems [96]. Accordingly, incompletely theorized agreements in this area could converge on policies to shape researcher norms around improved precaution or reflection around the impact or potential misuse of research [97].
2. Another example might be found in the domain of the *global regulation of military uses of AI*. This area has already seen years of shared efforts and even collaboration amongst a coalition of activists, lawyers, and institutions departing from both a near-term as well as longer-term perspective, such as the Future of Life Institute [52].
3. Incompletely theorized agreements could ground productive policy cooperation on policy interventions aimed at *preserving the integrity of public discourse and informed decision-making in the face of AI systems*. Policies aimed at combating AI-enabled disinformation would be a natural site for incompletely theorized collaboration, because a society's epistemic security [98] is relevant from both a near-term and long-term perspective alike.
4. Similarly, incompletely theorized agreements surrounding the promotion of policies aimed at *securing citizens' (political) autonomy and independence from unaccountable perception control* could be promising. After all, practices of opaque technological management [99] or perception control [100] can enable authorities to increasingly shape individuals' and societies' behaviour and values. Regulation to restrict the deployment of such tools, to facilitate privacy-preserving AI, or to ensure transparency and accountability of the principals of such tools, are important from a near-term perspective concerned with the role of 'hyper nudges' [101], algocracy [102], or surveillance capitalism [103]. Simultaneously, such policies are also critical to avert long-term worries over a 'value lock-in', whereby one generation or party might someday "invent a technology that will enable the agents alive at that time to maintain their values indefinitely into the future, controlling the broad sweep of the entire rest of the future of civilisation" [104].

Although the underlying motives for each group to pursue policies on the abovementioned domains may be partially distinct, these technical differences are arguably thwarted by the benefits derived from achieving impactful and effective policy measures together. In many of these cases, the practical benefits of an incompletely theorized agreement would be at least fourfold (1) to reduce public confusion around these topics; (2) to present policymakers with an epistemic community delivering integrated policy proposals; (3) to support the articulation of regulations or governance instruments for specific policy problems, which need not assume further advances in AI capabilities, but which are also not reliant on provisions or assumptions that are vulnerable to 'obsolescence' if or when such advances do occur [105–107], giving such policies a longer shelf life; (4) to improve engagement of particular AI policies with the steadily increasing cross-domain nature of AI, which could help inform regulatory responses across domains. This is especially relevant because different fields (such as content moderation, health-care, or the military) often confront different yet similar versions of underlying problems [108].

4.2 Limitations of incompletely theorized agreements

We do not wish to suggest that incompletely theorized agreements are an unambiguously valuable tool across all AI policy cases, or even a definite solution for any one policy case. Such agreements do suffer from a number of potential drawbacks or trade-offs, which both communities should consider before invoking them in any particular case.[4]

First, depending on one's assumptions around the expected degree of change in AI or in its societal impacts, incompletely theorized agreements could prove brittle.

[4] We thank one reviewer for prompting this discussion of the drawbacks of incompletely theorized agreements.

Incompletely theorized agreements bypass a full examination of the underlying disagreements to facilitate pragmatic and swift action on particular policies on which both communities find themselves in practical alignment in a specific moment in time. This may create a lack of clarity over the boundary conditions of that practical agreement, along with opacity over whether, or where (i.e. for which future particular questions around AI policy) the practical agreement might suddenly break down for either or all parties.

Second, an incompletely theorized agreement is, in an important sense, a 'stopgap' measure more than a general ideal or permanent fix. As discussed above, an incompletely theorized agreement might be most suited to situations where (a) practical policy action is urgently needed, and (b) underlying theoretical agreement by stakeholders on all engaged principles or questions does not seem close. However, over longer timeframes, deeper inquiry and debate do appear necessary [73]. In addition, there is always the possibility that agreement was not, in fact, intractable within the near term. As such, a premature leap by the community into an incompletely theorized agreement to achieve some policy X might inadvertently curb the very conversations amongst the communities which could have led both to eventually prefer policy Y instead, had their conversation been allowed to run its course.

