Legal Priorities Research: A Research Agenda

Christoph Winter, Jonas Schuett, Eric Martínez,
Suzanne Van Arsdale, Renan Araújo, Nick Hollman,
Jeff Sebo, Andrew Stawasz, Cullen O’Keefe, Giuliana Rotola

Legal Priorities
Project

Legal Priorities Research: A Research Agenda

Christoph Winter[1,2,3], Jonas Schuett[1,4], Eric Martínez[1,5],
Suzanne Van Arsdale[1], Renan Araújo[1], Nick Hollman[1],
Jeff Sebo[6], Andrew Stawasz[3], Cullen O'Keefe[1,7,8], Giuliana Rotola[9,10]

[1] *Legal Priorities Project*
[2] *Instituto Tecnológico Autónomo de México (ITAM)*
[3] *Harvard University*
[4] *Goethe University Frankfurt*
[5] *Massachusetts Institute of Technology (MIT)*
[6] *New York University (NYU)*
[7] *OpenAI*
[8] *University of Oxford*
[9] *Open Lunar Foundation*
[10] *Foresight Institute*

January 2021

We are particularly grateful for valuable comments, feedback and support from Alfredo Parra, Mark Eccleston-Turner, Leonie Koessler, Sean Richardson, Tyler John, Olavo Bittencourt, Lisa Forsberg, Kevin Tobia, Bradly Condon, Markus Anderljung, Seth Baum, Tobias Baumann, Haydn Belfield, Alexis Carlier, Devin Mauney, Peter Cihon, Frans von der Dunk, Dov Greenbaum, Ameen Jauhar, David Koplow, Sebastien Krier, David Manheim, Tanja Masson-Zwaan, Cecil Abungu, Gabriel Bankman-Fried, Gregory Lewis, Janvi Ahuja, Sarah Carter, Alasdair Phillips-Robins, Caleb Huffman, and the participants at presentations of the individual parts of this agenda. All views and errors are our own.

Table of Contents

Introduction

Part 1: Foundations of Legal Priorities Research

Part 2: Exploration by Cause Areas

Part 3: Cause Areas for Further Engagement

References

Appendix: Closely Related Areas of Existing Academic Research

Introduction

Humanity has seen relatively stable improvements in quality of life over time. Present generations benefit from the accomplishments of past generations, and future generations benefit from advanced knowledge, economic growth, stronger institutions, and other improved conditions for welfare created by present generations. This trend, however, might change.

Our ever-advancing knowledge, based on the exchange of ideas throughout space and time, has led to technologies that threaten the very existence of future generations. Yet, while humanity has been aware of the first anthropogenic existential threat for some time (the use of nuclear weapons) and is slowly realizing the dangers of climate change, the ongoing COVID-19 pandemic has shown that we are not prepared for some of the greatest threats of this century. For example, although scientific knowledge allowed us to encode the genome of the novel coronavirus within days, and an effective vaccine was discovered shortly thereafter, most national and international institutions have not been able to challenge the spread of the virus effectively.

More deadly and contagious pandemics, natural or engineered, may well pose much greater, possibly even existential threats to the future of humanity. Whether we address these and other risks—such as those resulting from advanced artificial intelligence, runaway climate change, or synthetic biology—will drastically affect the well-being of future generations, so much so that we may be at a very unusual point in history: For the first time, the future of sentient life heavily depends on those in the present. Even more so, its very existence may be at stake during what has been referred to as "the precipice" (Ord, 2020). Although our actions (and inactions) may have historically unique consequences for future generations, their interests are not represented in current political and economic systems, and human intuitions have not yet been updated accordingly. This calls for fundamental legal change.

Given that some of the risks and opportunities to positively shape the lives of countless future individuals are much greater than others, prioritization is of utmost importance. What are the greatest risks and opportunities for humanity, and what is the role that multidisciplinary-informed legal research can take? How can we prioritize so as to increase the chance of a flourishing and long-lasting future of humanity? How can we cooperate most effectively with those whom we will never meet, but whose lives lie in our hands? Choosing to address these questions and prioritizing carefully among them may be one of the great opportunities of our time to positively change the human trajectory, and will be the guiding theme of this agenda.

Part 1 outlines the various empirical and philosophical foundations underlying both our research agenda and legal priorities research generally. In particular, we highlight prioritization efforts as an important and neglected tool for legal scholarship (Section 1) and emphasize the importance of taking into account the long-term consequences of laws and legal research during prioritization (Section 2). Finally, we offer a rigorous yet flexible, and potentially ever-evolving methodological framework for deciding which problems to work on and how to tackle them (Section 3).

In Part 2 of this agenda, we explore a number of specific cause areas in more detail and identify promising research projects within each. We recognize that many of these projects are relatively broad, and further work is often needed to articulate a more specific research question that would naturally correspond to an individual research paper. We also provide an overview of relevant literature at the end of each individual subsection. This Part covers the law and governance of artificial intelligence (Section 4), synthetic biology and biorisk (Section 5), and institutional design (Section 6). Since choosing the right research project is one of the most important factors that determines the impact of legal research, we have also identified a number of meta-research projects (Section 7). Research in this area tackles problems that legal researchers encounter when prioritizing, such as whether to focus on international, comparative, or national law.

Part 3 follows the structure of Part 2. Here, we outline further cause areas that also fit our methodology criteria but for which further research is needed to more precisely compare them with other cause areas. This Part covers space governance (Section 8) and animal law (Section 9). Though we refer to these as cause areas for further engagement, we encourage interested researchers to pursue projects in these fields, both at the meta- and object-level, and may integrate them into our main cause areas in future iterations of this agenda.

Legal priorities research is by its very nature an interdisciplinary affair. We therefore include an appendix which aims to give an overview of some of the most closely related areas of existing literature that are likely to be particularly useful for legal priorities research. This appendix is organized around the general academic disciplines of philosophy (A), economics (B), psychology (C), macrohistory (D), and political science (E). Within each discipline we identify both general examples of interdisciplinary research between law and that respective discipline, as well as more specific research areas within those disciplines.

Identifying the most important research projects is accompanied by high degrees of both normative and empirical uncertainty. Although we develop specific criteria in Section 3 to account for this, a substantial amount of holistic uncertainty remains and must be acknowledged. This leads us even more so to appreciate feedback from the wider community of legal scholars who are interested in prioritization, law, and the long-term future. In fact, it would not have been possible to write this agenda without the helpful feedback and comments from and conversations with various experts in the first place. The transparency of this agenda's philosophical and empirical assumptions in its first Section is very much motivated by the idea of continuing and encouraging a fruitful culture of feedback. This said, the agenda is a common project in a different way as well: We aim at inspiring and encouraging the legal community to take up the outlined challenges. Anyone interested in using the agenda to get ideas and guidance on potential projects should feel free to do so.

Part 1

Foundations of Legal Priorities Research

There are various empirical and philosophical foundations underlying both our research agenda and legal priorities research generally. Here, we present and defend each of these main foundations in turn, including the notion of prioritization as an important and neglected tool for legal scholarship (Section 1), the importance of taking into account long-term consequences as the basis for this prioritization (Section 2), and a rigorous yet flexible (and potentially ever-evolving) methodological framework for deciding which problems to work on and how to best tackle them (Section 3).

1 The Case for Legal Priorities Research

One of the central foundations of our research agenda is the idea that legal scholarship should prioritize among possible research questions using first-principles and evidence-based reasoning. We refer to this practice (as well as the scholarship resulting from it) as *legal priorities research*. In this Section, we introduce and define the concept of legal priorities research, argue for its importance, explain potential reasons for its neglectedness (Section 1.1), and defend it against potential objections (Section 1.2). Note that the purpose of this Section is to give a broad overview (rather than an exhaustive account) of legal priorities research, which we further discuss in later parts of the agenda (in particular, Sections 2, 3, and 7).

1.1 The Importance and Neglectedness of Legal Priorities Research

As alluded to above, legal priorities research involves prioritizing among sets of possible research questions using first-principles and evidence-based reasoning, as well as engaging in and producing scholarship informed by said prioritization. Legal priorities research can also be thought of as a form of prioritization research, which involves applying techniques from philosophy, economics, mathematics, and social science to help individuals and organizations decide (a) which problems they should work on and (b) how best to work on them to do the most good (Stafforini, 2014). The importance of such prioritization arises from the mismatch between the myriad problems in the world and the paucity of resources available to solve them. Given the severity of this mismatch, an actor seeking to improve the world as much as possible must prioritize among both the problems themselves and the means for tackling them. [1] Assuming informal prioritization methods are insufficient, a formal prioritization program becomes necessary. Over the past decade, various formal, altruistically motivated prioritization efforts have emerged within a variety of contexts, including Open Philanthropy, 80,000 Hours, and, most recently, the Global Priorities Institute at the University of Oxford, the first academic institution of its kind to conduct and promote foundational prioritization research on global issues.

While such efforts have sought to address global issues generally, there currently exists no such formal prioritization program in the context of legal academia. Assuming one accepts the importance of prioritization at the global level, however, it naturally follows that legal researchers should apply the same principles to prioritize the projects they undertake. As in the global context, there is a discrepancy between the plethora of issues that the law could address and the limit to the resources available, making it necessary to prioritize among potential problems. Moreover, this need for careful prioritization seems to be acknowledged by those within academia; in a recent survey of law professors, the overwhelming majority responded that they would prefer legal academia to prioritize more carefully than it does currently (Martinez & Winter, 2021). Indeed, given the traditional view of law as a powerful instrument for social change, with the potential to improve (or exacerbate) some of the world's most pressing problems, it may come as a surprise that systematic prioritization of potentially impactful research has not received significant attention within legal academia. Potential explanations for such neglect relate to (a) *methodology*: lagging behind other fields with regard to interdisciplinary and evidence-based standards, (b) *scope*: focusing on national and near-term issues, and (c) *incentive structure*: research topics driven more by publication record than positive impact. For the remainder of this subsection, we discuss each of these possible explanations for neglectedness in turn.

With regard to methodology, legal research in many jurisdictions has been largely, if not strictly, unidisciplinary. For example, in many civil law jurisdictions, where law has traditionally been a very isolated discipline, lawyers and researchers often receive little to no formal training outside of the law and consequently tend to lack (a) interest in collaborating with outside researchers and (b) familiarity with other methods and perspectives most helpful for prioritization research

[1] For a general illustration of the fact that interventions/causes can vary in impact by orders of magnitude, see Ord (2013).

(see, e.g., Merryman, 1975; Merryman & Pérez-Perdomo, 2018). While it may not seem immediately obvious why unidisciplinary research is suboptimal from a prioritization standpoint, note that, as alluded to above, prioritization by its very nature involves drawing from techniques outside of law, such that a legal research program that does not sufficiently incorporate interdisciplinary methods is almost by definition incapable of engaging in prioritization as defined above.

Unidisciplinary legal scholarship was similarly ubiquitous in common-law jurisdictions prior to the 1970s and has been observed to have remained prevalent among traditional doctrinal scholars (cf. Posner, 1993, pp. 1653–1654).[2] Although researchers in some common-law jurisdictions are increasingly open to interdisciplinary work—as in the United States, where lawyers and legal academics receive more cross-disciplinary training than perhaps anywhere else in the world, and which has seen an explosion in interdisciplinary legal research movements in the last few decades (see, e.g., Eisenberg 2011; Posner, 1987)—the majority of lawyers and legal researchers remain most familiar with methods limited to the humanities rather than more quantitative fields, even in the most interdisciplinary-research-friendly jurisdictions. For example, according to data from the U.S. Census Bureau's Survey of Income and Program Participation (SIPP 1996 to 2013) and American Community Survey (ACS 2009 to 2014), roughly half of those with an American JD had majored in the humanities or social sciences for their undergraduate degree, whereas just 18% had majored in a science, technology, engineering, and mathematics (STEM)- or business-related discipline (Simkovic & McIntyre, 2014; Simkovic & McIntyre, 2018).[3] Meanwhile, although recent hiring reports reveal that tenure-track law professors in the United States are increasingly more likely to hold a non-JD doctorate (as many as 50% in recent years), not a single entry-level tenure-track professor in the last five years was reported to have held a doctorate in a STEM-related field, and, despite the apparent rise in law and economics, less than 5% of new professors held a PhD in economics (see Lawsky, 2020).[4] These figures are likely to far exceed those of other common-law jurisdictions, such as the United Kingdom, and civil-code jurisdictions, where law is generally studied from the undergraduate-level onwards, and where legal academics are very unlikely to

[2] For an overview of the differences and similarities between common-law and civil-law systems, see generally Dainow (1966), Merryman (1981), Merryman & Pérez-Perdomo (2018), Pejovic (2001), and Tetley (1999).

[3] Note that this 18% figure is likely to skew much lower if business-related majors are removed from the sample. For example, Harvard Law School reported that roughly 10% of its incoming students for the class of 2023 had majored in STEM-related disciplines for their undergraduate degree.

[4] Note that, while Lawsky (2020) is based on a compilation of official data and is considered to be generally reliable, the report itself is unofficial (and acknowledges the possibility that it is incomplete).

hold a non-law-related doctorate or even a non-law-related bachelor's degree (cf. Merryman & Pérez-Perdomo, 2018). Thus, it seems safe to suppose that most lawyers and legal academics in common- and civil-law jurisdictions alike are likely to be unfamiliar with many of the prioritization methods and may be reluctant to adopt the cutting edge of other research fields.[5]

Indeed, to the extent that it has engaged with and drawn from the methods of other disciplines, law has still tended to lag behind the other fields on which it draws. For example, *behavioral law and economics* was developed multiple decades after the field of *behavioral economics* (cf. Kahneman & Tversky, 1979; Simon, 1972 with Jolls et al., 1998; Posner, 1998) and continues to lag far behind, while the *Journal of Empirical Legal Studies*, among the first peer-reviewed, interdisciplinary legal journals of its kind, was not established until 2004. The failure to incorporate new methods and findings may extend beyond the realm of interdisciplinary research; the emphasis on the case method in common-law legal education, for example—and the similar focus in common-law legal scholarship—in some sense inherently orient legal thought towards issues that have appeared before courts and therefore tend to neglect issues that may arise only in the future or that extend beyond the boundaries of the common law's more established rules.

Aside from methodology, a second potential explanation relates to the substance and scope of the issues that legal academia has chosen to work on. First, while it stands to reason that international and global issues are, *ceteris paribus*, more important than national ones, legal academia has disproportionately given prominence to the latter more than the former.[6] Insofar as global issues are likely to be addressed, at least in part, by international law, it is revealing that Shapiro

[5] A similar pattern can also be observed among common-law judges. For example, in a survey of 400 American state court judges, Gatowski et al. (2001) found that only 5% of the respondent judges demonstrated a clear understanding of the concept of falsifiability in science, and only 4% demonstrated a clear understanding of error rate in statistics.

[6] It is worth pointing out, of course, that while international and jurisdiction-independent legal issues tend to be more important and neglected than national legal issues, the latter tend to be more tractable. Additionally, it seems plausible that at least some issues of international law and institutions are less important than national law and institutions of the largest and most influential legal systems (for example, the United States, China, and the European Union), such that in many cases it may be preferable to work on issues of national law as opposed to those of international law. However, even in its strongest form, this would not be an argument against prioritization as presented in this Section but rather serve as either (a) an objection to the necessity of formal prioritization methods or (b) a critique of a prioritization methodology that emphasizes international legal research questions at the expense of high-impact national issues. We further address these concerns in our Sections on objections to prioritization (Section 1.2), methodology (Section 3), and meta-research (Section 7).

and Pearse's (2012) compilation of the most cited law review articles of all time found that no article on international law made the top 100, with the most cited article on international law receiving less than half the citations as number 100 on the list (see also Shapiro, 1996; Shapiro, 2000a; Shapiro, 2000b). Second, with regard to national and international scholarship the vast majority of work is near-term oriented (cf. Shapiro & Pearse, 2012), whereas many of the most impactful legal issues appear likely to concern the long term (Section 2). Third, comparatively little scholarship appears to be done in a comparative context,[7] which is potentially suboptimal in terms of (a) evaluating proposed legal solutions based on their effectiveness (or lack thereof) in other jurisdictions, and (b) identifying existing legal interventions in other jurisdictions that might be effectively implemented in one's own jurisdiction, independent of the scale or substance of the issues that are chosen to focus on.

This revealed preference towards national, near-term, and unidisciplinary research questions is likely influenced by the incentive structure of legal academia. Legal scholars are rewarded above all for a strong publication record, which may incentivize a disproportionate focus on near-term issues, especially those most likely to be cited in high-profile court cases and journal articles during the course of one's career, even if long-term issues would ultimately be higher impact. One potentially notable instance of this in the United States (whose legal academy, as mentioned above, is among the prioritization-friendliest in the world) relates to the Chevron doctrine, a seemingly narrow topic of United States administrative law that has garnered widespread attention and resources from within American legal academia (*Chevron U.S.A. v. Natural Resources Defense Council*, 467 U.S. 837 (1984)). Performing a search for "Chevron" on the legal databases from HeinOnline yields over 55,000 results, more than three times as many results as a search for "human welfare" and over three hundred times as many results as a search for "human extinction" or "existential risk." Moreover, the top articles and cases relating to Chevron likewise receive significantly more citations than those mentioning human welfare, and several orders of magnitude more citations than those

[7] Although little systematic data is available, self-reported numbers on individual law school websites suggest a relatively low emphasis on both international and comparative scholarship, even at institutions that have historically been considered most open to these practices. For example, Yale Law School famously takes a more meta approach to legal education and may train the greatest number of future legal academics of perhaps any legal institution in the world, yet just 28 of the 112 faculty are listed as teaching or researching law in either a comparative or international context. Harvard lists just 56 of 392 law faculty members as part of its international legal studies program. At other law schools, this number is often several times lower; the University of Virginia Law School, for instance, lists 2 out of 90 of its resident faculty members in the "comparative law" faculty.

mentioning human extinction and existential risk, despite the latter three examples being of plausibly greater importance.[8]

Relatedly, many researchers may find that they are unable to find journals willing to publish prioritization-informed and -related research. As alluded to above, the relative dearth of interdisciplinary legal research may be explained by the lack of high-quality journals focused on publishing such research. For example, according to Washington & Lee's (2018) Law Journal Rankings, no interdisciplinary legal journal was among its top 30 (the *Journal of Legal Studies*, for example, was ranked 72nd). In Google Scholar Metrics' top 20 ranking of law journals, there are likewise no journals dedicated to interdisciplinary research. Similar patterns are found in rankings performed by HeinOnline, InCites Journal Citation Reports, and Scimago Journal and Country Rank.[9]

Even if they could ultimately publish such research, many legal scholars may simply choose not to work on it in the first place. This may be due to a lack of available information regarding which issues are the most important (and how best to work on them), or simply a lack or loss of motivation to pursue such issues. For example, although a survey of over 22,000 United States undergraduate students found that public-spirited motivations were the top reasons for considering a law degree (Association of American Law Schools & Gallup, 2018), a separate survey suggested that such motivations tend to significantly decrease as early as the first year of law school, particularly among those who perform the best academically. Further evidence suggests that this may depend on the competitive structure of the specific law school (Sheldon & Krieger, 2004).

[8] All searches were performed in October 2020. The original and most-cited Chevron case (*Chevron U.S.A. Inc. v. Natural Resources Defense Council, Inc.*, 467 U.S. 837 (1984)) was cited by a combined 27,649 articles and cases, compared to 10,542 for the most-cited "human welfare" case (*Planned Parenthood of Southeastern Pennsylvania v. Casey, Governor of Pennsylvania*, 505 U.S. 833 (1992)). The most-cited "Chevron" article (Kagan, 2001) was cited by 1,095 articles and cases, compared to 893 for the most-cited "human welfare" article (Sullivan, 1992). Meanwhile, the most-cited "human extinction" result of any type (Doremus, 2000) was cited by fewer than 100 articles and zero cases, as was the most-cited result for "existential risk" (Tribe & Gudridge, 2003).

[9] It is worth pointing out that this does not take into account the willingness of non-interdisciplinary journals to publish interdisciplinary work, which, as alluded to earlier in the Section, is often the case in common-law jurisdictions, such as the United States (though, as also emphasized earlier in the Section, not necessarily scholarship most likely to be useful from a prioritization standpoint). On the other hand, given the relative tendency of civil-law scholarship to eschew interdisciplinary research in general, from a global standpoint these rankings (which skew heavily towards American journals and legal scholarship) are, if anything, more likely to be a generous portrayal of the status of interdisciplinary legal journals in the legal academy.

Given these historical trends and the current state of legal academia, it is unsurprising that law might be slow to adopt prioritization methods. Legal priorities research specifically aims to address this issue by developing and promoting rigorous approaches to the question of how legal scholars can do the most good to maximize law's potential to solve some of the world's most pressing problems.

1.2 Objections to Legal Priorities Research

Even if one accepts the importance and neglectedness of legal priorities research in the abstract, one may still object to the practice on various grounds, including those relating to (a) the relative impact of different legal research questions, (b) the efficacy of existing prioritization methods in law, (c) the responsibility of legal academia to perform legal priorities research, and (d) the methodological limitations of such research. Although evidence suggests that most legal academics do not find these objections to be even somewhat compelling (Martinez & Winter, 2021),[10] here we briefly discuss each of them in turn, including reasons for why they do not appear to be particularly convincing. We also further discuss the fourth objection in Sections 2 and 3.

Under the first objection, legal priorities research is not worthwhile because potential research questions do not differ widely in their relative impact, such that, for example, the highest-impact legal research questions are not much more impactful than lower-impact ones. While it is important to take this sort of objection seriously, evidence from other contexts suggests that it is unfounded. For example, in the context of global health, we know that certain interventions—such as supporting the distribution of insecticide-treated bed nets to prevent malaria, deworming programs, and cataract surgery for the blind in developing countries—can be several orders of magnitude more effective than other interventions, even those that also seem very highly effective (see, e.g., Ord, 2019). To the extent that a similar phenomenon exists with regard to solutions developed through legal scholarship (see Section 2), this would likewise dictate in favor of prioritizing research questions that are more likely to result in highest-impact legal solutions (which, as alluded to previously, is a central aim of legal priorities research).

[10] In a survey of legal academics from around the world, legal academics were asked to rate, on a scale of 1 to 7 (with 1 representing "extremely uncompelling," 4 representing "neutral," and 7 representing "extremely compelling") how compelling they found four different objections to the idea of prioritizing among research questions based on ethical importance. The four objections included the idea that such prioritization was either (a) not very important, (b) not feasible/tractable, (c) not neglected, and (d) not the responsibility of legal academics. The mean rating was below 4 for each of these objections, and none of the four objections were rated as at least somewhat compelling (i.e., a 5 or higher) by more than 35% of academics.

According to the second objection, although prioritization itself may be an important task, the current method of prioritization in law simply works sufficiently well. Consequently, there is no need for a formal research program dedicated to the task. Though it remains an open question to what extent individual legal academics or academic legal journals attempt to prioritize research questions based on first-principles and evidence-based reasoning, neither the explicit and systematic prioritization of legal research projects nor evaluation of their effectiveness has received significant attention in legal academia. This suggests that legal academics who do attempt to prioritize among research questions do so largely on the basis of intuition alone. While it is true that, historically, consulting one's intuitions has been a commonly accepted practice, research in a variety of disciplines, most notably in behavioral economics and psychology, has shown intuition to be susceptible to a host of cognitive biases that often result in unreliable statistical, moral, and economic judgments (e.g., Alexander & Weinberg, 2014; Gilovich et al., 2002; Kahneman & Tversky, 1973; Kahneman & Tversky, 1982), as well as impaired strategic planning (e.g., Barnes, 1984; Das & Teng, 1999; Friesen & Weller, 2006). It seems reasonable to infer that this demonstrated unreliability of intuition in cognitive processes would interfere with informal prioritization of legal research, thus necessitating the existence of a formal program.

The third objection is that, although prioritization may be an important goal overall, it is not a goal for which legal research and academia are responsible. It is worth pointing out, of course, that insofar as this objection is seriously raised at all, it is much more likely to come from a lawyer trained in the civil tradition, which views (a) law as a more autonomous, isolated discipline; (b) the role of a lawyer as more of a technician or operator of a machine designed by others, their work being important but narrowly uncreative; and (c) legal research as pure and abstract, relatively unconcerned with the solution of concrete social problems or the operation of legal institutions (Merryman, 1975; Merryman & Pérez-Perdomo, 2018; Ostertag, 1993). In common-law jurisdictions, and particularly in the United States, where the perceived role of a lawyer or judge is more akin to that of a social engineer or omnicompetent problem solver, a legal researcher is much more accustomed to thinking about meta-questions such as "what is the purpose of law and legal research?" and accordingly may be much more likely to accept the task of prioritization as an appropriate one. Indeed, in a survey of (mostly) common-law-trained legal academics, fewer than 20% of participants rated this sort of objection to be even somewhat compelling (Martinez & Winter, 2021). However, even in jurisdictions where law is viewed as separate from the task of prioritization, to the extent that legal researchers are in favor of doing good and believe that prioritization is an effective means of doing so, they should likewise be in favor of prioritization, regardless of whether the prioritization itself counts as legal research.

The fourth objection holds that, although the idea of prioritization is important, empirically it is too intractable to reasonably determine or estimate which forms of legal research are more impactful than others. Although it might be difficult to verify the impact of every research question *ex ante*, it seems feasible to significantly increase the expected positive impact of legal research, given legal academia's disproportionate focus on seemingly narrow research questions. We deal with this objection more extensively as we further lay out our philosophical foundations (Section 2) and methodological framework (Section 3) for legal priorities research.

Setting aside these objections, the task of prioritization seems to be not only an important concept in the abstract, but also an important issue which should be appropriately addressed by legal research.

2 Longtermism

Legal priorities research, in the broadest sense, only requires that some actions are better than others according to some evaluative criteria, and that we can get rigorous evidence about which actions are among the best. Our own approach to legal priorities research is shaped by our shared commitment to the view that the long-term future is overwhelmingly important. The associated view in moral philosophy has been referred to as "longtermism."[11] Considering that longtermism is central to our research approach, it is worth detailing the arguments both in its favor and against it. In the following, we outline the foundations of longtermism (Section 2.1) and defend it against plausible objections (Section 2.2).

2.1 Foundations of Longtermism

Longtermism is the view that the primary determinant of the value of our actions and policies today is the effect of those on the very long-term future—hundreds, thousands, or even millions of years from now (Greaves et al., 2020, p. 7).[12] This view is based on two assumptions. The first is normative; the second is empirical.

The normative assumption is that the value of the effects of our actions and policies does not depend on when, where, or how those effects occur (Greaves & MacAskill, 2019, p. 5). A life in a distant country is not worth less than a life in our

[11] There has been some academic work on the philosophical foundations of longtermism (Beckstead, 2013a; Greaves & Pummer, 2019; Parfit, 1984), objections to longtermism (Greaves & MacAskill, 2019; Tarsney, 2020), and reducing existential risk (Bostrom, 2002, 2003a, 2013; Ord, 2020). There are also longtermist research agendas by the Global Priorities Institute (Greaves et al., 2020), Center on Long-Term Risk and the Forethought Foundation. Additionally, there has been extensive informal discussion on longtermism (see, e.g., Todd, 2017a; Whittlestone, 2017b; Wiblin, 2017a; MacAskill, 2019a).

[12] One may further distinguish among different versions of longtermism. At a first level, we can distinguish between weak and strong longtermism. "Weak longtermism" is the view that we should be particularly concerned with ensuring that the long-run future goes well, whereas "strong longtermism" holds that impacts on the long run are the *most important* feature of our actions and policies. The definition in the main text refers to strong longtermism. For more information, see Greaves and MacAskill (2019).

neighborhood.[13] Analogously, a life lived in 100 years is not worth less than a life lived now. This also implies that we must consider both the direct and indirect consequences of our actions and policies. For example, a direct consequence of distributing insecticide-treated bed nets in sub-Saharan Africa is a reduction of malaria incidents and child mortality (Pryce et al., 2019). Yet, that does not imply that those consequences are "better" *per se* or more important morally speaking than some of the indirect consequences of distributing insecticide-treated bed nets, such as improved education (Kuecken et al., 2014) and increased GDP growth (Gallup & Sachs, 2001; Sachs & Malaney, 2002).

However, while the attention to the consequences of actions and policies might imply that longtermism is an inherently consequentialist theory, this is just one approach to justifying longtermism. Alternatively, from a deontological perspective, it can be argued that we owe a duty to future generations, independent of what a consequentialist or even utilitarian calculus might demand.[14] Or, from a virtue ethics perspective,[15] that it is a virtue to act in such a way that protects future generations by exercising patience, self-discipline, benevolence, and taking responsibility for our actions (Gaba, 1999, pp. 283–287; cf. also Ord, 2020).[16] Further, empirical evidence suggests that caring for future generations, including those in the far future, is a view held by most legal scholars independent of their preferred moral theory (Martinez & Winter, 2021).[17] Many non-consequentialist

[13] See De Lazari-Radek and Singer (2014), Pogge (2002) and Unger (1996).

[14] Baier (1980), for instance, argues for the protection of future persons from a rights perspective.

[15] Ord (2020) refers to these as "civilizational virtues." See also Schell (2000) and Brand (2000).

[16] In addition to the perspectives outlined above, one might also value the long-term future from a purely aesthetic or intellectual achievement standpoint (Todd, 2017a). The robust case for caring about the long-term future, and about existential risk in particular, can be illustrated by the following passage from Ord (2020, p. 56): "[W]e could understand the importance of existential risk in terms of our present, our future, our past, our character or our cosmic significance. I am most confident in the considerations grounded in the value of our present and our future, but the availability of other lenses shows the robustness of the case for concern: it doesn't rely on any single school of moral thought, but springs naturally from a great many. While each avenue may suggest a different strength and nature of concern, together they provide a wide base of support for the idea that avoiding existential catastrophe is of grave moral importance."

[17] When asked "On a scale of 0 to 100, how much *should* your country's legal system protect the welfare of humans living in the far future (100+ years from now)" legal scholars, on average, answered about 68 (with 0 representing "not at all," and 100 representing "as much as possible"). This is roughly the level that researchers ascribe to

moral theories might maintain that, while consequences aren't the only thing that matters morally, they do, *ceteris paribus*, matter to some degree. This general acceptance about the (*ceteris paribus*) importance of consequences grounds our focus on consequences throughout our discussion of longtermism.

The empirical assumption is that, in expectation, the future is vast in size. One could approach this assumption by comparing the human species with other mammalian species (Greaves & MacAskill, 2019, p. 4). The lifespan of a typical mammalian species is about 1 million years.[18] Since the modern human species (*Homo sapiens*) is at least 200,000 years old, possibly 300,000 years old (see Galway-Witham & Stringer, 2018; Schlebusch et al., 2017),[19] we should expect, on average, to persist for another 700,000 to 800,000 years. However, the human species seems to have succeeded so far in protecting ourselves from most of the usual extinction threats[20] that mammals face, which, all else equal, would lead to humans persisting for much longer. Further, current estimates suggest that the Earth could remain habitable for around one billion years.[21] This would translate to 30 million future generations,[22] should humanity manage to survive this long. Moreover, it seems at least possible that humanity will one day leave the Earth and settle to the stars (Beckstead, 2014). In this case, the ultimate limits to human flourishing are set by the laws of physics and the expected end of the universe in quintillions of years' time (Adams & Laughlin, 1997; Adams & Laughlin, 1999). Such predictions suggest that future generations could vastly outnumber current generations in expectation. We may also expect many positive trends of moral, political, and technological progress to continue, such as curing diseases (Ortiz-Ospina & Roser, 2016), reducing extreme poverty (Roser & Ortiz-Ospina, 2013), increasing the number of democracies and access to equal rights (Pinker, 2018), and further scientific discoveries creating enormous value in the future if said trends continue.[23] Overall, the vast size and potential of the future could allow for unprecedented amounts of flourishing. However, major risks (see Section 3.2.1) threaten this potential by

the current legal protection of humans living in the present, which they estimate at 70 (Martinez & Winter, 2021).

[18] Estimates range from 0.6 million (Barnosky et al., 2011) to 1.7 million years (Foote & Raup, 1996).

[19] The earliest divergence between human populations may have occurred 350,000 to 260,000 years ago (Schlebusch et al., 2017).

[20] The usual threats of extinction faced by species include environmental, demographic and genetic factors (Benson et al., 2016).

[21] The end of complex life on Earth is expected to come in between 0.9 and 1.5 billion years (Caldeira & Kasting, 1992).

[22] If we assume three generations per 100 years.

[23] For greater discussion, see also Rosling (2018), Pinker (2012), and Pinker (2018).

curtailing positive trends and creating unprecedented disvalue or endangering our very existence. The future therefore seems to be the locus of most expected value and disvalue, warranting our attention.

Supposing one accepts the plausibility of these assumptions, one may consider longtermism from a legal standpoint. Longtermism is still highly neglected relative to its importance within legal research and the law, which is surprising given that, as alluded to above, legal scholars appear to be in favor of their countries providing much stronger legal protection to future generations (see Section 1.1; Martinez & Winter, 2021).[24] Legal scholars are also in favor of prioritizing legal research questions based on longtermist considerations (Martinez & Winter, 2021).[25] The reasons for this are likely multifaceted. A possible explanation is that our current political and economic systems are optimized for present generations on account of the fact that those are the people who vote, buy products, and advocate for their own interests (González-Ricoy & Gosseries, 2016; John & MacAskill, 2020). Even in a system that attempts to take into account the interests of future generations, operationalizing is arguably challenging, considering they cannot represent themselves. Some jurisdictions have recently tried different formats, such as parliamentary groups, commissioners, and funds, but with limited success (see Rose, 2018; John, forthcoming). Besides that, the measurement of harms in the future might be deemed too difficult, making it practically challenging to assign liability. Furthermore, protecting future generations might be perceived as too politically charged, which might contribute to the wariness of legal actors in approaching the subject, particularly courts**.** There may also be underlying psychological reasons. In particular, the neglectedness of longtermism could result from a number of heuristics and biases (Beckstead, 2013a, pp. 41–46; see also Yudkowsky, 2008b), such as people's insensitivity to quantitative differences when dealing with large numbers, a phenomenon referred to as "scope insensitivity" or "scope neglect" (Baron &

[24] When asked about how much their legal system currently protects the welfare of humans living in the far future (understood as 100+ years from now), a set of several hundred legal academics from around the world responded, on average, 22 on a scale of 0 to 100 (with 0 representing "not at all," and 100 representing "as much as possible"). When asked how much their legal system *should* protect the welfare of humans living in the far future, legal academics responded, on average, around 68, suggesting that legal academics, on average, believe that their respective legal systems should provide future generations with three times as much protection as they do currently (Martinez & Winter, 2021).

[25] With regard to legal academia specifically, around 75% of legal academics responded that legal research should prioritize research questions that, if worked on, would most positively influence the long-term future of humanity, whereas only 30% responded that legal academia currently prioritizes working on such questions (Martinez & Winter, 2021).

Greene, 1996; Greene & Baron, 2001; Stucki & Winter, 2019). Humans seem to have further difficulties thinking about the vastness of the future, in particular about human extinction scenarios (Schubert et al., 2019; Yudkowsky, 2008b). Similarly, humans are particularly bad at thinking impartially across time (O'Donoghue & Rabin, 1999) which, for instance, is likely to influence the (intuitive) evaluation of criminal environmental protection laws given its (partial) justification of protecting a large number of future lives (Winter, 2020b).

2.2 Objections to Longtermism

In this Section, we provide an overview of popular objections to longtermism and briefly discuss their plausibility. Although we do not aim at making novel contributions to this subject here, we present the objections to provide our readers (many of whom may be encountering longtermism for the first time) with a more thorough discussion of the subject. The objections we outline below refer to (a) the tractability of longtermism, (b) its decision-theoretic assumptions, and (c) its underlying population ethics. Throughout the analysis of objections, we put an emphasis on their specific application to the legal context.

With regards to the tractability of longtermism, one could argue that it is impossible to influence the far future.[26] This would be the case if the effects of our actions and even policies decay over time making the effects in the short-term outweigh any in the long-term. Greaves and MacAskill (2019, pp. 7–8) refer to this as the "washing-out hypothesis."[27] Although this hypothesis might be true for some trivial actions, there appear to be at least some non-trivial actions whose effects do not diminish over time or, in other words "wash out" (Beckstead, 2013a, pp. 3–8; Greaves & MacAskill, 2019, pp. 7–15). This includes, for instance, many aspects of law and legal institutions such as the persistence of the common law (Berman, 1985), Eastern legal institutions (Kuran, 2011; Rosett et al., 2003), and the similarly long-lasting effects of Roman law (Watson, 1991). Law may also be capable of influencing the far future through long-lasting legislation (e.g., the German criminal code of 1871), the role of precedent (Gerhardt, 1991), and path-dependent features of legal change (Hathaway, 2003).[28]

[26] For more information on the intractability objection, see Beckstead (2013a, pp. 3–8) and Greaves and MacAskill (2019, pp. 7–14).

[27] Greaves and MacAskill make the point that the washing out hypothesis may only apply to *ex ante* effects, as opposed to *ex post* effects.

[28] This list of long-term effects of law is non-exhaustive. The question is of central importance to many aspects of the research agenda and is explored in more depth in parts of Sections 6 and 7.

But even if one assumes that it is possible to influence the far future, one could still argue that it is impossible to *predict* our influence.[29] This objection has a grain of truth insofar as it is impossible to make such predictions with certainty. However, certainty is not necessary for sound ethical decision-making; the effects of most actions are uncertain to some degree. Rather, it is sufficient that, under the orthodox account of decision-making under uncertainty (see Briggs, 2019), the *expected value* of actions and policies (the sum of the value of each potential outcome multiplied by the probability of that outcome occurring) is high relative to alternative actions. Strikingly, in the above-mentioned survey on the long-term effects of law, over 70% of legal academics responded they agreed with the statement that there are "predictable, feasible mechanisms through which the law can influence the long-term future (understood as at least 100 years from now)" (Martinez & Winter, 2021). Despite this, the predictability of the effects of law on the long-term future remains one of our key uncertainties and is later addressed directly as a research question within the Section on meta-research (Section 7).

Additionally, the use of expected value theory itself may be challenged. In particular, one may object to expected value theory on the grounds that it runs into problems in circumstances with arbitrarily low probabilities of outcomes with arbitrarily high value.[30] One might worry that we are in a similar position with respect to the long-term future. To put it simply, our chances of affecting the far future are tiny, and the payoff immense. However, expected value theory seems to logically follow from hard-to-deny axioms.[31] It is also not clear what a preferable alternative model that deals with small probabilities of large payoffs might look like. On closer inspection, we may also find some of the implications of expected value theory in such scenarios rather intuitive. For instance, marginal improvements to the safety of nuclear reactors make only vanishingly small differences to the probability of meltdown, but may still be worthwhile given the large costs in the event of a catastrophe. In general, longtermist interventions appear to embrace only ordinary tolerance for this susceptibility to the dominance of low-probability events (see Tarsney, 2019).

[29] For more information, see Whittlestone, (2017b).

[30] For more information on this so-called "fanaticism" problem of expected value theory, see Beckstead (2013a, Chapter 7), Bostrom (2009, 2011a), Ross (2006), Tarsney (2019), and Wilkinson (2020). More precisely, the problem requires that there is an outcome with arbitrarily high value and arbitrarily low probability that outweighs, in expected value terms, a guaranteed outcome of high value. One is then faced with a "reckless" or "risky" situation in maximizing expected value (Beckstead & Thomas, 2020; Wilkinson, 2020).

[31] See Steele & Stefánsson (2020) for discussion on the axioms of completeness, transitivity, continuity, and independence.

One could further argue that it is impossible to calculate the expected value for long-term interventions—we are "clueless."[32] While expected value theory is useful in the ordinary challenges to predictability, where probability distributions and the magnitude of our actions are available, longtermists, as this objection states, often find themselves in situations where both such variables of an expected value calculus are undetermined given our limited foresight. Cluelessness, then, raises a challenge for the expected value approach to longtermism. Although this objection might seem appealing *prima facie*, it has been argued that it is in fact not an objection to longtermism (Greaves, 2016).[33] To illustrate, cases of cluelessness are arguably most salient in evaluating the long-term effects of interventions primarily focused on improving the short-term, for example, charities that address global health and extreme poverty in developing nations. Such interventions exhibit various indirect long-term effects that are not included in their cost-effectiveness analyses, whose net impact is greater in aggregate and places one in a clueless position to evaluate their impact on the far future (Greaves, 2020; Karnofsky, 2013).[34] For instance, on the one hand, reducing extreme poverty produces lower rates of disease, increase in economic growth, and improved quality of living for all those affected. But on the other hand, greater economic development tends to contribute more carbon emissions and produces risks associated with dangerous technologies (Hubacek et al., 2017; Karnofsky, 2013).[35] Likewise, laws focused on the short-term may face similar challenges. For example, environmental regulations[36] that banned leaded gasoline may have contributed to a dramatic reduction in crime (Nevin, 2007; Stretesky & Lynch, 2004; Marcus et al., 2010). This could plausibly increase economic growth (Goulas & Zervoyianni, 2015) and therefore influence the long-term future. While still facing concerns over cluelessness, one may favor longtermist interventions over short-term oriented ones as they arguably offer greater predictability of their effects due to their direct focus on the far future (Greaves, 2020).

Notwithstanding the above, the case for longtermism does not depend on expected value theory (Whittlestone, 2017b). For instance, it has been argued that the maximin principle, the decision theory that maximizes the minimum payoff, or

[32] For more information on the problem of "cluelessness", see Greaves (2016), Lenman (2000), Tomasik (2015a), and Wiblin and Harris (2018a); see also Feldman (2006), Lindblom (1959), McGee (1991), and Smith (2010).

[33] For recent informal discussion, see Greaves (2020).

[34] Karnofsky (2013) refers to these as "flow-through effects".

[35] For more discussion on the relationship between economic growth and risks from emerging technology see Bostrom (2019) and Aschenbrenner (2020).

[36] For instance, 38 Fed. Reg. 33734, implementing the Clean Air Act (42 U.S.C. § 7545(c)(1)).

in other words chooses the best worst-case scenario, may be well equipped to address exceptionally bad worst-case outcomes from a regulatory perspective, such as pandemics, climate change, emerging technology, and other risks that threaten the long-term future (Sunstein, 2020).[37] Additionally, Lempert (2019) has proposed decision support tools[38] to address the "deep uncertainty" in policy decisions through Robust Decision Making (RDM).[39] RDM endorses a norm of robust satisficing that maximizes satisfactory outcomes across many different, possible futures of the world. Mogensen & Thorstad (2020) have pointed out the connection between robust satisficing and addressing challenges of shaping the far future such as reducing existential risks. Finally, Mogensen (2020) argues that any novel account that gives plausible guidance in quotidian cases will support longtermism. Even if one rejects expected value theory, there remain a number of alternatives for decision-making under uncertainty that support the case for longtermism.