Moreover, there is a key related point here, on which we should reflect. By advocating for the adoption of incompletely theorized agreements on AI policy today, we ourselves are in a sense assuming or importing an implicit judgment about the urgency of AI issues today, and about the intractability of the underlying debates. Yet these are two positions which others might contest. For example, by arguing that 'AI issues do not today meet that threshold of urgency that the use of an incompletely theorized agreement is warranted'. We wish to make this assumption explicit. At the same time, we expect that it is an assumption widely shared by many scholars working on AI policy, many of whom may well share a sense that the coming years will be a sensitive and even critical time for AI policies.

Third, a sloppily formulated incompletely theorized agreement on an AI policy issue may not actually reflect convergence on particular policies (e.g. 'certification scheme for AI products with safety tests X, Y, Z'). Instead, it might solidify on apparent agreement on vague mid-level principles or values (e.g. 'AI developers should ensure responsible AI development'). These may be so broad that they do not ground clear action at the level of actual policies. If this were to happen, incompletely theorized agreements might merely risk contributing to the already-abundant proliferation of broad AI principles or ethical frameworks on AI that have little direct policy impact. While the ecosystem of AI codes of ethics issued in recent years have certainly shown some convergence [109–111], they have been critiqued as being hard to operationalize, and for providing only the appearance of agreement while masking underlying tensions in the principles' interpretation, operationalization, or practical requirements [93, 112]. Situations where an incompletely theorized agreement does not manage to root itself at the level of concrete policies but only mid-level principles would be a worst-of-both-worlds scenario: it would reduce the ability of actors to openly reflect upon and resolve inconsistencies amongst- or disagreements about high-level principles, while not even affording improvements at facilitating concrete policies or actions in particular AI domains. To mitigate this risk, incompletely theorized agreements should, therefore, remain closely grounded in concrete and clearly actionable policy goals or outputs.

Nonetheless, while limitations such as these should be considered in greater detail, we argue that they do not categorically erode the case for implementing, or at least further examining the promise of this principle and tool for advancing responsible AI policy.

5 Conclusion

AI has raised multiple societal and ethical concerns. This highlights the urgent need for suitable and impactful policy measures in response. Nonetheless, there is at present an experienced fragmentation in the responsible AI policy community, amongst clusters of scholars focusing on 'near-term' AI risks, and those focusing on 'longer-term' risks. This paper has sought to map the practical space for inter-community collaboration, with a view towards the practical development of AI policy.

As such, we briefly provided a rationale for such collaboration, by reviewing historical cases of scientific community conflict or collaboration, as well as the contemporary challenges facing AI policy. We argued that fragmentation within a given community can hinder progress on key and urgent policies. Consequently, we reviewed a number of potential (epistemic, normative or pragmatic) sources of disagreement in the AI ethics community, and argued that these trade-offs are often exaggerated, and at any rate do not need to preclude collaboration. On this basis, we presented the novel proposal for drawing on the constitutional law principle of an 'incompletely theorized agreement', for the communities to set aside or suspend these and other disagreements for the purpose of achieving higher-order AI policy goals of both communities in selected areas. We, therefore, non-exhaustively discussed a number of promising shared AI policy areas which could serve as the sites for such agreements, while also discussing some of the overall limits of this framework.

This paper does not suggest that communities should fully merge or ignore differences whatever their source may be. To be sure, some policy projects will be relevant to one group within the community but not the other. Indeed, community heterogeneity and diversity is generally a good thing for a scientific paradigm. Instead, the paper proposes to question some possible reasons for conflicting dynamics which could stall positive progress for policy making, and suggests an avenue for a higher-order resolution. Most of all, the paper hopes to pragmatically encourage the exploration of opportunities for shared work and suggested that work on such opportunities, where it is found, can be well grounded through an incompletely theorized agreement. We invite scholars in the ethical AI community to explore the strengths and limits of this tool.

Acknowledgements For valuable comments and feedback, the authors would like to thank Seth Baum, Haydn Belfield, Ashwin Acharya, John Danaher, Alex Lintz, Vincent Müller, Carina Prunkl, Rohin Shah and Jess Whittlestone, as well as detailed comments from two anonymous reviewers. Any remaining errors are all our own.