Apart from concerns with decision-theoretic assumptions, one could argue that future welfare matters less than current welfare.[40] One could weigh future welfare less by applying a positive discount rate (for example, 5% per annum), as is often done in economic cost–benefit analyses of policies affecting the future.[41] This way,

[37] More precisely, Sunstein argues that maximin may be attractive from a regulatory policy perspective under the following four conditions: "(1) The worst-cases are very bad, and not improbable, so that it may make sense to eliminate them under conventional cost-benefit analysis. (2) The worst-case outcomes are highly improbable, but they are so bad that even in terms of expected value, it may make sense to eliminate them under conventional cost-benefit analysis. (3) In circumstances of Knightian uncertainty, where observers (including regulators) cannot assign probabilities to imaginable outcomes, the maximin rule may make sense. (4) The probability distributions may include "fat tails," in which very bad outcomes are more probable than is usual…" (Sunstein, 2020, p. 3).

[38] Robust Decision Making is often used for framing and exploring decision problems rather than proving normative criteria for solving them and thus does not, strictly speaking, constitute a decision theory (see Mogensen & Thorstad, 2020; Helgeson, 2020). Robust satsficing, however, offers one example of normative criteria to use when solving RMD problems.

[39] Thorstad and Mogensen (2020) discuss other decision support tools for decision making under deep uncertainty (DMDU) such as Dynamic Adaptive Policy Pathways (Haasnoot et al., 2013) and Info-Gap Decision Theory (Ben-Haim, 2006).

[40] For more information on discounting future welfare, see Beckstead (2013a, pp. 63–64), Greaves et al. (2020, pp. 32–35), and Greaves and MacAskill (2019, p. 5).

[41] If one took the commonly used discount rate of 5% per year and applied it to our future, there would be strikingly little value left. Applied naïvely, this discount rate would suggest that our entire future is worth only about twenty times as much as the coming year, and that the period from 2100 to eternity is worth less than the coming year. As Ord (2020) illustrates, this would also mean that if we can save one person from a

future welfare would decrease in importance each year. This would significantly limit the value of future welfare and thereby undermine the core argument for longtermism. While discounting may make sense in economic contexts (given inflation and the time value of money due to interest or other investment potentials), discounting future *welfare as such* is incompatible with impartiality and the intuitive notion that moral value is relevant independent of when and where it occurs. This impartial view of discounting or the "zero rate of pure time preference" is endorsed by many philosophers and economists.[42]

Another objection refers to the underlying population ethics,[43] i.e., assessing the moral value of actions that influence both who is born and how many people are born (Greaves, 2017b). Given that longtermist interventions often directly deal with maintaining humanity's survival (especially in the case of mitigating extinction risk), population ethics becomes highly relevant. In this regard, one may argue that there is no value in bringing people into existence. Consequently, longtermist interventions would only be valuable insofar as they ensure that the future is good for whatever beings happen to exist but not for ensuring the very existence of future beings. This "person-affecting" view (cf. Arrhenius et al., 2017; Greaves, 2017b) would undermine the importance of preventing human extinction insofar as the argument for preventing human extinction depends on the number of potential beings that could inhabit the future.[44] However, since preventing human extinction is not the only action that influences the long-term future (see Section 3.2.1), people leaning towards person-affecting views can still endorse longtermism

headache in a million years' time, or a billion people from torture in two million years, we should save the one from a headache.

[42] See Greaves (2017a) for a survey of discounting in public policy, including a survey of the arguments for and against a positive rate of pure time preference. One could further argue, for instance, that the positive rate of pure time preference is incompatible with the Pareto principle. She also points out that a zero rate of pure time preference is endorsed by, among others, Broome (2008), Buchholz & Schumacher (2010), Cline (1992), Cowen & Parfit (1992), Dasgupta (2008), Dietz et al. (2008), Gollier (2013), Harrod (1948), Pigou (1932), Ramsey (1928), Sidgwick (1907), Solow (1974), and Stern (2007). Other related philosophical work includes Cowen and Parfit (1992), Mogensen (2019), and Parfit (1984). In a survey of experts on social discounting, 38% accepted a zero rate of pure time preference (Drupp et al., 2018).

[43] This Section only outlines a small fraction of the views and ongoing debate in population ethics. For a full discussion, see Greaves (2017), Parfit (1984), and Thomas (2017). Other views on population ethics include Totalist and Averagist theories, 'Variable Value' theories, and 'Critical Level' theories.

[44] It has been argued by MacAskill (2020a) that preventing extinction at least in the near term can be based on the importance of preserving option value and compatible with person-affecting views. See Lewis (2018a) for a defense of extinction risk mitigation from a person-affecting view.

(MacAskill, 2020a).[45] Person-affecting views could still see the very long-run effects on welfare as overwhelmingly important, so long as we have sufficient confidence in future beings existing in the first place. Indeed, various organizations explicitly endorsing person-affecting or otherwise “downside-focused” value systems, that is to say, value systems that place relatively less importance on ensuring that net positive beings come into existence, embrace a longtermist approach.[46]

Overall, longtermism seems fairly robust against the “traditional” objections outlined above, especially as a result of its compatibility with multiple value systems and decision theories. Furthermore, longtermism appears defensible in light of potential law-specific objections. In particular, legal academics seem to largely agree with the responsibility of law and legal research to positively shape the far future (Martinez & Winter, 2021). Although a majority of legal scholars (over 70%) agree that there are indeed feasible mechanisms through which the law can influence the long-term future (at least 100 years from now) (Martinez & Winter, 2021), and we ultimately expect that the law may have predictable effects on positively changing the trajectory of humanity, such as by reducing various risk (see Section 3), we find this empirical objection the most plausible. This calls for further evaluation of the predictability of the long-term effects of the law mentioned before: persistent legal institutions, legislation, precedent, and path dependence, to name a few.[47]

[45] There are a number of further responses to this objection. For example, there are serious objections to each of the several versions of the person-affecting view (see Greaves, 2017b, pp. 9–10). Furthermore, *average* population ethics causes unsolvable problems, namely the sadistic conclusion (see Greaves, 2017b, p. 3) and the violation of the mere addition principle (see Greaves, 2017b, p. 3; Parfit, 1984, pp. 417–421).

[46] See, for example, the Center on Long Term Risk, the Center for Reducing Suffering, and the Center for Emerging Risk Research.

[47] Cf. also Section 7 (Meta-Research).

3 Methodology

In this Section, we describe our methodological approach to legal priorities research. This will provide guidance for new researchers entering the field. Additionally, the high transparency of our methodology will make it easier for other researchers to critique our current approach, which will help us to improve it. As noted in Section 1.2, we are aware of certain methodological limitations. In the following, we describe our methodology for cause prioritization (Section 3.1) and for identifying research projects within cause areas (Section 3.2).

3.1 Methodology for Cause Prioritization

Simply put, cause prioritization is about finding problems where additional legal research can do the most good. A "cause" is a broad field around a particular problem or opportunity, such as fighting climate change, or improving the governance of artificial intelligence (Open Philanthropy, 2020a). "Cause prioritization" can be defined as the task of identifying causes with the highest expected marginal benefit of additional resources. The "marginal benefit" of a cause refers to the amount of "good done" per unit of additional resources, such as labor and funding, invested in that cause.

At the current stage, we primarily rely on existing literature. Only when there is sufficient reason to think that priorities in law deviate from global priorities will we engage in prioritization research. For instance, this may be the case with regards to climate change, which we hypothesize is significantly more neglected in legal research than in other fields (see Section 7). There are a few organizations, including the Global Priorities Institute at Oxford University, Open Philanthropy, and 80,000 Hours, which focus on prioritization research and whose values we share. Hence, it is necessary to follow the related research outcomes closely and update our priorities on an ongoing basis according to the best evidence available.

The relevant research conducted by the aforementioned organizations relies on the so-called "ITN framework."[48] According to this framework, one ought to prioritize cause areas that are (a) important, (b) tractable, and (c) neglected. *Importance*

[48] For a more detailed explanation of the ITN framework, see MacAskill (2018) and Wiblin (2019). For an analysis of its limitations, see Dickens (2016) and Wiblin (2016).

refers to the number of sentient beings affected and the degree to which they are affected by a given problem. To put it simply: "If we solved this problem, by how much would the world become a better place?" *Tractability* refers to the possibility of actually solving the problem. As a heuristic, one might ask: "If we doubled direct effort on this problem, what fraction of the remaining problem would we expect to solve?" Finally, *neglectedness* refers to the question: "How many resources will be dedicated to solving the problem before it is too late?"[49]

In Part 2 of this agenda, we explore a number of cause areas in more detail. This includes the law and governance of artificial intelligence (Section 4), synthetic biology and biorisk (Section 5), and institutional design (Section 6). Since choosing the right research project is one of the most important factors that determines the impact of legal research, we are also engaging in a number of meta-research projects (Section 7). Instead of competing with the existing organizations, our research in this area is significantly more specific in that it exclusively tackles problems that legal researchers encounter when prioritizing, such as whether to focus on international, comparative, or national law.[50] As part of Section 7, we are continuing to question and update our criteria to identify research projects within cause areas to which we will turn now.

3.2 Methodology for Identifying Research Projects

Within cause areas, we identify concrete research projects by applying two sets of criteria. Here we discuss each of these criteria in turn, including an obligatory primary criterion (Section 3.2.1), and a more holistic set of secondary criteria (Section 3.2.2) which ought to be interpreted in light of the primary criterion.

3.2.1 Primary Criterion

The primary criterion is to focus on research questions that positively shape *humanity's long-term trajectory* (Baum et al., 2019; Beckstead, 2013a; Beckstead, 2019).[51] If one takes humanity's current trajectory as a reference basis (*status quo*

[49] Note that the response to the related question of "how many resources are currently being dedicated to solving the problem?" is only an indicator for how many resources are going to be dedicated to the problem before it is too late. Consequently, the analysis involves a greater degree of uncertainty with regards to problems whose negative consequences would unfold decades from now, in comparison to those whose negative consequences would unfold very soon. For a better understanding of "long-term neglectedness," see also Section 3.2.2.

[50] On the importance of further prioritization research, see Todd, 2020b.

[51] The notions "humanity's trajectory," "world's development trajectory," "human trajectory," and "civilization's trajectory" are used interchangeably in the literature and this

trajectory),[52] it is conceivable that the conditions for welfare might drastically improve or deteriorate over time depending on humanity's actions today.[53] From this perspective, legal research should focus on research questions that increase the probability of entering a positive trajectory or decrease the probability of entering a negative trajectory. Important aspects within this framework are (a) the timing of different trajectory changes and (b) different types of risks that threaten a persistent or even permanent negative trajectory change. We will deal with these in turn.

A crucial consideration of any trajectory change is its timing. In general, greater amounts of future welfare require more or earlier entries into positive trajectories, and fewer or later entries into negative trajectories. However, the extent to which the timing of trajectory changes affects the far future also depends on the shape of the progress curve (Greaves & MacAskill, 2019, p. 9). For example, if we expect welfare gains to plateau at some point (s-curve), then the impact of a change in timing on the far future would be relatively low and bounded (Figure 1). Conversely, if we expect welfare gains to rise steadily (Figure 2) or exponentially (Figure 3), then the impact would be comparatively higher. Thus, legal research should focus on research questions that speed up positive trajectory changes or delay negative trajectory changes when future progress is thought of as linear or even exponential. The graphics below illustrate the importance of timing when it comes to trajectory changes.

agenda. One should bear in mind that "human" or "humanity's" trajectory does not entail that other sentient beings do not matter, but rather emphasizes the impact human action might have on the world's trajectory.

As Baum et al. (2019) note, even "trajectories of human civilization" may not only include civilizations led by genetic descendants of *Homo sapiens sapiens*, but also civilizations led by biological or non-biological beings that are engineered by *Homo sapiens sapiens* or its genetic descendants. To be precise, this agenda will instead refer to "human-originating civilizations," but one should note that the definition of this terminology does not differ from Baum et al.'s (2019) understanding of "human civilization."

Technically speaking, if one assumes that the future does not evolve sufficiently deterministically, it follows that there is no single trajectory. Instead, there would be a probability distribution over many possible trajectories (Beckstead, 2013a, p. 6; Beckstead, 2019, p. 91). Thus, the goal of "changing humanity's trajectory" would be better described as "changing the probability distribution over many possible trajectories." For the sake of simplicity, we will continue to refer to this more complex phenomenon simply as "changing humanity's trajectory."

[52] For more information on the status quo trajectory, see Baum et al. (2019, pp. 6–9).

[53] But see also Section 2.2 regarding empirical uncertainties of influencing the far future.

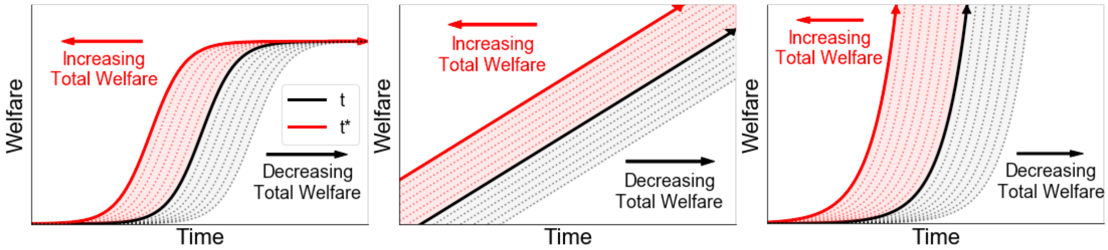

Figure 1. Diminishing Progress

Figure 2. Linear Progress

Figure 3. Exponential Progress

Figures 1–3. The total welfare, as measured by the area under the curve, for each trajectory, *t*, increases with an earlier trajectory change, *t**. The graphs also show the difference in welfare totals *across* the trajectories. The increase in total welfare as a result of earlier trajectory changes for *exponential* and *linear* growth curves is greater than earlier trajectory changes for *diminishing* progress curves, as shown by the greater area under the curve.

Given the high degree of both normative and empirical uncertainty as to what a positive trajectory change might look like, we are particularly concerned about preserving options.[54] Hence, legal research should aim at avoiding entering *persistent* or even *permanent trajectories*.[55] One example in this regard we are particularly concerned about is human extinction, which can be classified as a permanent trajectory because recovery from extinction seems unlikely. However, other scenarios, such as the rise of a new global (digital) authoritarian power, ought not to be neglected.

The following list contains different types of risks that legal research might aim at reducing in order to shape the human trajectory into a more positive direction. One should note that this list does not indicate that all options mentioned below to influence the long-term future are of equal value. It may well be the case that, as of now, existential risks ought to be prioritized even among the different risks listed due to the clear lock-in nature of their occurrence. Although the list is non-exhaustive, it can serve as a starting point.[56]

Reducing Existential Risks

"Existential risks (x-risks)"[57] are risks where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its

[54] See also MacAskill (2020a).

[55] Relatedly, Greaves and MacAskill (2019, pp. 9–10) refer to states of the world that have the property that "once they are entered they tend to remain in that state for a very long time" as "attractor states."

[56] See also generally Baum et al. (2019), Beckstead (2013a), Greaves & MacAskill (2019).

[57] For more information on existential risks, see Baum et al. (2019), Beckstead (2013a, pp. 5–6), Bostrom (2002; 2013), Cotton-Barratt and Ord (2015), Cotton-Barratt et al.

potential (Bostrom, 2002).[58] By their very nature, x-risks affect the course of a human-originating civilization for all time to come, either by premature human extinction or by locking in the conditions for welfare on an extremely low level. The exact threshold regarding what kind of event or consequences would satisfy this requirement has not been clearly identified within the existing literature. However, a minimum condition would be that the greater part of the potential of a human-originating civilization is lost (cf. Ord, 2020). Consequently, even a small reduction of such risks has an enormous expected value (Bostrom, 2013). Arguably the first anthropogenic existential risk emerged in the mid-twentieth century when the USA and the USSR started to build up their nuclear arsenals (Bostrom, 2002, p. 3; Ord, 2020). Particularly concerning existential risks may arise from synthetic biology (Lewis, 2020) and advanced artificial intelligence (Bostrom, 2014; Wiblin, 2017b; cf. Sections 4.1 & 4.2). To a significantly lesser degree, this may also be the case for runaway climate change (Duda & Koehler, 2016; Ord, 2020; Todd, 2017b).[59]

Reducing Risks of Astronomical Suffering

"Suffering risks (s-risks)"[60] are risks where an adverse outcome would bring about suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far (Althaus & Gloor, 2016).[61] Given our focus on longtermism (see

(2020), Greaves and MacAskill (2019, pp. 10–11), Liu et al. (2018), Ord (2020), Matheny (2007), Parfit (1984), and Todd (2017b). Academic institutions focusing on existential risks include the Future of Humanity Institute, the Centre for the Study of Existential Risk, and the Global Catastrophic Risk Institute.

[58] While some fields, such as AI safety research, make frequent use of the terminology "existential risk," other fields, such as synthetic biology, frequently discuss "global catastrophic risk (GCR)" or "global catastrophic biological risk (GCBR)," which includes some or all existential risks, depending on the definition, and can be considered at least a risk factor. For the related definitions, see footnote 103. It is for this reason that Section 5 of this agenda ("Synthetic Biology and Biorisk") will refer to both GCBRs and x-risks. Note, however, that we are particularly concerned with those GCBRs which can be classified as x-, s-, or p-risks.

[59] For an attempt of more specific quantifications, see Ord (2020) who estimates that the probability of existential risks to humanity this century is one-sixth. Pamlin and Armstrong (2015) give probabilities between 0.00003% and 5% for different scenarios that could eventually cause irreversible civilizational collapse.

[60] For more information on suffering risks, see Althaus and Gloor (2016), Daniel (2017), Gloor (2016a), and Tomasik (2011). Institutes that primarily focus on suffering risks are the Center on Long-Term Risk, Center for Reducing Suffering, and the Organization for the Prevention of Intense Suffering.

[61] Note that there are some overlaps between non-extinction existential risks and suffering risks. However, there are still sufficient differences to justify the distinction. For example, a future that contains both vast amounts of happiness and vast

Section 2), we are particularly concerned about long-term suffering risks, in the sense that the adverse outcome would persist for a very long time. It is important to note that if such a catastrophe is permanently locked in, this would be an outcome even worse than extinction. From this perspective, s-risks might be the worst kind of existential risks (Daniel, 2017). Risks of astronomical suffering may result from whole brain emulation technologies (Eckersley & Sandberg, 2013), the development of synthetic sentience (Bostrom, 2014),[62] and conflict involving powerful forms of artificial intelligence (Clifton, 2020).[63]

Reducing Risks of Losing Astronomical Pleasure

Let us define "pleasure risks (p-risks)"[64] as risks where an adverse outcome would prevent pleasure on an astronomical scale, vastly exceeding all pleasure that has existed on Earth so far. Just as greater development and progress in science and technology might lead to astronomical suffering or curtail humanity's potential, they may also enable experiences of astronomical pleasure and allow humanity to reach its full potential. P-risks include, although not exclusively, a range of human trajectories that avoid x- and s-risks, yet fail to reach astronomical pleasure and allow humanity to reach its full potential. Simply put, there is still a vast difference between being reasonably well-off and a trajectory of immense pleasure.[65] It should be noted that p-risks would be captured by x-risks, if one believes that the best case scenario of a human-originating civilization will be reached on any trajectory as long as humanity does not go extinct or suffer some other destruction of potential. However, merely avoiding such risks leaves open a vast range of possible trajectories for humanity, where an optimal outcome does not seem guaranteed.[66] Whereas

amounts of suffering would constitute an s-risk but not necessarily an x-risk (Althaus & Gloor, 2016).

[62] Bostrom (2014) refers to the latter as "mind crime."

[63] For an overview of potential s-risks, see Tomasik, 2019b; see also Section 4.2 for research projects on reducing s-risks from AI.

[64] An argument for why "disappointing futures" (p-risks) may be as important as existential risks, can be found in Dickens (2020). For a somewhat related argument, see also Bostrom, 2003a. A further related concept is that of "existential hope" (Cotton-Barratt & Ord, 2015).

[65] Unless one does not value pleasure itself. Cf., for instance, strong forms of negative utilitarianism, or variants of "tranquilism" (Gloor, 2017).

[66] But see Ord, 2020: "Given a long enough time with our potential intact, I believe we have a very high chance of fulfilling it: that setbacks won't be permanent unless they destroy our ability to recover. If so, then most of the probability that humanity fails to achieve a great future comes precisely from the destruction of its potential—from existential risk."

one may perceive x- or s-risks as threats or potential losses, very few would consider missing out on an opportunity of immense pleasure ("p-opportunities") as a threat or potential loss, but it may still be of utmost importance.[67] Having said this, prioritizing x- and s-risks over p-risks may still be justifiable at this stage of human history. This is especially the case as long as there are no identified immediate threats that would exclusively threaten a future of astronomical pleasure and would otherwise not be considered by x- or s-risks.

The graphic below illustrates the direct risks and opportunities introduced thus far.[68]

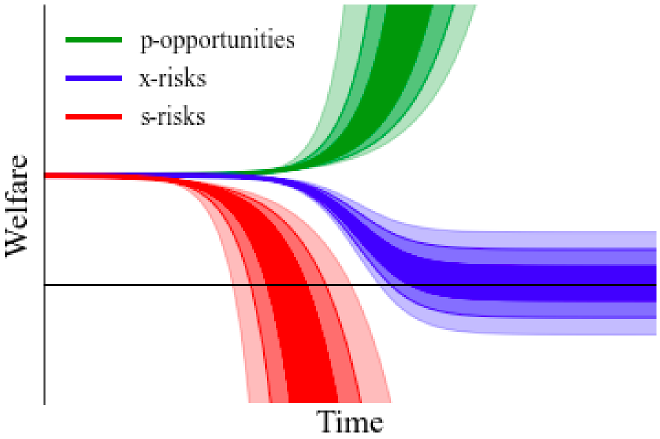

Figure 4. Possible trajectories and distributions for p-opportunities, and x- and s- risks.