Funding During most of the writing process of this paper, Matthijs Maas was funded under a PhD grant by the University of Copenhagen Faculty of Law. During the final stages of re-editing, he was a postdoctoral Research Associate at the Centre for the Study of Existential Risk, University of Cambridge.

Compliance with ethical standards

Conflict of interest Both authors are fellows at the Centre for the Future of Intelligence (University of Cambridge). During the writing of this paper, Charlotte Stix acted as Coordinator of the European Commission's High Level Expert Group on Artificial Intelligence. She does not disclose or use any confidential information they were privy to during that time, and hence does not identify a conflict of interest between that role and her ongoing academic research. Matthijs Maas does not identify a conflict of interest.

References

1. Trajtenberg M. AI as the next GPT: a Political-Economy Perspective. National Bureau of Economic Research; 2018. https://doi.org/10.3386/w24245
2. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O.: Viewpoint: when will AI exceed human performance? Evidence from AI experts. J Artif Intell Res. **62**, 729–754 (2018)
3. Gruetzemacher R, Whittlestone J. Defining and unpacking transformative AI. 2019. Available: http://arxiv.org/abs/1912.00747
4. Gruetzemacher R, Whittlestone J. The transformative potential of artificial intelligence. Commun ACM. 2020.
5. Goertzel B, Pennachin C, editors. Artificial general intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007.
6. Baum, S.D.: Reconciliation between factions focused on near-term and long-term artificial intelligence. AI Soc. **33**, 565–572 (2018)
7. Baum, S.D.: Medium-term artificial intelligence and society. Information. **11**, 290 (2020)
8. Prunkl C, Whittlestone J. Beyond near- and long-term: towards a clearer account of research priorities in AI ethics and society. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York NY USA: ACM; 2020. pp. 138–143.
9. Cave, S., ÓhÉigeartaigh, S.S.: Bridging near- and long-term concerns about AI. Nat Mach Intell **1**, 5–6 (2019)
10. Critch A, Krueger D. AI research considerations for human existential safety (ARCHES). 2020. Available: http://acritch.com/arches/
11. Drexler KE. Reframing superintelligence: comprehensive AI services as general intelligence. Oxford: Future of Humanity Institute, University of Oxford; 2019 Jan p. 210. Report No.: 2019–1. Available: https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf
12. Christiano P. Prosaic AI alignment. In: AI alignment [Internet]. 19 Nov 2016 [cited 2 Sep 2020]. Available: https://ai-alignment.com/prosaic-ai-control-b959644d79c2
13. Clark J. Import AI #83: Cloning voices with a few audio samples, why malicious actors might mess with AI, and the industry-academia compute gap. In: Import AI [Internet]. 26 Feb 2018 [cited 23 Jul 2018]. Available: https://jack-clark.net/2018/02/26/import-ai-83-cloning-voices-with-a-few-audio-samples-why-malicious-actors-might-mess-with-ai-and-the-industryacademia-compute-gap/
14. Floridi, L., Cowls, J., King, T.C., Taddeo, M.: How to design AI for social good: seven essential factors. Sci Eng Ethics. (2020). https://doi.org/10.1007/s11948-020-00213-5
15. Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling climate change with machine learning. 2019. Available: http://arxiv.org/abs/1906.05433
16. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, et al. The role of artificial intelligence in achieving the sustainable development goals. 2019. Available: http://arxiv.org/abs/1905.00501
17. Calo R. Artificial Intelligence Policy: A Primer and Roadmap. 2017;51: 37
18. Dafoe A. AI Governance: A Research Agenda. 2018; 52.
19. Müller VC. Ethics of artificial intelligence and robotics. In: Zalta EN, editor. Stanford Encyclopedia of Philosophy. Palo Alto: CSLI, Stanford University; 2020.
20. Barocas S, Selbst AD. Big Data's Disparate Impact. Calif Law Rev. 2016;671. Available: https://papers.ssrn.com/abstract=2477899
21. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of Machine Learning Research. 2018. p. 15.
22. Doran D, Schulz S, Besold TR. What does explainable AI really mean? A new conceptualization of perspectives. 2017 [cited 9 Oct 2017]. Available: http://arxiv.org/abs/1710.00794
23. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability

of machine learning. 2019. Available: http://arxiv.org/abs/1806.00069

24. Anderson JM, Kalra N, Stanley K, Sorensen P, Samaras C, Oluwatola TA. Autonomous vehicle technology: a guide for policymakers. RAND Corporation; 2016. Available: https://www.rand.org/pubs/research_reports/RR443-2.html
25. Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. 2018 [cited 21 Feb 2018]. Available: http://arxiv.org/abs/1802.07228
26. King, T.C., Aggarwal, N., Taddeo, M., Floridi, L.: Artificial intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions. Sci. Eng. Ethics. (2018). https://doi.org/10.1007/s11948-018-00081-0
27. Hayward, K.J., Maas, M.M.: Artificial Intelligence and crime: a primer for criminologists. Crime Med Cult. (2020). https://doi.org/10.1177/1741659020917434
28. Helbing D, Frey BS, Gigerenzer G, Hafen E, Hagner M, Hofstetter Y, et al. Will democracy survive big data and artificial intelligence? Sci Am. 2017. Available: https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/. Accessed 29 May 2017.
29. Nemitz, P.: Constitutional democracy and technology in the age of artificial intelligence. Philos Trans A Math Phys Eng Sci. **376**, 20180089 (2018)
30. Chesney R, Citron DK. Deep Fakes: A looming challenge for privacy, democracy, and national security. Calif Law Rev. 2019;107. Available: https://papers.ssrn.com/abstract=3213954
31. Raso F, Hilligoss H, Krishnamurthy V, Bavitz C, Kim L. Artificial intelligence and human rights: opportunities and risks. In: Berkman Klein Center for Internet and Society at Harvard University; 2018. Available: https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf?
32. Molnar, P.: Technology on the margins: AI and global migration management from a human rights perspective. Camb Int Law J. **8**, 305–330 (2019)
33. Feldstein, S.: The road to digital unfreedom: how artificial intelligence is reshaping repression. J Demo. **30**, 40–52 (2019)
34. Danzig R. An irresistible force meets a moveable object: the technology tsunami and the liberal world order. Lawfare Research Paper Series. 2017;5. Available: https://assets.documentcloud.org/documents/3982439/Danzig-LRPS1.pdf
35. Bostrom, N.: Superintelligence: paths, dangers. Oxford University Press, Strategies (2014)
36. Russell S. Human compatible: artificial intelligence and the problem of control. Viking; 2019.
37. Parson E, Re R, Solow-Niederman A, Zeide A. Artificial intelligence in strategic context: an introduction. PULSE, UCLA School of Law; 2019. Available: https://aipulse.org/artificial-intelligence-in-strategic-context-an-introduction/
38. Kaplan, S., Radin, J.: Bounding an emerging technology: Para-scientific media and the Drexler–Smalley debate about nanotechnology. Soc Stud Sci. **41**, 457–485 (2011)
39. Russell S, Dafoe A. Yes, the experts are worried about the existential risk of artificial intelligence. In: MIT Technology Review [Internet]. 2 Nov 2016 [cited 26 Feb 2017]. Available: https://www.technologyreview.com/s/602776/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/
40. Baum, S.D.: Countering superintelligence misinformation. Information. **9**, 244 (2018)
41. Future of Life Institute. AI Safety Myths. In: Future of Life Institute [Internet]. 2016 [cited 26 Oct 2017]. Available: https://futureoflife.org/background/aimyths/
42. Selin, C.: Expectations and the emergence of nanotechnology. Sci Technol Human Values. **32**, 196–220 (2007)