Reducing Risk Factors

Risk factors are factors that contribute indirectly to x-risks, s-risks, and p-risks through increasing their probability of occurrence or combining with other risks to increase the severity of their consequences (cf. Baumann, 2019; Koehler, 2020; Ord, 2020).[69] A straightforward way to identify risk factors is to consider stressors for

[67] This may be explained by the tendency of human psychology to value losses more than gains, and to consider such trajectories of immense pleasure rather as a potential gain than loss of not entering it. Nevertheless, it would be a mistake to directly draw any normative conclusions from this. For an overview of the concept of loss aversion, see Kahneman et al. (1991).

[68] For further visualizations of the relationship between x-risks, s-risks, and the related concept of "global catastrophic risks," see also Aird (2020).

[69] Risk factors have also been referred to as "structural risks." The opposite of a risk factor has been coined a "security factor," i.e., factors that indirectly reduce x-, s-, or p-risks. Examples include strong institutions for avoiding existential risk, improvements in civilizational virtues, or (potentially) becoming a multi-planetary species (Ord, 2020). However, classifying something as a risk or security factor in this way seems to heavily depend on the framing. For instance, while the existence of strong institutions

humanity, such as those which threaten global peace, international decision-making and coordination efforts (Ord, 2020). Accordingly, many problems that cannot be classified as x-, s-, or p-risks themselves but that could still lead to a global catastrophe may pose substantially more indirect risk than direct risk and ought not to be underestimated. For instance, while runaway climate change is unlikely to cause extinction itself, it could cause conflict between major powers and threaten any global coordination efforts to tackle x-, s-, or p-risks.[70] One may be inclined to think that tackling direct risks should always be prioritized over working on the reduction of risk factors. Note, however, that what ultimately matters is how much we can reduce risk via a particular intervention all things considered. If, for instance, the greatest reduction in risk from the use of bioweapons could be achieved by decreasing the odds of conflict due to climate change or via improving global cooperation more generally, then research should focus on that.

3.2.2 Secondary Criteria

If multiple research questions meet the primary criterion, we apply a number of secondary criteria to analyze their importance. These secondary criteria are optional and do not stand on their own, but ought to be interpreted in light of the more abstract goal of positively shaping the long-term future. Because we necessarily have to consult our intuitions when estimating the importance of solving a problem for the long-term future, and such intuitions as well as the sheer number of considerations can be misleading, the secondary criteria should guide one's evaluation process by disentangling different concerns one might have. To put it simply, they serve as a checklist. The following list is non-exhaustive:[71]

for avoiding existential risk could be classified as a "security factor," the absence of such institutions would fall in the category of "risk factors." Primarily for the sake of simplicity, we will only refer to risk factors. Arguably, this may even be preferable from many perspectives, if one keeps in mind the human tendency to value risks (or risk factors) and opportunities (security factors) differently depending on their framing. The relationship between "risk factors" and "structural risks" is outlined in Section 4.1.

[70] On the relation between climate change and conflict, see generally Hiasang et al. (2013).

[71] The criteria themselves will have to be reviewed based on their effectiveness (cf. Section 7: Meta-Research). Also note that some factors will necessarily overlap to some degree. Given that the common denominator is the impact on the long-term future, this should not come as a surprise.

Taking Uncertainty into Account

When choosing a research question, we rely on empirical and normative assumptions, both of which can be wrong. Our choices need to reflect this risk by preferring research questions which are, all things equal, also important from the point of view of diverging moral theories and take empirical predictions of the far future into account which are, from our perspective, less compelling. This, for instance, strengthens the case for work on the risks associated with digital authoritarianism.

Bridging the Near- and Long-Term Future

All else equal, focusing on questions that are also relevant in the near-term is preferable.[72] The view that some projects are important from both a short- as well as a long-term perspective has recently been articulated with regards to risks arising from artificial intelligence (see Baum, 2018a, 2020; Cave & Ó hÉigeartaigh, 2019; Prunkl & Whittlestone, 2020). As we shall see in Section 9, this also (partially) applies to the study of non-human animal law.

(Indirect) Practical Significance

In many cases, it may be easier to come up with a solution to a research question than to implement that solution. Because the ultimate goal of legal priorities research is to maximize real-world impact, questions where practical implementation seems feasible are preferable when deciding between questions of otherwise equal importance. That said, "practical significance" ought to be interpreted broadly.[73] That is to say, some questions that, at first glance, seem only theoretically interesting, may help to guide humanity's long-term vision. Consequently, they may be of great *indirect practical significance*. One example of this would be to solve open questions with regards to future global governance systems. Although the chances of implementing advanced global governance mechanisms are low in the near-term, research findings may still be valuable from a practical perspective. First, one would want to be as prepared as possible in case opportunities for implementation

[72] "Bringing the near- and long-term future" is a specific application of the first point ("taking uncertainty into account"). We refer to this application explicitly due to its importance throughout the agenda.

[73] Connecting this more explicitly to the ITN framework (see above, footnote 48 and accompanying text), practical significance of a research question could be thought of as analogous to tractability, which we can further break down into two dimensions: (a) tractability of coming up with a solution to a research question that, if implemented, would address a problem, and (b) tractability of implementing a solution to a research question, once that solution has been proposed/developed. Our secondary criterion of "(indirect) practical significance" focuses on the implementation part (b).

arise, especially if there might be only a short window of opportunity, for instance, as a result of a major global catastrophe. Second, solving problems of global governance could guide humanity's long-term vision which, in turn, might have practical implications for the here and now and the path forward.

Unlocking Research Opportunities

In some cases, a research problem that addresses one question may facilitate or "unlock" opportunities to research and address other problems, thereby increasing the expected impact of the initial research project in comparison to other projects that only address their initial question.[74] Unlocking research opportunities can be extremely beneficial in new areas of research, such as legal priorities research, where the core concepts, questions, and methodologies are relatively underdefined (cf. Flynn, 2017). In this vein, "unlocking" should be interpreted broadly and may include, but not be limited to, structuring the research field, clarifying concepts, or identifying key questions. The identification of key questions and crucial considerations may be particularly important, as it can have multiplicative effects by opening up large areas of work for many more researchers and implementers to pursue (Flynn, 2017). One example of unlocking research opportunities is this agenda.

High Cross-Jurisdictional Value

All things equal, focusing on questions that are not specific to a particular jurisdiction, but would contribute to the solution of cross-jurisdictional problems ought to be given priority. This said, it may be the case that research relevant only to specific jurisdictions sometimes trumps these considerations, given the overarching importance of certain jurisdictions to fight existential threats, such as those of the United States, China, or the European Union. Notably, high cross-jurisdictional value does not indicate that research on international law is more important than research on national law or comparative legal approaches. On the contrary, in practice, it may sometimes be the case that existential threats can be more effectively addressed via national rather than international law, even though the latter would, in theory, be the most desirable solution.[75] An example for research questions of high cross-jurisdictional value is research on legal concepts and principles which are relevant to many jurisdictions, such as novel approaches to "balancing," "proportionality," or "public interest."[76]

[74] Flynn (2017) refers to this as "disentanglement research."

[75] See Section 7 with regards to our uncertainties on this matter.

[76] Cf. Section 6: Institutional Design.

Focusing on Long-Term Neglectedness

Some of the cause areas analyzed in the second part of this agenda, such as the law and governance of artificial intelligence and synthetic biology, have received very little attention despite their overwhelming importance for the long-term. However, what matters most from a longtermist perspective is not how many resources are currently being spent on these issues (short-term neglectedness), but rather how many resources will be spent on it in total before it is too late (long-term neglectedness) (Ord, 2020). For instance, AI has already been receiving more attention recently, and it is a rather straightforward task to imagine that the current COVID-19 pandemic will cause more resources to be spent on preventing risks from synthetic biology. From this perspective, one might hope that funding for the chronically and profoundly underfunded Biological Weapons Convention (BWC) of 1972 and other organizations in this area will increase.[77] However, relying on future spending based on the effectiveness of warning shots is a risky endeavor, especially because media and political attention can quickly shift towards less crucial areas. Furthermore, one needs to carefully distinguish between a cause area broadly defined, for example, "Law & Governance of AI" and more specific research projects. This leads us to the final criterion.

Focusing on Specific Neglectedness

More resources being spent on AI and synthetic biology, all things considered, does not mean that these resources will be spent on the most important risks and research projects within these areas. Just as humanity's efforts are often not focused on the most important cause areas, it is not clear that, within these areas, resources will be spent wisely once they receive the appropriate attention. For instance, one can imagine that, due to the current pandemic, more resources will be spent on preventing natural pandemics while engineered pandemics remain neglected, even though the latter are estimated to pose a far greater threat to humanity than the former.[78] Similarly, most research on climate change concentrates on the consequences of a rise of 1.5–2 °C, and very little resources have been spent on tackling the more extreme scenarios, such as a rise of more than 4.5 °C by the end of this century.[79] Consequently, while looking at the resources spent on a broadly

[77] According to Ord (2020), the BWC has a smaller budget than an average McDonald's.

[78] Ord (2020) estimates that engineered pandemics are roughly 330 times more likely to cause existential catastrophe than natural pandemics. He considers the x-risk posed by natural pandemics as 1 in 10,000 and 1 in 30 for engineered pandemics (Ord, 2020, p. 167).

[79] For instance, Sheerwood et al. (2020) suggest that there is a 6–18% chance of increases in temperature of at least 4.5 °C (8.1 degrees Fahrenheit) per doubling of atmospheric

defined field can serve as an indicator, it is important to estimate *long-term* neglectedness of *specific* risks when choosing research questions. This also leads one to prioritize risks that strike soon (cf. Ord, 2020).

CO_2 which could well happen before the end of the century. The Intergovernmental Panel on Climate Change states that there is at least a two-thirds chance that temperature increases will be somewhere between 1.5 °C and 4.5 °C while acknowledging that, if we end up between one and two doublings from pre-industrial levels, the range of eventual warming is 1.5 °C to 9 °C. Rogelj et. al. (2016) estimate that, if all countries fulfill their Intended Nationally Determined Contributions of the Paris Agreement but they do not grow more aggressive over time, there is still a 10% chance of exceeding a rise of 4.7 ºC.

Part 2

Exploration by Cause Areas

In this part, we explore a number of cause areas in more detail. This includes the law and governance of artificial intelligence (Section 4), synthetic biology and biorisk (Section 5), and institutional design (Section 6). Since choosing the right research project is one of the most important factors that determines the impact of legal research, we are also engaging in a number of meta-research projects (Section 7). Instead of competing with the existing organizations, our research in this area is significantly more specific in that it exclusively tackles problems that legal researchers encounter when prioritizing, such as whether to focus on international, comparative, or national law.

4 Artificial Intelligence

Artificial intelligence (AI)[80] could significantly shape the long-term future. On the one hand, it could enable scientific breakthroughs[81] and the accumulation of unprecedented wealth. On the other hand, it could pose existential risks and cause

[80] There is no generally accepted definition of the term "artificial intelligence." Since its first usage by McCarthy et al. (1955), a vast spectrum of definitions has emerged. Popular definitions have been proposed, among others, by Kurzweil et al. (1990), McCarthy (2007), Minsky (1969), Nilsson (2009), and Russell and Norvig (2020). For surveys of AI definitions, see Legg and Hutter (2007a, 2007b) and Monett and Lewis (2018). Recently, policy makers have started to develop their own definitions (European Commission, 2018; Federal Government of Germany, 2019; High-Level Expert Group on AI, 2019; Organisation for Economic Co-operation and Development [OECD], 2019; Office for AI, 2020). For more information about the term "AI" in the legal context, see Martinez (2019), Scherer (2016), Schuett (2019), and Turner (2019). More advanced AI systems have been referred to as "Transformative AI, TAI" (Dafoe, 2018; Gruetzemacher et al., 2019; Gruetzemacher & Whittlestone, 2019; Karnofsky, 2016b), "Artificial General Intelligence, AGI" (Goertzel & Pennachin, 2007; Muehlhauser, 2013), and "Superintelligence" (Bostrom, 1998, 2003b, 2014).

[81] See, for instance, DeepMind's latest progress in solving the "protein folding problem" (Jumper et al., 2020).

suffering on an astronomical scale. There seems to be a general consensus in prioritization research that positively shaping the development of AI is one of the world's most pressing problems (Gloor, 2016b; Karnofsky, 2016a; Wiblin, 2017). Even though the law seems to play an important role in this respect, there is surprisingly little legal research focused on the long-term implications of AI.[82] We have identified four areas of research which seem particularly promising: reducing existential risks from AI (Section 4.1), reducing suffering risks from AI (Section 4.2), sharing the benefits of AI (Section 4.3), and meta-research in AI (Section 4.4).

4.1 Reducing Existential Risks from AI

It has been argued that AI could pose existential risks for humanity (Bostrom, 2014; Christian, 2020; Ord, 2020; Russell, 2019).[83] Ord (2020) estimates that there is a 10% chance that AI will cause an existential catastrophe within the next 100 years. Similarly, Wiblin (2017) estimates that the risk of a serious catastrophe caused by machine intelligence within the next 100 years is between 1% and 10%. A recent survey of leading AI safety and governance researchers reveals similar estimates (Carlier et al., 2020).[84] Risks from AI have been conceptualized as (a) accident risks, (b) misuse risks, and (c) structural risks (Zwetsloot & Dafoe, 2019).[85] The following research projects detail promising mechanisms through which the law could help to reduce each of these risks.[86]

[82] Notable exceptions include Flynn (2020), Liu et al. (2018), Maas (2019a, 2019b), O'Keefe (2018, 2020a, 2020b), and O'Keefe et al. (2020).

[83] Recall that "existential risks" are risks where an adverse outcome would either annihilate Earth-originating intelligent life, or permanently and drastically curtail its potential (Bostrom, 2002). For more information on existential risks, see Section 3.2.1.

[84] Note that subjective probability estimates of existential catastrophes should be taken with a grain of salt (Baum, 2020b; Beard et al., 2020a; see also Morgan, 2014). We therefore advise against putting too much emphasis on the precise numbers. However, the estimates do suggest that leading experts think that the probability is sufficiently high to take the risks seriously.

[85] It is worth noting that accident and misuse risks are dichotomous (unintentional vs. intentional harm), whereas structural risks can overlap with both accident and misuse risks.

[86] For a more general analysis of potential responses to extinction risks, see Cotton-Barratt et al. (2020).

RESEARCH PROJECTS

4.1.1 Reducing Accident Risks

"AI accidents" can be defined as any unintended and harmful behavior of an AI system (Amodei et al., 2016).[87] Specific scenarios in which AI accidents cause an existential catastrophe have been described by Bostrom (2014) and Yudkowsky (2008a),[88] as well as Christiano (2019).[89] A major challenge in all scenarios is to ensure that advanced AI systems are properly aligned with human values (Bostrom, 2014; Christian, 2020; Christiano, 2018b; Gabriel, 2020; Russell, 2019; Soares, 2016a; Soares & Fallenstein, 2014; Taylor et al., 2016). This problem, which is typically called the "alignment problem," involves a technical and a normative challenge (Gabriel, 2020).

The technical challenge is how to encode values in a given AI system so that it reliably does what it ought to do.[90] Proposed solutions include "iterated amplification" (Christiano, 2018; Cotra, 2018) and "debate" (Irving et al., 2018), though the problem ultimately remains unsolved. The law can help to ensure that these or

[87] Accident risks can be further broken down into (a) specification problems, (b) robustness problems, and (c) assurance problems (Ortega & Maini, 2018). Specification ensures that an AI system's behavior aligns with the operator's true intentions. For more information on specification problems, see Clark and Amodei (2016), Everitt et al. (2019), Krakovna et al. (2019a, 2019b), and Leike et al. (2018). Robustness ensures that an AI system continues to operate within safe limits upon encountering perturbations. For more information on robustness problems, see García and Fernández (2015), Goodfellow et al. (2015), Kohli et al. (2019), Quiñonero-Candela et al. (2009), and Szegedy et al. (2014). Assurance ensures that we can understand and control AI systems during operations. For more information on assurance problems, see Orseau and Armstrong (2016) and Doshi-Velez and Kim (2017).

[88] In this scenario a single AI system with goals that are hostile to humanity quickly becomes sufficiently capable of complete world domination and causes the future to contain very little of what we value. The scenario has been criticized, among others, by Baum (2018b), Baum et al. (2017), Calo (2017), Christiano (2018a), Davis and Marcus (2019), Drexler (2019), Goertzel (2015), and Shah (2018). For reviews of *Superintelligence* in academic journals, see Brundage (2015), Thorn (2015), and Thomas (2016). For informal discussion, see Fodor (2018) and Grace (2014).

[89] This scenario, which Christano refers to as "part 2," involves multiple AIs accidentally being trained to seek influence, and then failing catastrophically once they are sufficiently capable, causing humans to become extinct or otherwise permanently lose all influence over the future. For informal discussion, see Hubinger et al. (2019) and in parts Carlier and Davidson (2020). See Manheim (2019) on the dynamics that make the multi-agent scenario more complex and difficult to understand even in the short run.

[90] Bostrom (2014, p. 185) calls this the "value loading problem."

other solutions are actually implemented or slow down the development before certain safety standards are met. For example, there could be corresponding AI safety regulations.[91] How should such regulations be formed? Will EU regulation diffuse globally via the so-called “Brussels effect” (Bradford, 2020), or will there be a global race to the bottom with regards to minimum safety standards (Askell et al., 2019; Smuha, 2019)? Is there a need for new regulatory bodies (Calo, 2014; Erdélyi & Goldsmith, 2018; Scherer, 2016)? How should the scope of AI safety regulations be defined (Schuett, 2019)? Do we need new regulatory instruments (Clark & Hadfield, 2018)? How can compliance be monitored and enforced? Is there a need for stronger forms of supervision (Bostrom, 2019; Garfinkel, 2018)? If so, would they violate civil rights and liberties? What is the relationship between hard and soft law (Villasenor, 2020)? In particular, what role should professional self-regulation (O’Keefe, 2020a) and other forms of soft-law play (Cihon, 2019; Cihon et al., 2020; Jobin et al., 2019)? Do existing criminal law provisions penalize the (concrete or abstract) increase of existential accident risks (e.g., Section 221 of the German Criminal Code)? How effective are liability regimes to tackle existential accident risks? Which other legal mechanisms are conceivable (e.g., Farquhar et al., 2017)?

The normative challenge is what values, if any, we ought to encode in a given AI system. A possible answer to this question is to use some aggregate of the ethical views of society (Baum, 2017).[92] How can legal research contribute to the related challenges, such as whose ethical views to include, how to identify their views, and how to combine individual views to a single view? What can we learn from techniques to balance conflicting legal interests, such as the principles of “proportionality” or “balancing” respectively? To what extent can the law itself be used as a proxy for desirable values?

[91] See the discussion around the “White Paper on AI” (European Commission, 2020) in the EU, for example, Abecassis et al. (2020), Belfield et al. (2020), Centre for the Governance of AI (2020), and Future of Life Institute (2020), as well as the “Ethics Guidelines for Trustworthy AI” (High-Level Expert Group on AI, 2019), for example, Avin and Belfield (2019). Also see the responses to planned government regulation in the UK, for example, Beard et al. (2017), Belfield and Ó hÉigeartaigh (2017), Belfield et al. (2020), and Cave (2017).

[92] More precisely, one could seek to have the AI derive its values from the values of other ethical agents. This mechanism has been called “coherent extrapolated volition” (Bostrom, 2014; Muehlhauser & Helm, 2012; Yudkowsky, 2004). Alternatively, one could follow a “bottom-up” approach, i.e., AI designed to learn ethics as it interacts with its environment and with other ethical agents (Allen et al., 2000; Allen et al., 2005; Wallach & Allen, 2008; Wallach et al., 2008).

4.1.2 Reducing Misuse Risks

"AI misuse" means any use of an AI system with the intention of causing harm (Brundage et al., 2018).[93] A possible risk scenario involves a malevolent actor (for example, a terrorist organization or rogue state) who gains control over powerful AI-based weapons (for example, lethal autonomous weapons). How can the law be used to reduce existential risks in this scenario? In particular, what role should criminal law and law enforcement play? Is there a need to legally restrict certain types of scientific knowledge to prevent malevolent actors from gaining control over potentially dangerous AI technologies (Bostrom, 2017; Ovadya & Whittlestone, 2019; Shevlane & Dafoe, 2020; Whittlestone & Ovadya, 2020)? If so, how could this be done most effectively? To what extent is restricting scientific knowledge consistent with the relevant provisions of constitutional law?

Another misuse scenario involves an authoritarian government that uses AI-based surveillance techniques to permanently suppress opposition (Caplan, 2011; Garfinkel, 2018; Ord, 2020; Young et al., 2019). For such Orwellian surveillance states, the term "digital authoritarianism" has been coined. If they lock in the conditions for welfare on an extremely low level, they could constitute an existential risk (see Section 3.2.1). How can the law prevent the emergence of such regimes? Should certain surveillance techniques be banned?[94] Which limits does constitutional law place on the use of facial recognition technologies for state surveillance purposes (Ferguson, 2019)? Inversely, in which cases can stronger forms of state surveillance be justified in order to reduce other types of risk (Bostrom, 2019; Garfinkel, 2018)? How should international law respond to such threats?