43. Shew, A.: Nanotech's History: An Interesting, Interdisciplinary Ideological Split. Bull Sci Technol Soc. **28**, 390–399 (2008)
44. Baum, R., Drexler, K.E., Smalley, R.E.: Point-counterpoint: nanotechnology. Chem Eng News. **81**, 37–42 (2003)
45. Berg, P., Baltimore, D., Brenner, S., Roblin, R.O., Singer, M.F.: Summary statement of the Asilomar conference on recombinant DNA molecules. Proc Natl Acad Sci USA **72**, 1981–1984 (1975)
46. Grace K. The Asilomar conference: a case study in risk mitigation. Berkeley, CA: Machine Intelligence Research Institute; 2015 Jul. Report No.: 2015–9. Available: https://intelligence.org/files/TheAsilomarConference.pdf
47. Berg, P., Singer, M.F.: The recombinant DNA controversy: Twenty years later. Proc Natl Acad Sci USA **92**, 9011–9013 (1995)
48. Baratta, J.P.: Was the Baruch plan a proposal of world government? Int Hist Rev. **7**, 592–621 (1985)
49. Bartel, F.: Surviving the years of grace: the atomic bomb and the specter of world government, 1945–1950. Diplom His **39**, 275–302 (2015)
50. Adler, E.: the emergence of cooperation: national epistemic communities and the international evolution of the idea of nuclear arms control. Int Organ. **46**, 101–145 (1992)
51. Maas, M.M.: How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. Contemp Secur Policy. **40**, 285–311 (2019)
52. Belfield H. Activism by the AI community—analysing recent achievements and future prospects. In: Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics and Society 2020. 2020.
53. Cave S, Ó hÉigeartaigh SS. An AI race for strategic advantage: rhetoric and risks. In: AAAI/ACM Conference on Artificial Intelligence, Ethics and Society. 2018. Available: http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf
54. Lee, K.-F.: AI superpowers: china, silicon valley, and the new world order. Houghton Mifflin Harcourt, Boston (2018)
55. Thompson N, Bremmer I. The AI cold war that threatens us all. Wired. 2018. Available: https://www.wired.com/story/ai-cold-war-china-could-doom-us-all/
56. Auslin M. Can the Pentagon Win the AI Arms Race? Foreign Aff. 2018. Available: https://www.foreignaffairs.com/articles/united-states/2018-10-19/can-pentagon-win-ai-arms-race
57. Imbrie A, Dunham J, Gelles R, Aiken C. Mainframes: a provisional analysis of rhetorical frames in AI. In: Center for security and emerging technology; 2020. Available: https://cset.georgetown.edu/research/mainframes-a-provisional-analysis-of-rhetorical-frames-in-ai/
58. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. 2016 [cited 13 May 2017]. Available: http://arxiv.org/abs/1606.06565
59. Krakovna V, Uesato J, Mikulik V, Rahtz M, Everitt T, Kumar R, et al. Specification gaming: the flip side of AI ingenuity. Deepmind. 2020. Available: https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity
60. Kumar RSS, Brien DO, Albert K, Viljöen S, Snover J. Failure modes in machine learning systems. arXiv [cs.LG]. 2019. Available: http://arxiv.org/abs/1911.11034
61. Turner AM. Optimal farsighted agents tend to seek power. arXiv [cs.AI]. 2019. Available: http://arxiv.org/abs/1912.01683
62. Bostrom N, Dafoe A, Flynn C. Public policy and superintelligent AI: a vector field approach. In: Liao SM, editor. Ethics of artificial intelligence. Oxford University Press; 2020.
63. Brooks R. The seven deadly sins of predicting the future of AI. 7 Sep 2017 [cited 13 Sep 2017]. Available: http://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/

64. Sutton R. The bitter lesson. 2019. Available: http://www.incompleteideas.net/IncIdeas/BitterLesson.html
65. Brooks R. A Better lesson. 2019. Available: https://rodneybrooks.com/a-better-lesson/
66. Marcus G. Deep learning: a critical appraisal. 2018. Available: http://arxiv.org/abs/1801.00631
67. Hernandez-Orallo, J., Martınez-Plumed, F., Avin, S., Whittlestone, J.: AI paradigms and AI safety: mapping artefacts and techniques to safety issues, p. 8. Santiago de Compostela, Spain (2020)
68. Manheim D, Garrabrant S. Categorizing variants of goodhart's law. 2018. Available: http://arxiv.org/abs/1803.04585
69. Thomas R, Uminsky D. The problem with metrics is a fundamental problem for AI. arXiv [cs.CY]. 2020. Available: http://arxiv.org/abs/2002.08512
70. McDonald H. Home Office to scrap "racist algorithm" for UK visa applicants. The Guardian. 4 Aug 2020. Available: http://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants. Accessed 2 Sep 2020.
71. Illinois general assembly—full text of HB2557. 2019 [cited 2 Sep 2020]. Available: https://www.ilga.gov/legislation/fulltext.asp?DocName=&SessionId=108&GA=101&DocTypeId=HB&DocNum=2557&GAID=15&LegID=&SpecSess=&Session=
72. SB-1001 Bots: disclosure. In: California Legislative Information [Internet]. 2018 [cited 2 Sep 2020]. Available: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001
73. Sunstein, C.R.: Incompletely theorized agreements. Harv Law Rev. **108**, 1733–1772 (1995)
74. Sunstein, C.R.: Incompletely theorized agreements in constitutional law. Soc Res. **74**, 1–24 (2007)
75. Sunstein CR. Holberg Prize 2018, Acceptance Speech. Holberg Prize 2018; 2018; Bergen, Norway. Available: https://www.holbergprisen.no/en/cass-sunsteins-acceptance-speech
76. Taylor C. Conditions of an unforced consensus on human rights. Bangkok; 1996. Available: https://www.iilj.org/wp-content/uploads/2016/08/Taylor-Conditions-of-an-Unforced-Consensus-on-Human-Rights-1996.pdf
77. Ruger, J.P.: Pluralism, incompletely theorized agreements, and public policy. Health and Social Justice. Oxford University Press, Oxford (2009)
78. Rawls J. Political Liberalism. Columbia University Press; 1993.
79. Benjamin M. The value of consensus. Society's choices: social and ethical decision making in biomedicine. National Academy Press; 1995.
80. Søraker, J.H.: The role of pragmatic arguments in computer ethics. Ethics Inf Technol. **8**, 121–130 (2006)
81. Hongladarom, S.: Intercultural Information Ethics: a pragmatic consideration. In: Kelly, M., Bielby, J. (eds.) Information cultures in the digital age: a festschrift in honor of rafael capurro, pp. 191–206. Springer Fachmedien, Wiesbaden (2016)
82. ÓhÉigeartaigh, S.S., Whittlestone, J., Liu, Y., Zeng, Y., Liu, Z.: Overcoming barriers to cross-cultural cooperation in AI ethics and governance. Philos Technol. (2020). https://doi.org/10.1007/s13347-020-00402-x
83. Baum, S.D.: The far future argument for confronting catastrophic threats to humanity: practical significance and alternatives. Futures. **72**, 86–96 (2015)
84. Perry, B., Uuk, R.: AI governance and the policymaking process: key considerations for reducing AI risk. Big Data and Cogn Comput. **3**, 26 (2019)
85. Hallsworth M, Parker S, Rutter J. Policymaking in the real world: evidence and analysis. Institute for Government; 2011 Apr. Available: https://www.instituteforgovernment.org.uk/sites/default/files/publications/Policy%20making%20in%20the%20real%20world.pdf
86. Cihon P, Maas MM, Kemp L. Should artificial intelligence governance be centralised?: Design lessons from history. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: ACM; 2020. pp. 228–234.
87. Jelinek, T., Wallach, W., Kerimi, D.: Policy brief: the creation of a G20 coordinating committee for the governance of artificial intelligence. AI Ethics. (2020). https://doi.org/10.1007/s43681-020-00019-y
88. Baum SD. On the promotion of safe and socially beneficial artificial intelligence. AI Soc. 2016 [cited 13 May 2017]. doi:https://doi.org/10.1007/s00146-016-0677-0
89. Raji ID, Buolamwini J. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. 2019. p. 7.
90. McNamara A, Smith J, Murphy-Hill E. Does ACM's code of ethics change ethical decision making in software development? In: Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018. Lake Buena Vista, FL, USA: ACM Press; 2018. pp. 729–733.
91. Cleek, M.A., Leonard, S.L.: Can corporate codes of ethics influence behavior? J Bus Ethics. **17**, 619–630 (1998)
92. Shevlane T, Dafoe A. the offense-defense balance of scientific knowledge: does publishing ai research reduce misuse? In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery; 2020. pp. 173–179.
93. Whittlestone J, Ovadya A. The tension between openness and prudence in AI research. 2020. Available: http://arxiv.org/abs/1910.01170
94. Solaiman I, Brundage M, Clark J, Askell A, Herbert-Voss A, Wu J, et al. Release strategies and the social impacts of language models. 2019. Available: http://arxiv.org/abs/1908.09203
95. McGuffie K, Newhouse A. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. Middlebury Institute of International Studies; 2020 Sep p. 13. Available: https://www.middlebury.edu/institute/academics/centers-initiatives/ctec/ctec-publications-0/radicalization-risks-gpt-3-and-neural
96. Bostrom N. Strategic Implications of Openness in AI Development. Glob Policy. 2017 [cited 18 Feb 2017]. https://doi.org/10.1111/1758-5899.12403
97. Ashurt C, Anderljung M, Prunkl C, Leike J, Gal Y, Shevlane T, et al. A guide to writing the NeurIPS impact statement. Medium. 2020. Available: https://medium.com/@operations_18894/a-guide-to-writing-the-neurips-impact-statement-4293b723f832
98. Seger E, Avin S, Pearson G, Briers M, Ó hÉigeartaigh S, Bacon H. Tackling threats to informed decision-making in democratic societies: promoting epistemic security in a technologically-advanced world. The Alan Turing Institute; 2020. Available: https://www.turing.ac.uk/research/publications/tackling-threats-informed-decision-making-democratic-societies
99. Brownsword, R.: In the year 2061: from law to technological management. Law Innov Technol **7**, 1–51 (2015)
100. Susskind J. Future Politics: Living Together in a World Transformed by Tech. Oxford; New York: Oxford University Press; 2018.
101. Yeung, K.: "Hypernudge": Big data as a mode of regulation by design. Inf Commun Soc. **20**, 118–136 (2017)
102. Danaher J. The Threat of Algocracy: Reality, resistance and accommodation. Philos Technol. 9/2016;29: 245–268.