The judicial system will likely play an important role in a digital authoritarian state. With the development of advanced artificial judicial intelligence (Winter, 2021a), values, laws, and other norms could be implemented into a primarily AI-based judiciary that becomes resistant to change. This type of lock-in effect has been called "technological-legal lock-in" (Crootof, 2019) and has been argued to result from current limitations of AI systems to adapt to social changes and institutional factors such as path dependence (Bernstein, 2006; Crootof, 2019; Re & Solow-Niederman, 2019). How does this conception of technological-legal lock-in scale with advancements in AI capabilities and potential solutions to the alignment problem, in particular to the normative challenge (Gabriel, 2020)? What other

[93] Brundage et al. (2018) prefer the term "malicious use," but there seems to be no difference. For more information on misuse risks, see Belfield (2019), Dafoe (2018), and Karnofsky (2016a).

[94] In the US, some municipalities have already started to ban state use of facial recognition technology for law enforcement purposes, including San Francisco (Conger et al., 2019) and Boston (Johnson, 2020).

institutional factors contribute to technological-legal lock-in? Which challenges would artificial judicial decision-making pose for liberal democracy (Winter, 2021a)? How can we uphold liberal democratic values in general and the separation of powers in particular within an AI judiciary? How should these long-term risks be balanced with potential short-term benefits, such as improved access to justice, transparency and fairness (Winter, 2020a; Winter, 2021a)? Which other long-term effects from AI in the judiciary are conceivable (Hollman et al., 2021)?

4.1.3 Reducing Structural Risks

AI could also shape the broader environment in harmful ways that do not fall into the accident-misuse dichotomy. These risks have been called "structural risks" (Zwetsloot & Dafoe, 2019). They typically result from the destabilizing effects of AI and could also be seen as risk factors (see Section 3.2.1).[95] A possible scenario involves some kind of war exacerbated by developments in AI (Aguirre, 2020; Allen & Chan, 2017; Avin & Amadae, 2019; Boulanin et al., 2020; Dafoe, 2018; Geist & Lohn, 2018; Horowitz, 2019; Horowitz et al., 2019; Jayanti & Avin, 2020; Lieber & Press, 2017; Maas, 2019a, 2019b).[96] For example, if AI systems could be used to detect retaliation capabilities, the equilibrium of mutual assured destruction would be disturbed, which would drastically increase the risk of a nuclear war (Bostrom, 2019; Horowitz, 2019; Lieber & Press, 2017). How effective are international treaties at banning certain AI applications (Castel & Castel, 2016; Maas, 2019a; Nindler, 2019; Wilson, 2013)? Can operators of lethal autonomous weapons be held criminally responsible (Bo, 2020)? How else can the law be used to reduce structural risks in a war scenario?

Race dynamics are another destabilizing factor (Armstrong et al., 2016; Askell et al., 2019; Bostrom, 2017; Hogarth, 2018; Naudé & Dimitri, 2020; Soares, 2016b). If competing actors think that AI could lead to some kind of economic, military or technological supremacy, and gains from AI result from their relative strength over other actors, then a race dynamic will commence in which actors might be willing to sacrifice safety in order to "win the race" (Askell et al., 2019). Such a dynamic could increase the risk that advanced AI systems are unaligned, thereby increasing the risk of an existential accident. How can the law reduce such race dynamics?

[95] For more information on AI risk factors, see Hernández-Orallo et al. (2019) and Burden & Hernández-Orallo (2020).

[96] Another scenario has been described in Part 1 of *What failure looks like* (Christiano, 2019). This scenario involves multiple AIs pursuing easy-to-measure goals, rather than the goals humans actually care about, causing us to permanently lose some influence over the future. For informal discussion, see Clarke (2020), Grue_Slinky (2019), Hanson (2019), and Pace (2020).

Which legal mechanisms can help to increase trust among competing actors (Brundage et al., 2020)? For example, there could be regulations intended to prevent a race to the bottom with regards to minimum safety standards (Smuha, 2019). There could also be auditing and certifications schemes (Cihon et al., 2020), or contractual obligations to develop AI responsibly (Askell et al., 2019). What are the most effective means to reduce structural risk?

As governments realize the power of AGI, they may seek to gain control over its development and deployment, leading to a new kind of geopolitics which has been referred to as "AI nationalism" (Hogarth, 2018). Increasing economic and political tensions between states like the US and China could then increase other types of risks, such as the risk of great power wars. How can the law reduce such tensions and foster cooperation between states? How effective are economic treaties at preventing related protectionist trade policies? How can the law help to make AI a global public good (Hogarth, 2018)? Does this require a new global organization (Cihon et al., 2020a, 2020b; Erdélyi & Goldsmith, 2018; Kemp et al., 2019)? It is worth noting that private actors currently dominate AI research and development, which leads to the question of who should govern the development of advanced AI systems (Leung, 2019). When is governmental control desirable (Leung, 2018) and what form should it take? To the extent that government control or influence is undesirable, which modes of influence (O'Keefe, 2020b) and possible defensive measures exist? Under what circumstances would it be preferable if governments were unaware of the development of advanced AI systems?

EXISTING ACADEMIC LITERATURE

Allen, G., & Chan, T. (2017, July). *Artificial intelligence and national security*. Belfer Center for Science and International Affairs, Harvard Kennedy School. https://www.belfercenter.org/publication/artificial-intelligence-and-national-security

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv. https://arxiv.org/abs/1606.06565

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI and Society*, *31*, 201–206. https://doi.org/10.1007/s00146-015-0590-y

Askell, A., Brundage, M., & Hadfield, G. (2019). *The role of cooperation in responsible AI development*. arXiv. https://arxiv.org/abs/1907.04534

Avin, S., & Amadae, S. M. (2019). Autonomy and machine learning as risk factors at the interface of nuclear weapons, computers and people. In V. Boulanin, (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk* (pp. 105–118). Stockholm International Peace Research Institute. https://doi.org/10.17863/CAM.44758

Baum, S. D. (2017). Social choice ethics in artificial intelligence. *AI and Society*, *35*, 165–176. https://doi.org/10.1007/s00146-017-0760-1

Bernstein, G. (2006). When new technologies are still new: Windows of opportunity for privacy protection. *Villanova Law Review*, *51*(4), 921–950. https://digitalcommons.law.villanova.edu/vlr/vol51/iss4/8

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

Bostrom, N. (2017). Strategic implications of openness in AI development. *Global Policy*, *8*(2), 135–148. https://doi.org/10.1111/1758-5899.12403

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455–476. https://doi.org/10.1111/1758-5899.12718

Boulanin, V., Saalman, L., Topychkanov, P., Su, F., & Carlsson, M. P. (2020). *Artificial intelligence, strategic stability and nuclear risk*. Stockholm International Peace Research Institute. https://www.sipri.org/publications/2020/other-publications/artificial-intelligence-strategic-stability-and-nuclear-risk

Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. Oxford University Press.

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

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

Calo, R. (2014, September 15). *The case for a federal robotics commission*. Brookings Institution. https://www.brookings.edu/research/the-case-for-a-federal-robotics-commission

Caplan, B. (201). The totalitarian threat. In N. Bostrom, & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 504–519). Oxford University Press.

Carlier, A., Clarke, S., & Schuett, J. (2020). *AI risk survey* [Unpublished manuscript].

Castel, J.-G., & Castel, M. E. (2016). The road to artificial superintelligence: Has international law a role to play? *Canadian Journal of Law and Technology*, *14*(1), 1–15. https://ojs.library.dal.ca/CJLT/article/view/7211

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W. W. Norton & Company.

Christiano, P., Shlegeris, B., & Amodei, D. (2018). *Supervising strong learners by amplifying weak experts*. arXiv. https://arxiv.org/abs/1810.08575

Cihon, P. (2019). *Standards for AI governance: International standards to enable global coordination in AI research & development* [Technical report]. Center for the Governance of AI, Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Cihon, P., Maas, M. M., & Kemp, L. (2020a). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy*, *11*(5). https://doi.org/10.1111/1758-5899.12890

Cihon, P., Maas, M. M., & Kemp, L. (2020b). Should artificial intelligence governance be centralised? Design lessons from history. *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–234. https://doi.org/10.1145/3375627.3375857

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., & Baum, S. D. (2020). *AI certification: Advancing ethical practice by reducing information asymmetries* [Manuscript submitted for publication].

Clarke, J., & Hadfield, G. K. (2018). *Regulatory markets for AI safety*. arXiv. https://arxiv.org/abs/2001.00078

Crootof, R. (2019). "Cyborg Justice" and the risk of technological-legal lock-in. *Columbia Law Review Forum*, *119*(7), 233–251. https://columbialawreview.org/content/cyborg-justice-and-the-risk-of-technological-legal-lock-in

Dafoe, A. (2018). *AI governance: A research agenda*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf

Erdélyi, O. J., & Goldsmith, J. A. (2018). Regulating artificial intelligence: Proposal for a global solution. *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101. https://doi.org/10.1145/3278721.3278731

Farquhar, S., Cotton-Barratt, O., & Snyder-Beattie, A. (2019). Pricing externalities to balance public risks and benefits of research. *Health Security*, *15*(4), 401–408. https://doi.org/10.1089/hs.2016.0118

Ferguson, A. G. (2019). Facial recognition and the fourth amendment. *Minnesota Law Review*, *105*. http://dx.doi.org/10.2139/ssrn.3473423

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*. https://doi.org/10.1007/s11023-020-09539-2

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

Hollman, N., Winter, C. K., & Jauhar, A. (2021). *Long-term challenges of AI for the judiciary* [Unpublished manuscript].

Horowitz, M. C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies*, *42*(6), 764–788. https://doi.org/10.1080/01402390.2019.1621174

Horowitz, M. C., Scharre, P., & Velez-Green, A. (2019). *A stable nuclear future? The impact of autonomous systems and artificial intelligence*. arXiv. https://arxiv.org/abs/1912.05291

Irving, G., Christiano, P., & Amodei, D. (2018). *AI safety via debate*. arXiv. https://arxiv.org/abs/1805.00899

Jayanti, A., & Avin, S. (2020). *It takes a village: The shared responsibility of "raising" an autonomous weapon*. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/resources/it-takes-village

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

Kemp, L., Cihon, P., Maas, M. M., Belfield, H. Ó hÉigeartaigh, S., Leung, J., & Cremer, C. Z. (2019). *UN high-level panel on digital cooperation: A proposal for international AI governance*. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/resources/proposal-international-ai-governance

Leung, J. (2019). *Who will govern artificial intelligence? Learning from the history of strategic politics in emerging technologies* [Doctoral dissertation]. University of Oxford. https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665

Liu, H.-Y., Lauta, K. C., & Maas, M. M. (2018). Governing boring apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures*, *102*, 16–19. https://doi.org/10.1016/j.futures.2018.04.009

Maas, M. M. (2019a). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy*, *40*(3), 285–311. https://doi.org/10.1080/13523260.2019.1576464

Maas, M. M. (2019b). Innovation-proof global governance for military artificial intelligence? *Journal of International Humanitarian Legal Studies*, *10*(1), 129–157. https://doi.org/10.1163/18781527-01001006

Naudé, W., & Dimitri, N. (2020). The race for an artificial general intelligence: Implications for public policy. *AI and Society*, *35*, 367–379. https://doi.org/10.1007/s00146-019-00887-x

Nindler, R. (2019). The United Nation's capability to manage existential risks with a focus on artificial intelligence. *International Community Law Review*, *21*(1), 5–34.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.

Ovadya, A., & Whittlestone, J. (2019). *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*. arXiv. https://arxiv.org/abs/1907.11274

O'Keefe, C. (2020a). *Antitrust-compliant AI industry self-regulation* [Unpublished manuscript]. https://cullenokeefe.com/blog/antitrust-compliant-ai-industry-self-regulation

O'Keefe, C. (2020b). *How will national security considerations affect antitrust decisions in AI? An examination of historical precedents* [Technical report]. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/How-Will-National-Security-Considerations-Affect-Antitrust-Decisions-in-AI-Cullen-Okeefe.pdf

Re, R. M., & Solow-Niederman, A. (2019). Developing artificially intelligent justice. *Stanford Technology Law Review*, *22*(2), 242–289.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin Random House.

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law and Technology*, *29*(2), 353–400. http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf

Schuett, J. (2019). *A legal definition of AI*. arXiv. https://arxiv.org/abs/1909.01095

Shevlane, T., & Dafoe, A. (2020). The offense-defense balance of scientific knowledge: Does publishing AI research reduce misuse? *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 173–179. https://doi.org/10.1145/3375627.3375815

Smuha, N. A. (2019). *From a 'Race to AI' to a 'Race to AI Regulation': Regulatory competition for artificial intelligence*. SSRN. http://dx.doi.org/10.2139/ssrn.3501410

Soares, N. (2016a). *The value learning problem*. Machine Intelligence Research Institute. https://intelligence.org/files/ValueLearningProblem.pdf

Soares, N., & Fallenstein, B. (2014). Agent foundations for aligning machine intelligence with human interests: A technical research agenda. In V. Callaghan, J. Miller, R.

Yampolskiy, & S. Armstrong (Eds.), *The technological singularity: Managing the journey* (pp. 103–125). Springer. https://doi.org/10.1007/978-3-662-54033-6_5

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). *Alignment for advanced machine learning systems*. Machine Intelligence Research Institute. https://intelligence.org/files/AlignmentMachineLearning.pdf

Villasenor, J. (2020, July 31). *Soft law as a complement to AI regulation*. Brookings Institution. https://www.brookings.edu/research/soft-law-as-a-complement-to-ai-regulation

Whittlestone, J., & Ovadya, A. (2020). *The tension between openness and prudence in responsible AI research*. arXiv. https://arxiv.org/abs/1910.01170

Wilson, G. (2013). Minimizing global catastrophic and existential risks from emerging technologies through international law. *Virginia Environmental Law Journal*, *31*(2), 307–364. https://www.jstor.org/stable/44679544

Winter, C. K. (2020a). The value of behavioral economics for EU judicial decision-making. *German Law Journal*, *21*(2), 240–264. https://doi.org/10.1017/glj.2020.3

Winter, C. K. (2021a). Exploring the challenges of artificial judicial decision-making for liberal democracy [Forthcoming]. In P. Bystranowski, P. Janik, & M. Próchnicki (Eds.), *Judicial decision-making: Integrating empirical and theoretical perspectives*. https://www.christophwinter.net/s/AI-Judiciary.pdf

Young, M., Katell, M., & Krafft, P. M. (2019). Municipal surveillance regulation and algorithmic accountability. *Big Data and Society*, *6*(2), 1–14 https://doi.org/10.1177/2053951719868492

Yudkowsky, E. (2008a). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom, & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308–345). Oxford University Press. https://intelligence.org/files/AIPosNegFactor.pdf

EXISTING INFORMAL DISCUSSION

Aguirre, A. (2020, November 11). *Why those who care about catastrophic and existential risk should care about autonomous weapons* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/oR9tLNRSAep293rr5/why-those-who-care-about-catastrophic-and-existential-risk-2

Bo, M. (2020, December 18). *Meaningful human control over autonomous weapon systems: An (international) criminal law account*. Opinio Juris. http://opiniojuris.org/2020/12/18/meaningful-human-control-over-autonomous-weapon-systems-an-international-criminal-law-account

Christiano, P. (2018, April 7). *Clarifying "AI alignment*". AI Alignment. https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6

Christiano, P. (2019, March 17). *What failure looks like [*Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like

Cotra, A. (2018, March 4). *Iterated distillation and amplification*. AI Alignment. https://ai-alignment.com/iterated-distillation-and-amplification-157debfd1616

Garfinkel, B. (2018, October 12). *The future of surveillance*. Effective Altruism. https://www.effectivealtruism.org/articles/ea-global-2018-the-future-of-surveillance

Hogarth, I. (2018, June 13). *AI nationalism.* https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism

Karnofsky, H. (2016a, May 6). *Potential risks from advanced artificial intelligence: The philanthropic opportunity.* Open Philanthropy. https://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity

Leung, J. (2018, September 28). *Analyzing AI actors.* Effective Altruism. https://www.effectivealtruism.org/articles/ea-global-2018-analyzing-ai-actors

Soares, N. (2016b, July 23). *Submission to the OSTP on AI outcomes.* Machine Intelligence Research Institute. https://intelligence.org/2016/07/23/ostp

Wiblin, R. (2017b, March). *Positively shaping the development of artificial intelligence.* 80,000 Hours. https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence

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

4.2 Reducing Suffering Risks from AI

AI could cause suffering on an astronomical scale (Althaus & Baumann, 2020; Althaus & Gloor, 2016; Baumann, 2017a, 2017b, 2018a, 2018b; Clifton, 2020; Daniel, 2017; Gloor, 2016b; Tomasik, 2018, 2019b).[97] If a state of astronomical suffering is permanently locked in, it could be worse than extinction, making such scenarios the worst kind of existential risks (Daniel, 2017). Against this backdrop, it is worrying that suffering risks (s-risks) from AI are highly neglected, especially in legal research. Given our high degree of uncertainty, disentanglement research seems particularly important (see Flynn, 2017). Besides that, we think that the following research directions are worth considering.

Research Projects

4.2.1 Near Misses

A potential s-risk scenario involves an AGI which is only slightly misaligned with human values (Tomasik, 2018).[98] Such a scenario, which has been called "near miss," could cause astronomical amounts of suffering. For example, suppose an AGI has the goal of creating as many "happy minds" as possible, but its slightly askew interpretation of this goal results in vast numbers of minds with severe mental-

[97] Recall that "suffering risks" are risks where an adverse outcome would bring about suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far (Althaus & Gloor, 2016). For more information on suffering risks, see Section 3.2.1.

[98] For more information on the "alignment problem," see Section 4.1.

health problems like depression or anxiety. We have already outlined potential ways in which the law could help to solve the alignment problem in Section 4.1.

4.2.2 Mind Crime

S-risks could also result from situations in which artificial minds are made to suffer for instrumental purposes, for instance in order to simulate evolution or perform experiments (Tomasik, 2019b). This scenario has been called "mind crime" (Bostrom, 2014, p. 125). If one assumes that artificial minds have a relevant moral status (Danaher, 2019; Gunkel, 2018; Schwitzgebel & Garza, 2015; Shulman & Bostrom, 2020; Tomasik, 2014), and that there could be vast numbers of them, suffering can reach astronomical scales. What is the threshold above which artificial minds should be legally protected (Chesterman, 2020; Hubbard, 2011; Kurki, 2019)? How should the law deal with uncertainties about their moral status (cf. MacAskill et al., 2020)? What can we learn from the related debate on animal welfare (see Section 9)?

4.2.3 Agential S-Risks

"Agential s-risks" involve agents that actively and intentionally want to cause harm (Althaus & Baumann, 2020; Baumann, 2017b, 2018b).[99] It seems at least somewhat plausible that artificial agents might exhibit behavior that resembles malevolent traits like psychopathy or sadism.[100] Their occurrence in some humans suggests that they may have provided evolutionary fitness advantages (Book et al., 2015; Jonason et al., 2015; McDonald et al., 2012; Nell, 2006). If these traits prove useful in a given environment, then advanced AI systems that are trained on this environment might learn corresponding behavior with potentially catastrophic consequences. For example, an agent might cause suffering as a strategic threat in an escalating conflict (Baumann, 2018b). One possible intervention would be to expand the scope of extortion laws. To the extent that the agent making such threats is controlled by states, international treaties banning such strategies could be another lever. Besides that, it is unclear how the law could reduce such risks. There is a need for exploratory research that structures the problem and identifies relevant questions for legal research.

[99] This is the s-risk equivalent of existential misuse risks as outlined in Section 4.1.2.

[100] Note that one should not make the mistake of anthropomorphizing AI (see Salles et al. 2020). The notion of malevolence might be of limited value in the context of an artificial agent (Althaus & Baumann, 2020).

EXISTING ACADEMIC LITERATURE

Althaus, D., & Gloor, L. (2016, September). *Reducing risks of astronomical suffering: A neglected priority*. Center on Long-Term Risk. https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority

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

Chesterman, S. (2020). Artificial intelligence and the limits of legal personality. *International & Comparative Law Quarterly*, *69(4)*, 819–844. https://doi.org/10.1017/S0020589320000366

Clifton, J. (2020, March). *Cooperation, conflict, and transformative artificial intelligence: A research agenda*. Center on Long-Term Risk. https://longtermrisk.org/research-agenda

Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, *26*, 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Gloor, L. (2016b, November). *Altruists should prioritize artificial intelligence*. Center on Long-Term Risk. https://longtermrisk.org/altruists-should-prioritize-artificial-intelligence

Gunkel, D. J. (2018). *Robot Rights*. MIT Press.

Hubbard, F. P. (2011). "Do Androids Dream?": Personhood and intelligent artifacts. *Temple Law Review*, *83*(2), 405–474. https://www.templelawreview.org/article/83-2_hubbard

Kurki, V. A. J. (2019). *A theory of legal personhood*. Oxford University Press.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, *39*, 98–119. https://doi.org/10.1111/misp.12032

Tomasik, B. (2014). *Do artificial reinforcement-learning agents matter morally?* arXiv. https://arxiv.org/abs/1410.8233

Tomasik, B. (2019b, July 2). *Risks of astronomical future suffering*. Center on Long-Term Risk. https://longtermrisk.org/risks-of-astronomical-future-suffering

EXISTING INFORMAL DISCUSSION

Althaus, D., & Baumann, T. (2020, April 29). *Reducing long-term risks from malevolent actors* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/LpkXtFXdsRd4rG8Kb/reducing-long-term-risks-from-malevolent-actors

Baumann, T. (2017a, September 16). *Focus areas of worst-case AI safety*. Reducing Risks of Future Suffering. https://s-risks.org/focus-areas-of-worst-case-ai-safety

Baumann, T. (2017b, December 15). *Using surrogate goals to deflect threats*. Reducing Risks of Future Suffering. https://s-risks.org/using-surrogate-goals-to-deflect-threats

Baumann, T. (2018a, July 5). *An introduction to worst-case AI safety*. Reducing Risks of Future Suffering. https://s-risks.org/an-introduction-to-worst-case-ai-safety

Baumann, T. (2018b). *A typology of s-risks*. Center for Reducing Suffering. http://centerforreducingsuffering.org/a-typology-of-s-risks

Daniel, M. (2017). *S-risks: Why they are the worst existential risks, and how to prevent them (EAG Boston 2017)*. Center on Long-Term Risk. https://longtermrisk.org/s-risks-talk-eag-boston-2017

Tomasik, B. (2018, December 13). *Astronomical suffering from slightly misaligned artificial intelligence*. Essays on Reducing Suffering. https://reducing-suffering.org/near-miss

4.3 Sharing the Benefits of AI

AI could create wealth on an astronomical scale with far-reaching implications for every sector of the economy (Bostrom, 2003a; Hanson, 2001; Makridakis, 2017; Trajtenberg, 2018; Trammel & Korinek, 2020). However, by default those benefits may be captured by a small set of actors, and due to some lock-in effects, the initial distribution of wealth may be hard to change in certain circumstances. If the initial distribution is suboptimal, humanity could permanently lose a significant fraction of its potential, thus constituting a p-risk (see Section 3.2.1). The question of how the gains of AI ought to be distributed—and how to design mechanisms to approach an ideal distribution of gains—may therefore be one of the most important economic questions of our time.

Research Projects

4.3.1 Distributing Windfall Profits

It seems plausible that AI will enable the accumulation of unprecedented wealth in the hands of a few firms. “Windfall profits” are profits greater than a substantial fraction of the world’s total economic output (O’Keefe et al., 2020). How should these profits be distributed? A possible solution is the so-called “Windfall Clause,” a voluntary but binding agreement to donate a meaningful portion of profits if they earn a historically unprecedented economic windfall from the development of advanced AI (Bostrom, 2014; O’Keefe et al., 2020). Which other mechanisms are conceivable (see also the Shared Prosperity Initiative)?

4.3.2 Economic Regulation of AI

Technology industries are highly concentrated (Varian, 2001) and AI services may have features of a natural monopoly. Many competition authorities are therefore concerned with avoiding harm to consumers and deadweight loss associated with monopolized AI markets, especially if these markets dominate the world economy. How can antitrust/competition law (U.S. House Judiciary Subcommittee on Antitrust, Commercial and Administrative Law, 2020), utility ratemaking, and other options be used as a tool to check the power of large AI companies, and avoid excessive pricing of AI services without excessively reducing incentives to innovate (Belfield, 2020b; Hua & Belfield, 2020; O’Keefe, 2020b; see also Khan, 2016)? Another promising area concerns investor-state treaty disputes. As large AI companies and governments might use private arbitration to resolve disputes, how can we ensure that important implications for the long-term future are duly taken into consideration?

4.3.3 Corporate Governance and Firm Incentives

Firms' incentives shape their behavior. Still, profit-maximization alone seems unlikely to be the best incentive structure for firms aiming to develop advanced AI systems. What other firm structures might be desirable to ensure that safety and ethical concerns are given due consideration (Brockman et al., 2019; Feldman et al., 2020)? How can employees (Belfield, 2020a), investors (Belfield, 2020c, 2020d) and other actors (Cihon, et al., 2020) influence corporate decision-making? In particular, which legal instruments are at their disposal (for example, unionization, shareholder resolutions, replacing the board of directors)?

4.3.4 International Coordination and Distribution of Benefits

AI development is concentrated in a small number of already-wealthy countries, but is likely to affect the entire world in the long-run. A maximally beneficial distribution of the benefits from AI will necessarily cross borders (Ó hÉigeartaigh et al., 2020). Yet it is unclear whether existing international institutions responsible for equitably distributing benefits from AI are adequate for this task. What would adequate institutions look like? Will their form and mission vary geographically, and if so, how? How would they interact with governments, NGOs, private AI developers, and existing international bodies? What would beneficiaries' rights against such distributor bodies be?

4.3.5 Intellectual Property

IP regimes may have a significant influence over the development of advanced AI. AI is expensive to produce (Amodei & Hernandez, 2018), but comparatively cheap to copy once produced, making it a prototypical candidate for IP protections. Yet, the IP protections for AI are currently patchwork (Calvin & Leung, 2020), unsettled, and evolving. Reliance on trade secrets also means that AI may be protected indefinitely, unlike copyrighted or patented systems, thus potentially depriving the general public of gains from lower-cost copies of original systems after IP protections expire. It may also create difficulties for regulatory auditing of algorithms (Kroll et al., 2017, p. 658; Tsamados et al., 2020, p. 18). Furthermore, the data-intensity of training AI systems raises questions about infringement during training (O'Keefe et al., 2019). Structuring the IP of AI systems properly may influence both the rate of progress in the field and the magnitude and distribution of economic gains from IP-protected systems. Are the current IP regimes adequate to balance incentives for innovation and widespread adoption, or ought they be revised to accommodate for unique dynamics in AI? If so, will existing international IP treaties allow such tailoring?

EXISTING ACADEMIC LITERATURE

Belfield, H. (2020a). Activism by the AI community: Analysing recent achievements and future prospects. In *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 15–21). https://doi.org/10.1145/3375627.3375814

Belfield, H. (2020b). *From tech giants to a tech colossus: Antitrust objections to the Windfall Clause* [Manuscript submitted for publication].

Belfield, H. (2020c). *Financing our final hour (Pt. I): Institutional investors' Obligations to manage global risks* [Unpublished manuscript].

Belfield, H. (2020d). *Financing our final hour (Pt. II): Institutional investors' strategies for managing global risks* [Unpublished manuscript].

Bostrom, N. (2003a). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, *15*(3), 308–314. https://doi.org/10.1017/S0953820800004076

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

Calvin, N., & Leung, J. (2020). *Who owns artificial intelligence? A preliminary analysis of corporate intellectual property strategies and why they matter* (Working paper). Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-working-paper-Who-owns-AI-Apr2020.pdf

Cihon, P., Schuett, J., & Baum, S. D. (2020). *Corporate governance of artificial intelligence in the public interest* [Manuscript submitted for publication].

Hanson, R. (2001). *Economic growth given machine intelligence* (Technical Report). University of California, Berkeley. http://mason.gmu.edu/~rhanson/aigrow.pdf

Hua, S.-S., & Belfield, H. (2020). *AI & antitrust: Reconciling tensions between competition law and cooperative AI development* [Manuscript submitted for publication].

Khan, L. M. (2016). Amazon's antitrust paradox. *Yale Law Journal*, *126*(3), 710–805. https://digitalcommons.law.yale.edu/ylj/vol126/iss3/3

Korinek, A. (2019). *Integrating ethical values and economic value to steer progress in artificial intelligence* [Working paper]. National Bureau of Economic Research. https://doi.org/10.3386/w26130

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, *165*(3), 633–705.

Makridakis, S. (2017). The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

Ó hÉigeartaigh, S., Whittlestone, J., Liu, Y., Zeng, Y., & Liu, Z. (2020). Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy & Technology*, *33*, 571–593. https://doi.org/10.1007/s13347-020-00402-x

O'Keefe, C. (2020b). *How will national security considerations affect antitrust decisions in AI? An examination of historical precedents* [Technical report]. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/How-Will-National-Security-Considerations-Affect-Antitrust-Decisions-in-AI-Cullen-OKeefe.pdf

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., & Dafoe, A. (2020, February). *The windfall clause: Distributing the benefits of AI for the common good* [Technical report]. Centre for the Governance of AI Research Report. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/windfallclause

Trajtenberg, M. (2018). *AI as the next GPT: A political-economy perspective* (Working Paper 24245). National Bureau of Economic Research. https://doi.org/10.3386/w24245

Trammel, P., & Korinek, A. (2020, October). *Economic growth under transformative AI: A guide to the vast range of possibilities for output growth, wages, and the labor share* (GPI Working Paper No. 8-2020). Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/philip-trammell-and-anton-korinek-economic-growth-under-transformative-ai

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2020). *The ethics of algorithms: Key problems and solutions*. SSRN. http://dx.doi.org/10.2139/ssrn.3662302

Varian, H. R. (2001). High-technology industries and market structure. *Proceedings – Economic Policy Symposium – Jackson Hole, Federal Reserve Bank of Kansas City*, 65–101. Archived at https://perma.cc/DZ2B-E7GT

EXISTING INFORMAL DISCUSSION

Amodei, D., & Hernandez, D. (2018, May 16). *AI and compute*. OpenAI. https://openai.com/blog/ai-and-compute

Brockman, G., Sutskever, I., & OpenAI (2019, March 11). *OpenAI LP* [Press release]. https://openai.com/blog/openai-lp

O’Keefe, C., Lansky, D., Clark, J., & Payne, C. (2019). *Before the United States Patent and Trademark Office Department of Commerce: Comment regarding request for comments on intellectual property protection for artificial intelligence, Innovation Docket No. PTO–C–2019–0038, Comment of OpenAI, LP addressing question 3.* https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf

4.4 Meta-Research in AI

Shaping the development of advanced AI involves substantial uncertainties. For example, views on AI timelines vary widely (Baum et al., 2011; Grace et al., 2018; Müller & Bostrom, 2016) and researchers disagree on why AI might pose an existential risk (Adamczewski, 2019; Carlier et al., 2020; Cottier & Shah, 2019; Dai, 2018, 2019; Garfinkel, 2018, 2020; Ngo, 2019, 2020). There is no simple answer to the question of how legal scholarship can best contribute to the raised issues. Some resources should therefore be dedicated towards “meta-research,” that is to say, addressing high-level uncertainties and methodological questions that arise in prioritizing legal research. In the following, we list promising AI-specific meta-research projects, while Section 7 concerns meta-research in general.

RESEARCH PROJECTS

4.4.1 Improving Our Ability to Shape the Development of AI in the Future

If one believes that future generations will have more effective ways to shape the development of AI, then one should consider improving their ability to do so.[101] Should we, for example, wait to regulate AI in order to prevent a regulatory backlash (Baum, 2016; Gurkaynak et al., 2016)? How can we ensure that the law remains adaptive to future AI technologies (Maas, 2019b; Moses, 2011)? In particular, how can law-related path dependencies be prevented? What measures can we take today that make governing AI in the future easier? For example, it might be useful to establish AI registers which contain detailed information about potentially harmful AI systems (Floridi, 2020). The law could also help to accumulate resources over a substantial length of time (see Trammell, 2020). To this end, what role do foundation law and tax law play?

4.4.2 Predicting How the Law Will Shape the Development of AI

Predicting AI progress is an important challenge that has received considerable attention (Armstrong & Sotala, 2015; Cremer & Whittlestone, 2020; Etzioni, 2020; Gruetzemacher, 2020; Gruetzemacher et al., 2020; Page et al., 2020). However, there is much less work, if any, that tries to predict how the law will shape the development of AI, even though the law will likely have a significant influence. How has the law shaped the development of other general purpose technologies? To what extent should regulatory impact assessments (OECD, 2009) include long-term implications of AI (see Calvo et al., 2020)?

4.4.3 Clarifying Legal Researchers' Views on the Long-Term Implications of AI

Currently, legal research is mainly concerned with legal questions about today's AI systems (for example, regarding liability, data protection, or anti-discrimination). It is unclear what their views on the long-term implications of AI are, in particular on existential risks, suffering risks, and extreme benefits. Clarifying these views, for example, by conducting specific literature reviews or surveys,

[101] For more information on the underlying view called "patient longtermism," see MacAskill (2020b), Todd (2020a), and Trammell (2020).

would therefore be a valuable research project that could unlock future research opportunities (see Section 3.2.2).

EXISTING ACADEMIC LITERATURE

Armstrong, S., & Sotala, K. (2015). How we're predicting AI—or failing to. In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond Artificial Intelligence* (pp. 11–29). Springer. https://doi.org/10.1007/978-3-319-09668-1_2

Baum, S. D. (2016). On the promotion of safe and socially beneficial artificial intelligence. *AI & Society*, *32*(4), 543–551. https://doi.org/10.1007/s00146-016-0677-0

Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an expert assessment. *Technological Forecasting and Social Change*, *78*(1), 185–195. https://doi.org/10.1016/j.techfore.2010.09.006

Calvo, R. A., Peters, D., & Cave, S. (2020). Advancing impact assessment for intelligent systems. *Nature Machine Intelligence*, *2*, 89–91. https://doi.org/10.1038/s42256-020-0151-z

Carlier, A., Clarke, S., & Schuett, J. (2020). *AI risk survey* [Manuscript in preparation].

Cremer, C. Z., & Whittlestone, J. (2020). Canaries in technology mines: Warning signs of discontinuous progress in AI. *Evaluating Progress in AI Workshop – ECAI 2020.* http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_4.pdf

Floridi, L. (2020). Artificial intelligence as a public service: Learning from Amsterdam and Helsinki. *Philosophy & Technology*, *33*, 541–546. https://doi.org/10.1007/s13347-020-00434-3

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, *62*, 729–754. https://doi.org/10.1613/jair.1.11222

Gruetzemacher, R. (2020). *Forecasting transformative AI* [Doctoral dissertation]. Auburn University. https://etd.auburn.edu/handle/10415/7338

Gruetzemacher, R., Dorner, F., Bernaola-Alvarez, N., Giattino, C., & Manheim, D. (2020). *Forecasting AI progress: A research agenda.* arXiv. https://arxiv.org/abs/2008.01848

Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling artificial intelligence: Human perils. *Computer Law & Security Review*, *32*(5), 749–758. https://doi.org/10.1016/j.clsr.2016.05.003

Maas, M. M. (2019b). Innovation-proof global governance for military artificial intelligence? *Journal of International Humanitarian Legal Studies*, *10*(1), 129–157. https://doi.org/10.1163/18781527-01001006

Moses, L. B. (2011). *Recurring dilemmas: The law's race to keep up with technological change*. SSRN. http://dx.doi.org/10.2139/ssrn.979861

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 555–572). Springer. https://doi.org/10.1007/978-3-319-26485-1_33

Page, M., Aiken, C., & Murdick, D. (2020, October). *Future indices: How crowd forecasting can inform the big picture*. Center for Security and Emerging Technology, Georgetown University. https://cset.georgetown.edu/research/future-indices

Trammell, P. (2020). *Patience and philanthropy*. Global Priorities Institute, University of Oxford. https://philiptrammell.com/static/PatienceAndPhilanthropy.pdf

EXISTING INFORMAL DISCUSSION

Adamczewski, T. (2019, May 25). *A shift in arguments for AI risk*. Fragile Credences. https://fragile-credences.github.io/prioritising-ai

Cottier, B., & Shah, R. (2019, August 15). *Clarifying some key hypotheses in AI alignment* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/mJ5oNYnkYrd4sD5uE/clarifying-some-key-hypotheses-in-ai-alignment

Dai, W. (2018, December 16). *Two neglected problems in human-AI safety* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/HTgakSs6JpnogD6c2/two-neglected-problems-in-human-ai-safety

Dai, W. (2019, February 10). *The argument from philosophical difficulty* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/w6d7XBCegc96kz4n3/the-argument-from-philosophical-difficulty

Garfinkel, B. (2019, February 9). *How sure are we about this AI stuff? [*Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/9sBAW3qKpp-noG3QPq/ben-garfinkel-how-sure-are-we-about-this-ai-stuff

Garfinkel, B. (2020, July 9). *On scrutinising classic AI risk arguments*. 80,000 Hours. https://80000hours.org/podcast/episodes/ben-garfinkel-classic-ai-risk-arguments

Ngo, R. (2019, February 21). *Disentangling arguments for the importance of AI safety* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/w6d7XBCegc96kz4n3/the-argument-from-philosophical-difficulty

Ngo, R (2020, September 28). *AGI safety from first principles [*Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ

5 Synthetic Biology and Biorisk

Synthetic biology[102] has great potential to shape the long-term future, promising numerous beneficial applications in medicine, fuel, materials science, agriculture, and other industries. Synthetic biology also poses global catastrophic risks to human-originating civilization, threatening serious loss of well-being and life on a global scale and constituting a risk factor.[103] Some extreme cases of this are

[102] There is no generally accepted definition of "synthetic biology." The term emerged at the turn of the millenia as an extension of recombinant DNA and genetic engineering in the 1970s and has continued to evolve. For an overview of its development, see National Academies of Sciences, Engineering, and Medicine (NASEM, 2018a), Chapter 2, Acevedo-Rocha (2016), and Way et al. (2014). Today, synthetic biology is frequently defined as applying engineering concepts and approaches to biology (see, e.g., Agapakis, 2014). Another common definition offers two main elements: (a) the design and construction of new biological components and systems, and (b) the redesign of existing, natural biological organisms and systems for useful purposes (Engineering Biology Research Consortium, 2020; Evans, 2014). Policy makers have surveyed and proposed their own definitions (see, e.g., European Commission, 2014; Secretariat of the Convention on Biological Diversity, 2015). For a further survey of definitions, see Nature Biotechnology (2009), and for discussions on other core principles of synthetic biology, see Benner and Sismour (2005), Benner et al. (2011), Endy (2015), Le Feuvre and Scrutton (2018), and Oldham et al. (2012).

Synthetic biology encompasses diverse tools, techniques, and applications from a variety of scientific disciplines and industries. For a discussion of uses and applications, see, for example, König et al. (2013), Pray et al. (2011), and Schmidt and Pei (2010).

[103] There are several definitions for global catastrophic risk, such as those set forth in Bostrom and Ćirković (2007, p. 1) ("The term 'global catastrophic risk' lacks a sharp definition. We use it to refer, loosely, to a risk that might have the potential to inflict serious damage to human well-being on a global scale. On this definition, an immensely diverse collection of events could constitute global catastrophes: potential candidates range from volcanic eruptions to pandemic infections, nuclear accidents to worldwide tyrannies, out-of-control scientific experiment to climate changes, and cosmic hazards to economic collapse."), Cotton-Barratt et al. (2016, p. 1) ("risk of events or processes that would lead to the deaths of approximately a tenth of the world's population, or have a comparable impact."), Millett and Snyder-Beattie (2017) ("We loosely define global catastrophic risk as being 100 million fatalities, and existential risk as being the total extinction of humanity."), Open Philanthropy (2020b) ("We use the term 'global catastrophic risks' to refer to risks that could be globally destabilizing enough to permanently worsen humanity's future or lead to human extinction."), Palmer et al.

existential risks; Ord (2020) estimates that there is a 1 in 30 chance that engineered pandemics will cause an existential catastrophe within the next 100 years.[104] Prioritization research has identified the related fields of biosecurity and governance of synthetic biology and biotechnology as major global priorities (Centre for the Study of Existential Risk, 2020; Future of Humanity Institute, 2020; Lewis, 2020; Open Philanthropy, 2020b; Watson, 2018).[105] Such governance must bridge boundaries between legal and scientific disciplines, between national and international law, between international and national geopolitical areas, and between professionals and amateurs as technology, education, and information become increasingly accessible.

This Section begins with a focus on how the law can reduce existential risk, first by minimizing the likelihood of intentional or unintentional release through preventive measures (Section 5.1) and second by minimizing the negative outcomes upon release through coordination and response (Section 5.2).[106] While we believe legal research to address these existential risks is most important, it also seems worth considering how to steer scientific research and distribute benefits and risks,

(2017), Schoch-Spana et al. (2017, p. 1) ("The Johns Hopkins Center for Health Security's working definition of *global catastrophic biological risks* (GCBRs): those events in which biological agents—whether naturally emerging or reemerging, deliberately created and released, or laboratory engineered and escaped—could lead to sudden, extraordinary, widespread disaster beyond the collective capability of national and international governments and the private sector to control. If unchecked, GCBRs would lead to great suffering, loss of life, and sustained damage to national governments, international relationships, economies, societal stability, or global security."), Yassif (2017) ("A GCR is something that could permanently alter the trajectory of human civilization in a way that would undermine its long-term potential or, in the most extreme case, threaten its survival.").

[104] For additional estimates of existential risk, see footnote 109. Note that probability estimates of existential catastrophes should be taken with caution (Beard et al., 2020a; see also Baum, 2020b; Beard et al., 2020b; Morgan, 2014; Yudkowsky, 2008b). Ord (2020) acknowledges that there is "significant uncertainty in these estimates, and they should be treated as representing the right order of magnitude" (p. 167). For a discussion of types of uncertainties in estimating natural pandemic risk, see Manheim (2018), and for an estimate that addresses those concerns, see Snyder-Beattie et al. (2019).

[105] Governance of synthetic biology has also received considerable attention from the scientific community (see, e.g., Douglas & Stemerding, 2013; Kelle, 2013; Ribeiro & Shapira, 2019; Stirling et al. 2018; Wallach, 2018) and legal community (see, e.g., Mandel & Marchant, 2014), albeit with less attention to the far future.