103. Zuboff, S.: The age of surveillance capitalism: the fight for a human future at the new frontier of power, 1st edn. PublicAffairs, New York (2019)
104. MacAskill W. Are we living at the hinge of history? 2020. Available: https://www.academia.edu/43481026/Are_We_Living_at_the_Hinge_of_History
105. Crootof R. Jurisprudential space junk: treaties and new technologies. In: Giorgetti C, Klein N, editors. Resolving Conflicts in the Law. 2019. pp. 106–129.
106. Maas, M.M.: Innovation-proof governance for military ai? how i learned to stop worrying and love the bot. J Int Hum Legal Stud. **10**, 129–157 (2019)
107. Rosert, E., Sauer, F.: Prohibiting autonomous weapons: put human dignity first. Global Policy. **10**, 370–375 (2019)
108. Crootof R, Ard BJ. Structuring techlaw. Harv J Law Technol. 2021;34. Available: https://papers.ssrn.com/abstract=3664124
109. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019; 1–11.
110. Fjeld J, Hilligoss H, Achten N, Daniel ML, Feldman J, Kagay S. Principled artificial intelligence: a map of ethical and rights-based approaches. In: Berkman Klein Center for Internet & Society at Harvard University; 2019 p. 1. Available: https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf
111. Schiff D, Biddle J, Borenstein J, Laas K. What's next for AI ethics, policy, and governance? A global overview. In: Proceedings of the AAAI/ACM conference on AI, Ethics, and Society. New York: ACM; 2020. pp. 153–158.
112. Whittlestone J, Nyrup R, Alexandrova A, Cave S. The Role And Limits Of Principles in AI Ethics: towards a focus on tensions. Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics and Society 2019. 2019. p. 7.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.