[106] This categorization is presented in NASEM (2018a), Chapter 8, but other, similar typologies may be useful in considering the broad range of risks and how to address them (see Avin, 2018, p. 2; Cotton-Barratt et al., 2020; Farquhar et al., 2017, p. 17; Schoch-Spana et al., 2017).

which may implicate existential risks through loss of potential, as well as pleasure risks and suffering risks (Section 5.3).[107]

5.1 Preventing Intentional or Accidental Release of a Biological Agent

The most desirable outcome is to avoid a catastrophic event entirely (Cotton-Barratt et al., 2020, p. 273). If we can prevent the intentional or accidental release[108] of a biological organism that poses catastrophic or existential risk, human-originating civilization can avoid that harm and retain resources that would have been expended in responding to and mitigating the threat. It seems worthwhile to focus on these anthropogenic risks—those arising out of human activity, such as engineered pathogens—because they may pose much greater existential risk than natural ones (see Lewis, 2020; Ord, 2020, p. 167; Sandberg & Bostrom, 2008).[109] The following avenues of research seem promising:

RESEARCH PROJECTS

5.1.1 Reducing Misuse Risks (Biowarfare, Bioterrorism)

"Misuse" here means any use of synthetic biology with the intention of causing harm. One challenge of preventing misuse in synthetic biology is the evolving risk landscape. Over time, less powerful, non-state actors may pose existential risk, as increasingly powerful tools become more available, less expensive, and easier to use.[110] How might the law address this more distributed and democratized biology

[107] Notably, now may be a particularly good time for legal research to reduce biorisk. We may be in a window of opportunity for government and private interest and policy change in light of COVID-19; however, this window may be short, focused on natural risks, and tempered by the need to respond to immediate needs (Joshi, 2020; cf. World Bank, 2017, p. 17).

[108] For a portrayal of biological risks on a spectrum ranging from natural to accidental to intentional, see Husbands (2018, Figure 1).

[109] Ord (2020) estimates that engineered pandemics are roughly 330 times more likely to cause an existential catastrophe by 2120 than naturally arising pandemics. He estimates that the x-risk from natural pandemics is 1 in 10,000 (.01%) and from engineered pandemics is 1 in 30 (3.3%).

Similar results were found in an informal survey conducted at the 2008 Oxford Global Catastrophic Risk Conference, where participants estimated that an engineered pandemic was 40 times more likely to cause human extinction by 2100. The median risk estimate of participants for natural pandemics was .05% and for engineered pandemics was 2% (Sandberg & Bostrom, 2008).

[110] Sandberg and Nelson (2020) propose a risk chain model of biorisk to identify what kinds of actors pose the greatest risk. They suggest that in the near future we may be

community? What domestic criminal and civil laws exist to deter and prevent deployment of a biological weapon, and how could they be adapted to better address threats from synthetic biology? What can be learned from deterrence approaches in political science (Knopf, 2010)?[111] How well do traditional legal mechanisms effectively reach this growing set of actors (for example, related to attribution, information hazards, dual-use concerns, and restrictions and monitoring, discussed as separate research projects)? Given the current limitations of the international legal framework to address wrongful acts by non-state actors and biorisks in general,[112] how can international institutions or instruments, such as the Biological Weapons Convention, be strengthened (Means, 2019; Scrivner, 2018; Wilson, 2013)? What new institutions or instruments are desirable?

Similarly, motivations and corresponding sources of harm can vary widely.[113] Existential catastrophe could result from pandemic pathogens (known and recreated, novel, or modified to be more dangerous), widespread eradication of food sources, modified or novel organisms with broad capacity for harm (Schoch-Spana et al., 2017), or other threats that lead to risk factors such as global conflict (Section 3.2.1). There could be erosion of norms against biowarfare that would otherwise provide deterrence, through state dynamics or non-state actions. For example, smaller, targeted biological attacks could become commonplace, similar to cyber

more concerned about highly skilled researchers or other "insider" threats, while less sophisticated actors could pose a similar threat over time, as synthetic biology becomes more accessible through less expensive and easier to use tools and methods.

[111] Deterrence may also come from other sources, such as availability and use of a vaccine and other countermeasures. Kosal (2014) argues that improving public health infrastructure could serve as a deterrent to misuse. These are discussed as tools for responding to an event in Section 5.2.4.

[112] For example, the Biological Weapons Convention allows ample room for argument that particular research or biological agents have a peaceful purpose, and no mandatory verification or enforcement mechanisms exist. There are confidence-building measures—annual declarations of critical information on research, development, and more—which were introduced "in order to prevent or reduce the occurrence of ambiguities, doubts and suspicions and in order to improve international co-operation in the field of peaceful biological ambiguities" (United Nations Office for Disarmament Affairs, 2015); however, there are few, if any, consequences for failing to participate (Chevrier & Hunger, 2000, pp. 31–32). By comparison, the Chemical Weapons Convention (CWC) allows for strict verification of compliance following mandatory destruction of all declared chemical weapons and production sites, as well as possible "challenge inspections." However, the CWC has a similar issue with dual-use, and "chemical weapon" is defined by intended purpose rather than lethality or quantity.

[113] Possible motivations could be political, economic, or sociocultural, perhaps to seek attention, make a statement, blackmail, incapacitate, destabilize, retribute, or deter (Gandhi et al., 2011; Revill, 2017, Figure 2 at pp. 630–631).

attacks with economic motivations.[114] How can the law adapt to the changing risk landscape? Would different legal mechanisms be appropriate to deter release from different motivations, and are any of these motivations more concerning or likely to pose existential risk? Is there a risk of norms against biowarfare being eroded (Ilchman & Revill, 2014), and if so, how can the law promote "biopeace"? How could this look different for international and national law?

5.1.2 Reducing Accident Risks (Biosafety)

"Accident risks" here are defined as any unintentional release of a harmful biological agent.[115] Biosafety regulations and guidelines apply to research involving infectious agents, toxins, and other biological hazards, aiming to safeguard against accidental release, ensure reporting and transparency about accidents, and provide oversight and monitoring.[116] However, some have argued that even maximum containment labs are prone to error and thus inadequate for potential pandemic pathogens (Klotz, 2019). What kind of containment, reporting, and transparency mechanisms would be more effective? What could be learned from accident reporting in other industries, such as aviation (Gronvall, 2015, p. 6), or high reliability organizations (Roberts & Bea, 2001)? Do existing criminal law provisions penalize the (concrete or abstract) increase of existential accident risks (e.g., Section 221 of the German Criminal Code; see also Duff & Marshall, 2015; Simester & von Hirsh, 2009), discussed more in Section 6.1.9? What other legal mechanisms are conceivable to reduce accident risks, such as deterrence via civil liability?

While existential risk from accidents was once limited to academic and commercial labs, it is increasingly within the reach of other groups and individuals.

[114] As synthetic biology and biological agents are used for production in materials science and other industries, those same industries will also become susceptible to biowarfare.

[115] Compare to accident risks in artificial intelligence, which encompass "any unintended and harmful behavior of an AI system" (Section 4.1.1). In the discussion of synthetic biology, accident risk focuses on the specific risk of unintentional release, while unintentional consequences are discussed separately. For an informal discussion of historical accidental release of pandemic pathogens, see Shulman (2020).

[116] While the term "biosafety" has several accepted definitions (Beeckman & Rüdelsheim, 2020), here we use it to refer specifically to principles and practices to prevent unintentional release or exposure. Biosafety guidelines commonly specify different levels of biocontainment precautions required to isolate dangerous biological agents in a facility, referred to as biosafety level (BSL), containment level (CL), or pathogen/protection level (P), with BSL-1/CL1/P1 as the lowest and BSL-4/CL4/P4 as the highest. In the United States, the Centers for Disease Control and Prevention specify these levels. The same levels are defined in the European Union Directive 2000/54/EC, Biological Agents at Work, the Canadian Biosafety Standards and Guidelines, and elsewhere (National Academy of Sciences & National Research Council, 2012, Chapter 4 & Appendix E).

Synthetic biology no longer requires years of training and experience in laboratories, where biosafety and containment protocols are accompanied by certification programs and institutional oversight. A scientist could theoretically find themselves in safety situations that exceed their biosafety experience. Powerful equipment and technologies outside of a lab may go without regular maintenance or checks and result in bio-errors. In the context of accidents, existential risk seems most likely from release of a potential pandemic pathogen. How can the law reduce existential risk from accidents outside of traditional laboratories? What role should professional self-regulation, best practices and norms (Open Philanthropy, 2017), and other forms of soft law play?

Comparative law may offer insights on potential gaps and more effective measures, yet little research exists comparing biosafety governance in different countries, let alone the relative effectiveness of different strategies. What laws and regulations exist in different countries to minimize accident risks (Beeckman & Rüdelsheim, 2020, Appendix 1; National Academy of Sciences & National Research Council, 2012, Chapter 4 & Appendix E; Osman, 2018; Van Houten & Fleming, 1993)? To what extent have they been implemented in practice?[117] How might their effectiveness be measured, and what uncertainties exist in such an analysis? What do they reflect about biosafety norms? How have different nations attempted to regulate the DIY bio community, and with what result?

5.1.3 Restrictions and Monitoring Measures

Laws that impose lab safety requirements or place other limits on research, use, or access to materials and equipment reduce existential risk by making it more difficult to develop, produce, or accidentally release the most harmful biological agents. The effectiveness of those laws depends on the ability to verify and enforce compliance. However, biological weapons, including those made with synthetic biology, have characteristics that make verification and enforcement technically difficult, compared to nuclear and chemical weapons (Bakerlee et al., 2020; Bressler & Bakerlee, 2018). Biological weapons require fewer resources and are relatively easy to develop and manufacture in secret, due to the multiple-use nature of materials, equipment, and techniques used.[118]

[117] According to Gronvall (2015), “There is now adequate guidance for laboratories to develop oversight systems to catch and contain accidents, but not all research institutions adhere to such guidance, require adequate training, or have sufficient resources to dedicate to biosafety. There is also great variability from one research institution to another, even within a nation.”

[118] “The knowledge, materials, and technologies needed to make and use a biological weapon are readily accessible around the world.” Gronvall (2017).

Consider that nuclear weapons require highly enriched uranium, which emits readily detectable radiation, as well as specific equipment and infrastructure that is expensive, technologically advanced, large and difficult to hide, and has few other uses. In contrast, synthetic biology has no need for large facilities and uses materials and equipment that are widely used for a variety of research projects, without a clear indicator of malicious intent. Biological materials are widely available in labs and nature, and it is increasingly possible to synthesize materials and organisms *de novo*, allowing actors to circumvent screening requirements[119] and avoid attribution (e.g., Gronvall, 2016, pp. 36–41; Gronvall et al., 2009, p. 434).

In international law, the Biological Weapons Convention lacks effective monitoring and enforcement mechanisms (Means, 2019; Scrivner, 2018) and faces financial and political challenges.[120] What legal mechanisms have been used or proposed for the monitoring and enforcement of legal instruments, for example through verification, transparency, confidence-building measures, and other measures short of verification (Lentzos, 2019)? What can be learned from existing compliance and enforcement protocols for other weapons and controlled agents (Becker et al., 2005)? What measures are most effective to prevent proliferation when considering existential risk reduction, rather than considering the ability to strictly verify compliance?

In national law, what legal mechanisms might exist, such as screening and restrictions on providing dual-use technology and materials (Garfinkel, 2007; Kobokovich et al., 2019)? What might be effective for different points of intervention (for example, equipment, labs, vendors, institutional researchers, DIY bio community)? Is there a need for stronger forms of supervision (Bostrom, 2019)? If so, would they violate civil rights and liberties? What limits should exist on monitoring and surveillance, such as to prevent abuse or avoid an attractor state or lock-in to a totalitarian state? Do specific synthetic biology applications have adequate oversight (Gronvall, 2015, p. 8)? More broadly, how can oversight mechanisms adapt as circumstances change, such as with emerging technology or changing risks? What role could soft law, such as other guidance and norms, have in a monitoring regime at an international (Cameron et al., 2020) or national level?

[119] Early proposals and guidance sought to address concerns that pathogen or toxin DNA could be manipulated or created through the use of nucleic acid synthesis technologies by requiring commercial firms to screen purchases for synthetic DNA (e.g., Garfinkel et al., 2007; U.S. Department of Health and Human Services, 2020). However, changes to gene synthesis technologies and market conditions have reduced the efficacy of these biosecurity protections (Kobokovich et al., 2019), a trend likely to continue as technology develops.

[120] A joint NGO statement in 2018 described the Convention as "in a precarious state," with financial debts from certain state parties putting its operation at risk (Center for Global Health Science and Security et al., 2018).

5.1.4 Attribution

Attribution is the ability to identify or rule out the source of a biological threat. Attribution offers three main security benefits, which can reduce existential risk: (a) informing response efforts and mitigating consequences by providing information about the motive of the actor and capabilities of the biological agent, (b) identifying responsible parties for appropriate legal recourse, and (c) deterring reckless accident and misuse, and preventing future misuse by the same actors, if perpetrators are held accountable[121] (Lewis et al., 2020). In the context of synthetic biology, attribution involves determining whether a biological agent involved has been genetically engineered and, if so, where it was engineered, by whom, and why.

Attribution of synthetic biology agents poses unique technical challenges. Biological agents may be developed and deployed in a clandestine manner. Once released, they may propagate, replicate, and mutate in unpredictable ways, making it more difficult to identify the actor or location of release.[122] Technical forensics may aid in attribution,[123] but are not as reliable or complete as for nuclear and chemical weapons. As a result, attribution of synthetic biology agents may depend on non-technical indicators (for example, location, victims, epidemiological features) and intelligence (for example, human sources, communications, surveillance and monitoring data). How can the law ensure that several sources of information are available to support or supplement technical measures (for example, legal ability to collect samples, gather intelligence)? How can attribution methods meet standards for admissibility as evidence under national law or at an international tribunal (Bidwell & Bhatt, 2016, pp. 18–20)?

Development of attribution measures may have the unintended effect of increasing certain risks. First, the possibility of being found culpable may motivate concealment of misuse or accident in a way that could create or aggravate risk

[121] Attribution is only meaningful if it leads to some form of legal recourse, as described above for accident and misuse. Attribution is of limited value if an actor intends to claim responsibility or can avoid consequences for misuse or accidental release of a biological agent.

[122] Compare to chemical and nuclear weapons, which can generally be traced (cf. footnotes 119–120 and accompanying text, discussing technical challenges in monitoring the development and production of biological weapons compared to chemical and nuclear weapons).

[123] Attribution tools for synthetic biology include, for example, advanced sequencing to rapidly characterize an agent (NASEM, 2018a, Box 8-2 and accompanying text), forensics to detect engineering and identify the engineer (Lewis et al., 2020; Scoles, 2020; see also IARPA, 2020; NASEM, 2017a), machine-learning tools to predict lab-of-origin, nation-of-origin, and ancestor lab (Alley et al., 2020), and microbial forensics (National Research Council, 2014).

(Cotton-Barratt et al., 2020, p. 274).[124] Second, tools and techniques used for attribution are dual-use, meaning they might also be used to evade attribution. How can the law minimize these risks? What role can the law play in balancing the benefits of developing attribution measures with the risks from their dual-use nature?

5.1.5 Dual-Use Concerns

"Dual-use" refers to something that can be used for beneficial purposes or to cause harm.[125] The dual-use nature of synthetic biology poses an existential risk, from misuse as well as accident, as research to advance beneficial applications may have harmful applications or present other risks. Legal instruments create prohibitions based on these dichotomies (Millett, 2017), yet much research and technology exists along some spectrum of dual use; even extremely dangerous biotechnology has a plausible argument for how it could have a defensive or peaceful use, or may itself create the need for research on a countermeasure. Notably, dual-use concerns have been raised by gain-of-function research, in which a biological entity is given a new property.[126] What types of institutions and legal mechanisms have been used to reduce existential risk from dual-use concerns throughout the research life cycle—such as prohibitions on certain types of research or involving certain materials, limiting access to materials and equipment, export controls (Kanetake, 2018), intellectual property restrictions, oversight committees at different stages (NASEM, 2018b, pp. 43–58 & Table 3-1; Resnik, 2013), and advisory boards such as the National Science Advisory Board for Biosecurity (NASEM, 2017b, pp. 31–38)? How could international instruments or institutions be strengthened or created to address dual-use concerns (Millett, 2017; NASEM, 2017b, pp. 38–44)? What can we learn from existing regulations (Lev, 2019)? What other mechanisms are conceivable (Marcello & Effy, 2018)? What limits do constitutional law and other instruments or rights place on mechanisms to control research and development (Ram, 2017; Santosuosso et al., 2007)? What role can norms, codes of ethics, and other soft law play (NASEM, 2018b, pp. 58–78 & Table 3-2)?

[124] For case studies of how this incentive can weaken prevention and response, see Chernov and Sornette (2016).

[125] Dichotomies of dual use have been conceptualized as: (a) war or peace, (b) good or evil, (c) offense or defense, (Evans & Commins, 2017), and (d) military or civilian (Mahfoud, et al., 2018). For an informal discussion of understandings of dual-use, see Weiss Evans (2018).

[126] From a scientific perspective, not all gain-of-function research is concerning, such as research to confer pest resistance to crops. However, the term "gain-of-function" often refers specifically to gain-of-function research *of concern*, in the same way that "dual-use" often refers specifically to dual-use research, technology, or materials *of concern*.

By clearly identifying what is and is not prohibited, the law could set clear expectations and support decisive action. In an international agreement, clear definitions could also reduce doubt, suspicion, and proliferation throughout other countries seeking to protect themselves, and thereby reduce overall biorisk (cf. Enemark, 2017; Joint NGO statement, 2018). However, it is also important that a dual-use framework remain adaptable to changing risk considerations. In what ways can the law create bright-line rules to identify dual-use research, materials, and technology of concern? What about bright and fuzzy lines? What kind of framework or delineations would be useful (for example, categories for what is permitted, prohibited, or permitted with special oversight or regulatory requirements)? Given that the weight of considerations may change over time as new defense- and offense-enabling technologies come into play (cf. Lewis, 2019; NASEM, 2017a), what kind of process would be appropriate to (a) assign categories and (b) update these assignments with some frequency (Dubov, 2014, p. 251; Palmer, 2020)? How would this interact with legal mechanisms for addressing information hazards? What can we learn from other fields of law?

5.1.6 Information Hazards

Biorisks arise not only from biological materials, but also from biological information; information can also be dual use. “Information hazards” are risks that arise from dissemination or potential dissemination of true information that may cause harm or enable some agent to cause harm (Bostrom, 2011b). If published, they may give ideas or implementation details to those who would misuse or carelessly use it (Crawford et al., 2019). The dual-use nature of much biological information makes it difficult to draw clear lines around what information is a hazard or what scientific research could produce hazardous information (Lewis et al., 2019). How can the law anticipate and manage potential information hazards (Lewis, 2018b; Lewis et al., 2019, pp. 979–980)? What can be learned from discussions on broader dual-use concerns or on information hazards in other fields? What legal mechanisms or areas of law have been used or are conceivable to address information hazards—such as export controls (Hindin et al., 2017; NASEM, 2017b, pp. 47–50), administrative law, security classification, or intellectual property law? How could the regulation of such information adapt to the changing risks over time? To what extent is restricting scientific knowledge consistent with applicable constitutional law (Ram, 2017)? What role should professional self-regulation, journal policies (Casadevall et al., 2013), best practices and norms, and other forms of soft law play?

5.1.7 Reducing Unintended Consequences

Emerging synthetic biology technologies could pose risks that are unknown or difficult to anticipate with specificity at the time of deployment. Thus, even intentional release of organisms could carry a risk of unintended harmful consequences.[127] While the nature of some risks may be known, there could still be uncertainty about its likelihood and specific details. How can the law reduce the risk of harmful and unintended effects stemming from synthetic biology? What kind of analysis is appropriate to assess the risks and benefits (Section 6.3.1)? Is there a need for reporting, registration, other documentation of certain types of information, required containment and response strategies, ongoing monitoring following release, or liability schemes (e.g., Warmbrod et al., 2020)? If so, how can this be done effectively?

Gene drives present a specific need to reduce unintended consequences. A gene drive is a type of genetic element that improves its own chances of inheritance in future generations. Through genetic engineering, gene drive systems can be used to suppress a population (for example, disease vectors, plant pests) or alter most of a population to express a desired trait (for example, to increase traits that correspond with well-being or survival of desired species, to increase productivity of resources that are heavily harvested). Due to the nature of gene drives, they present a greater risk of competing with native species and acting like an invasive species, leading to greater concern for potential movement across political boundaries. If multiple gene drives target the same organism (or less likely, the same sequence), there could also be unexpected and unintended interactions (Warmbrod et al., 2020, p. 20 & Appendix 2). How can the law reduce risks from environmental release and transboundary movement of organisms with gene drives (Kuzma & Rawls, 2016; Warmbrod et al., 2020), at a national and international level? What national and international laws exist and might address release of organisms with gene drives (e.g., NASEM, 2016a, Chapter 8; Rabitz, 2019)? What other biosafety, risk assessment, and regulatory measures or legal institutions could address gene drive research and reduce risk of and mitigate unintended consequences (Kofler et al., 2018; Warmbrod et al., 2020)? What factors should be considered, such as persistence and reversibility (Eckerström Liedholm, 2019), and specific technical solutions to meet them, such as a self-extinguishing daisy-drive to make untested gene drives less persistent, or ensuring reversibility with a tested reversal drive

[127] For example, (a) modified microbes could have allergenic properties, transfer antibiotic resistance into a harmful strain of bacteria, or cause a microbial strain to become pathogenic, and (b) environmental release could have unforeseen consequences on the balance of functioning ecosystems, lead to competition with native species, or result in horizontal gene transfer (i.e., to non-target organisms) (e.g., Hewett et al., 2016).

(Warmbrod et al., 2020; see also Backus & Delborne, 2019)?[128] How can the law facilitate coordination and communication between researchers and stakeholders? What legal instruments exist that already apply? What areas are unsettled?

5.1.8 Flexible and Clear Regulatory Approach

Specific language in regulation of technology can limit its applicability to that which is known now. For example, list-based approaches that create bright lines allow emerging developments to escape regulation (Carter & Friedman, 2015, pp. 8-9 and throughout). What alternatives exist to list-based approaches,[129] which might create a more flexible safety net (Casadevall & Relman, 2010; DiEuliis et al., 2017; Lewis, 2020; Lewis et al., 2019; NASEM 2018a, Chapter 8)? What can we learn from related research on flexible constitutions (Section 6.1.3)?

However, ambiguity may limit enforceability, or even sow doubt and encourage proliferation in an international context (cf. Enemark, 2017). For example, the Biological Weapons Convention describes “microbial or other biological agents, or toxins” with no “protective” purpose, providing considerable room for argument. Especially for international agreements, how can a legal instrument ensure sufficient clarity to reduce doubt and corresponding defensive proliferation, while also allowing adaptability? How might these instruments and institutions be designed to facilitate easier consensus around updating provisions or interpretations?

EXISTING ACADEMIC LITERATURE

Bakerlee, C., Guerra, S., Parthemore, C. Soghoian, D., & Swett, J. (2020). *Common misconceptions about biological weapons*. Council on Strategic Risks. https://councilonstrategicrisks.org/2020/12/07/briefer-common-misconceptions-about-biological-weapons/

Becker, U., Müller, H., & Wunderlich, C. (2005). While waiting for the protocol. *The Nonproliferation Review*, *12*(3), 541–572. https://doi.org/10.1080/10736700600601194

[128] Reversal drives allow researchers to contain the damage and manage unforeseen consequences from release of an organism with a gene drive. By default a gene drive is persistent, requiring only a single process; however, a daisy-drive is self-extinguishing (Noble et al., 2019), providing a way to reduce geographic spread and conduct more limited field trials by limiting the number of generations it can spread (Eckerström Liedholm, 2019).

[129] List-based approaches include the Select Agent Regulations set forth by the U.S. Department of Health and Human Services and U.S. Department of Agriculture, which review and republish the lists at least every other year (Centers for Disease Control, 2020), and the seven experiments of concern (National Research Council, 2004; Rapport, 2014).

Beeckman, D. S. A., & Rüdelsheim, P. (2020). Biosafety and biosecurity in containment: A regulatory overview. *Frontiers in Bioengineering and Biotechnology*, *8*(650), 1–7. https://doi.org/10.3389/fbioe.2020.00650

Bidwell, C. A., & Bhatt, K. (2016, February). *Use of attribution and forensic science in addressing biological weapon threats: A multi-faceted study*. Federation of American Scientists. https://fas.org/pub-reports/biological-weapons-and-forensic-science/

Bostrom, N. (2011b). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, *10*, 44–79. https://nickbostrom.com/information-hazards.pdf

Carter, S. R., & Friedman, R. M. (2015, October). *DNA synthesis and biosecurity: Lessons learned and options for the future*. J. Craig Venter Institute. https://www.jcvi.org/research/dna-synthesis-and-biosecurity-lessons-learned-and-options-future

Casadevall, A., & Relman, D. (2010). Microbial threat lists: obstacles in the quest for biosecurity?. *Nature Reviews Microbiology*, *8*(2), 149-54. https://doi.org/10.1038/nrmicro2299

Cotton-Barratt, O., Daniel, M., & Sandberg, A. (2020). Defence in depth against human extinction: Prevention, response, resilience, and why they all matter. *Global Policy*, *11*(3), 271–282. https://doi.org/10.1111/1758-5899.12786

Dubov, A. (2014). The concept of governance in dual-use research. *Medicine, Health Care and Philosophy*, *17*, 447–457. https://doi.org/10.1007/s11019-013-9542-9

Enemark, C. (2017). *Biosecurity dilemmas*. Washington, DC: Georgetown University Press. http://press.georgetown.edu/book/georgetown/biosecurity-dilemmas

Gronvall, G. K. (2015, February). *Mitigating the risks of synthetic biology*. Council on Foreign Relations: Center for Preventive Action. https://www.jstor.org/stable/resrep24166

Gronvall, G. K. (2016). *Synthetic biology: Safety, security, and promise*. Baltimore, MD: CreateSpace Independent Publishing Platform.

Gronvall, G. K., Bouri, N., Rambhia, K. J., Franco., C., & Watson, M. (2009). Prevention of biothreats: A look ahead. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 7(4), 433–442. https://doi.org/10.1089/bsp.2009.1112

Ilchmann, K., & Revill, J. Chemical and biological weapons in the 'New Wars'. *Science and Engineering Ethics*, *20*, 753–767 (2014). https://doi.org/10.1007/s11948-013-9479-7

Kobokovich, A., West, R., Montague, M., Inglesby, T., & Gronvall, G. K. (2019). Strengthening security for gene synthesis: Recommendations for governance. *Health Security*, *17*(6), 419–429. http://doi.org/10.1089/hs.2019.0110

Kofler, N., Collins, J. P., Kuzma, J., Marris, E., Esvelt, K., Nelson, M. P., Newhouse, A., Rothschild, L. J., Vigliotti, V. S., Semenov, M., Jacobsen, R., Dahlman, J. E., Prince, S., Caccone, A., Brown, T., Schmitz, O. J. (2018, November 2). Editing nature: Local roots of global governance. *Science*, *362*(6414), 527–529. https://doi.org/10.1126/science.aat4612

Lentzos, F. (2019). Compliance and enforcement in the biological weapons regime. United Nations Institute for Disarmament Research. https://www.unidir.org/sites/default/files/2020-02/compliance-bio-weapons.pdf

Lewis, G., Millett, P., Sandberg, A., Snyder-Beattie, A., & Gronvall, G. (2019). Information Hazards in Biotechnology. *Risk Analysis*, *39*(5), 975–981. https://doi.org/10.1111/risa.13235

Lewis, G., Jordan, J. L., Relman, D. A., Koblentz, G. D., Leung, J., Dafoe, A., Nelson, C., Epstein, G. L., Katz, R., Montague, M., Alley, E. C., Filone, C. M., Luby, S., Churche, G. M., Millett, P., Esvelt, K. M., Cameron, E. E., Inglesby, T. V. (2020). The biosecurity benefits of genetic engineering attribution. *Nature Communications*, *11*(6294). https://doi.org/10.1038/s41467-020-19149-2

Marcello, I., & Effy, V. (2018). Dual use in the 21st century: emerging risks and global governance. *Swiss Medical Weekly*, *148*(14688). https://doi.org/10.4414/smw.2018.14688

Millett. P. D. (2017, January 17). Gaps in the international governance of dual-use research of concern. In National Academies of Sciences, Engineering, and Medicine, *Dual use research of concern in the life sciences*. DC: The National Academies Press. https://doi.org/10.17226/24761 (under the Resources tab)

National Academies of Sciences, Engineering, and Medicine (2018a). *Biodefense in the age of synthetic biology*. The National Academies Press. https://doi.org/10.17226/24890

National Academies of Sciences, Engineering, and Medicine (2018b). *Governance of dual use research in the life sciences: Advancing global consensus on research oversight: proceedings of a workshop*. The National Academies Press. https://doi.org/10.17226/25154

National Academies of Sciences, Engineering, and Medicine (2017a). *A Proposed Framework for Identifying Potential Biodefense Vulnerabilities Posed by Synthetic Biology: Interim Report*. The National Academies Press. https://doi.org/10.17226/24832

National Academies of Sciences, Engineering, and Medicine (2017b). *Dual use research of concern in the life sciences: Current issues and controversies*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24761

National Academies of Sciences, Engineering, and Medicine (2016a, July 28). *Gene drives on the horizon: Advancing science, navigating uncertainty, and aligning research with public values*. The National Academies Press. https://doi.org/10.17226/23405

National Academy of Sciences and National Research Council (2012). *Biosecurity challenges of the global expansion of high-containment biological laboratories: Summary of a workshop*. The National Academies Press. https://doi.org/10.17226/13315

Nouri, A., & Chyba, C. F. (2008). Biotechnology and biosecurity. In N. Bostrom, & M. M. Ćirković (Eds.), *Global catastrophic risks*. Oxford University Press.

Palmer, M. J. (2020). Learning to deal with dual use. *Science*, *367*(6482), 1057. https://doi.org/10.1126/science.abb1466

Rabitz, F. (2019). Gene drives and the international biodiversity regime. *Review of European, Comparative and International Environmental Law*, *28*(3), 339–348. https://doi.org/10.1111/reel.12289

Ram, N. (2017). Science as speech. *Iowa Law Review*, *103*(3), 1187–1238. https://ilr.law.uiowa.edu/print/volume-102-issue-3/science-as-speech/

Resnik, D. B. (2013). Scientific control over dual-use research: prospects for self-regulation. In B. Rappert, & M. J. Selgelid (Eds.), *On the dual uses of science and ethics. Principles, practices and prospects* (pp. 237–254). Canberra: Australian National University E-Press.

Sandberg, A., & Nelson, C. (2020, June 10). Who should we fear more: Biohackers, disgruntled postdocs, or bad governments? A simple risk chain model of biorisk. *Health Security*, *18*(3), 155–163. https://doi.org/10.1089/hs.2019.0115

Santosuosso, A., Sellaroli, V., & Fabio, E. (2007). What constitutional protection for freedom of scientific research? *Journal of Medical Ethics*, *33*(6), 342–344. http://dx.doi.org/10.1136/jme.2007.020594

Schoch-Spana, M., Cicero, A., Adalja, A., Gronvall, G., Kirk Sell, T., Meyer, D., Nuzzo, J. B., Ravi, S., Shearer, M. P., Toner, E., Watson, C., Watson, M., & Inglesby, T. (2017). Global catastrophic biological risks: Toward a working definition. *Health Security*, *15*(4), 323–328. https://doi.org/10.1089/hs.2017.0038

Scrivner, S. (2018). Regulations and resolutions: Does the BWC prevent terrorists from accessing bioweapons? *Journal of Biosecurity, Biosafety, and Biodefense Law*, *9*(1), 1–5. https://doi.org/10.1515/jbbbl-2018-0006

Warmbrod, K. L., Kobokovich, A., West, R., Ray, G., Trotochaud, M., & Montague, M. (2020, May 18). *Gene drives: Pursuing opportunities, minimizing risk*. Johns Hopkins Bloomberg School of Public Health, Center for Health Security. https://www.centerforhealthsecurity.org/our-work/publications/gene-drives-pursuing-opportunities-minimizing-risk

EXISTING INFORMAL DISCUSSION

Berger, A. (2014, June 26). *Potential global catastrophic risk focus areas*. Open Philanthropy. https://www.openphilanthropy.org/blog/potential-global-catastrophic-risk-focus-areas

Bressler, D., & Bakerlee, C. (2018, December 6). *"Designer bugs": How the next pandemic might come from a lab*. Vox. https://www.vox.com/future-perfect/2018/12/6/18127430/superbugs-biotech-pathogens-biorisk-pandemic

Centre for the Study of Existential Risk. *Global catastrophic biological risks*. University of Cambridge. https:// www.cser.ac.uk/research/global-catastrophic-biological-risks

Crawford, M., Adamson, F., & Ladish, J. (2019, September 16). *Bioinfohazards* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/ixeo9swGQTbYtLhji/bioinfohazards-1

Klotz, L. (2019, February 25). *Human error in high-biocontainment labs: a likely pandemic threat*. Bulletin of the Atomic Scientists. https://thebulletin.org/2019/02/human-error-in-high-biocontainment-labs-a-likely-pandemic-threat/

Lewis, G. (2020, March). *Reducing global catastrophic biological risks*. 80,000 Hours. https://80000hours.org/problem-profiles/global-catastrophic-biological-risks/

Lewis, G. (2018b, February 19). *Horsepox synthesis: A case of the unilateralist's curse?* Bulletin of the Atom Scientists. https://thebulletin.org/2018/02/horsepox-synthesis-a-case-of-the-unilateralists-curse/

5.2 Coordination and Response

Many have recognized the need for global cooperation in order to avoid existential risk (see, e.g., Bostrom, 2013; Farquhar et al., 2017, p. 6). This holds true for biological risks, where one nation can have a global impact, and often a multilateral approach is most effective (Heyman et al., 2009). Infectious disease, organisms, and knowledge are not confined to national borders; potential pandemic pathogens can spread with increasing ease due to globalization and air travel, and organisms with gene drives may travel across political boundaries. To respond effectively, there must be a shared and cooperative approach for the detection and mitigation of threats to global health. Nations must also coordinate local response and manage sharing of information across local and national boundaries.

Research to improve coordination and response is dual purpose, as many legal and technical measures to detect and respond to anthropogenic biological risks, such as robust surveillance systems, availability of medical countermeasures, and surge capacity for healthcare systems, are also relevant to natural pandemics (NASEM, 2018a, Chapter 8). The following research projects detail mechanisms through which the law could reduce existential risk by improving global and local coordination and response.

RESEARCH PROJECTS

5.2.1 Global Cooperation

While many actors can help address global catastrophic risk and existential risk, the international community will probably need to play a major role (Cotton-Barratt et al., 2016, p. 88). Promising legal research on global cooperation and response could first survey the landscape and identify areas for change. What international legal frameworks are relevant to synthetic biology (Keiper & Atanassova, 2020; Lai et al., 2019, Table 1), and what protocols or mechanisms do they have for ongoing review and changes? How might these mechanisms be strengthened?

It would also be useful to learn how international bodies could more easily reach consensus. Future implications of synthetic biology may be difficult to predict and warrant an adaptable method of governance (Zhang, 2011), and efforts to adapt or strengthen existing instruments have faced different limitations and challenges. What meta-level process could be used to reach consensus on topics such as dual-use, information hazards, and emerging technology risks? What flexible and evolving art of governance would facilitate effective interactions among current and emerging actors, with representation by various stakeholders? What would cultivate accountability, mutual trust, and responsiveness to emerging technologies and concerns? What role could an institution or protocols within an instrument play? What can we learn from more general research on mechanisms of cooperation and world governance (Section 6.1.2)?

5.2.2 Global Pandemic Response

A primary concern is international detection and response to a potential pandemic. An epidemic, which is an outbreak of disease affecting many people within a region, if not contained can become a pandemic, which is spread over a wider geographic area (usually multiple countries or continents) and affects a high proportion of the population (Merriam-Webster). There is a need for collective preparedness, as risky governance by one nation could endanger others and lead to global catastrophic or existential risk. Given the risk of a natural or engineered pandemic,[130] it seems worthwhile to investigate the specific question of pandemic detection and response. What can we learn from how nations and global institutions have responded to epidemics and pandemics in the past (e.g., Sirleaf, 2018a)? What legal institutions or tools can help with rapid anticipation, prevention, and response to outbreaks (Farquhar et al., 2017, Section 2.2; Sirleaf, 2018b)? How could existing institutions or instruments, such as the International Health Regulations,[131] be adapted to better address this need?

5.2.3 Pandemic Finance

Preventing and managing the spread of an epidemic requires both a source of funds and effective mobilization of these funds for response. Several funding sources exist but are problematic for responding to a potential pandemic; funds may be preallocated, distributed too slowly to prevent spread, dependent on private giving, take the form of undesirable loans (Bruns, 2019), or, as in the case of the World Bank Group's Pandemic Emergency Financing Facility, discontinued (Hodgson, 2020). What institution or legal mechanism could facilitate financing pandemic response and management, ensuring that funds are available, allocated to pandemic response, and distributed effectively? What kind of trigger will ensure that money and resources are delivered in a timely manner, to catch a potential pandemic as early as possible (see Meenan, 2020; NASEM, 2016b, ch. 6)? What kind of insurance (Taylor, 2008; Cotton-Barratt, 2014), reinsurance (Anthony & Neill, 2020),

[130] See above, footnote 109 and accompanying text.

[131] The International Health Regulations (IHR) were adopted by the World Health Assembly in 1969 and last revised in 2005, aim "to prevent, protect against, control and provide a public health response to the international spread of disease in ways that are commensurate with and restricted to public health risks, and which avoid unnecessary interference with international traffic and trade" (Article 2). They require members to assess events within their respective territories and use directives set forth in the IHR, including notice of initial assessment; public health information; measures taken to respond; and ongoing information regarding studies, cases and deaths, and spread of the disease (Article 6).

financial institution, capital market instrument, or other instruments are conceivable, and how might they interact (Farquhar et al., 2019; NASEM, 2016b)? What are their advantages and disadvantages? What can we learn from their usage in other fields?

5.2.4 National Public Health Preparedness

In an ideal response to a potential pandemic or other public health emergency, a nation detects the threat early and responds appropriately. Responsiveness hinges on several factors, including coordination among government agencies, officials, and non-government actors; clear roles and responsibilities; preparedness testing; surveillance, monitoring, and reporting capabilities for early detection (for example, epidemiological methods of identifying victims, agents, and modes of transmission); countermeasures and a robust supply chain for quick response; mitigation strategies, emergency response, availability of supportive health care facilities, and effective procedures for isolation and quarantine; and legal ability to enact and enforce pharmaceutical and nonpharmaceutical interventions (see Avin et al., 2018, Figure 3, p. 5; Khan, 2018; Kun, 2014; NASEM, 2017a, p. 34; NASEM, 2020a; Nelson et al., 2007). What institutions, framework, or infrastructure would allow for a quick and effective response to a biological threat? What are the barriers? What powers are useful or necessary for oversight, monitoring, and response? Should any of these be limited for use in certain circumstances, and if so, which ones and how? To what extent are they consistent with existing law?[132] Presented as a separate research project is the question of coordination among different actors.

5.2.5 National Coordination

Nations often rely on several actors to prepare for biorisk, detect a threat early, and respond appropriately. Coordination is a key factor, as detection and response

[132] More specifically, near-term research could address specific legal mechanisms or powers, bridging the near- and long-term: What specific legal mechanisms could be used to implement public health interventions as preventive or responsive measures (for example, vaccines, mask mandates, travel restrictions for individuals who are ill or traveling from a suspect country, quarantine, or isolation mandates, air filtration requirements for businesses remaining open during a pandemic, measures to prevent spread of misinformation)? What exemptions are or would be permitted under existing laws, and what is the impact on biorisk? To what extent are biomonitoring and contact tracing (for example, metadata on hospital visitation and symptoms, broader network effects bigger than individual level of contact tracing) consistent with applicable privacy laws?

could involve federal or local agencies, other government bodies, and the private sector.[133] Government actors may have overlapping responsibilities in biodefense efforts, and inadequate planning and coordination can increase the probability of a given risk reaching catastrophic levels. What are the legal barriers to national coordination, such as lack of clear jurisdiction or responsibility (cf. Kvinta, 2011) or lack of harmonized state or local laws? How could they be overcome? How might it look for a centralized body or command structure to take force during a pandemic or other bio-threats? What would be the limits on such a body? Would this look different for different nations, and if so, how? Given existing structures of governance, what approaches could optimally increase coordination in the near- and long-term?

These questions could be addressed through a broader comparative legal analysis, to examine what legal mechanisms for responding to biorisk have been effective in different contexts, and how. What are the characteristics of governments, institutions, and mechanisms that correspond to different outcomes? Do early and effective detection and response correspond to particular decision-making processes, emergency powers, clear structures for coordination, adaptability in an existing regulatory regime, or other factors? How does it vary by the type or scope of the threat?

Existing Academic Literature

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Cotton-Barratt, O., Farquhar, S., Halstead, J., Schubert, S., & Snyder-Beattie, A. (2016). *Global catastrophic risks 2016*. Global Challenges Foundation. https://globalchallenges.org/wp-content/uploads/2019/07/Global-Catastrophic-Risk-Annual-Report-2016.pdf

Farquhar, S., Cotton-Barratt, O., & Snyder-Beattie, A. (2017). Pricing externalities to balance public risks and benefits of research. *Health Security*, *15*(4), 401–408. https://doi.org/10.1089/hs.2016.0118

Farquhar, S., Halstead, J., Cotton-Barratt, O., Schubert, S., Belfield, H., & Snyder-Beattie, A. (2017). *Existential risk: diplomacy and governance*. Global Priorities Project. https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf

[133] In the United States several iterations of biodefense strategies accompany a large biodefense budget. The 2018 National Biodefense Strategy is currently in force, preceded by the Obama administration's 2009 National Strategy for Countering Biological Threats and the 2012 National Strategy for Biosurveillance, and the George W. Bush administration's 2004 Homeland Security Presidential Directive-10, which followed the 2001 anthrax attacks. It recognizes the importance of "multi-sectoral cooperation," through engagement and cooperation across all levels of government and partnership with non-governmental organizations and the private sector (p. 4).

Heyman, D., Epstein, G. L., & Moodie, M. (2009, December). *The Global Forum on Biorisks: Toward effective management and governance of biological risks*. Center for Strategic and International Studies. https://fas.org/programs/bio/resource/documents/The%20Global%20Forum%20on%20Biorisks.pdf

Kvinta, B. (2011). Quarantine powers, biodefense, and Andrew Speaker. *Journal of Biosecurity, Biosafety and Biodefense Law*, *1*(1), 1–17. https://doi.org/10.2202/2154-3186.1002

Lai, H-E., Canavan, C., Cameron, L., Moore, S., Danchenko, M., Kuiken, T., Sekeyová, Z., & Freemont, P. S. (2019). Synthetic biology and the United Nations. *Trends in Biotechnology*, *37*(11), 1146–1151. https://doi.org/10.1016/j.tibtech.2019.05.011

Larsen, R., Boddie, C., Watson, M., Gronvall, G. K., Toner, E., Nuzzo, J., Cicero, A., & Inglesby, T. (2015, July). *Jump start: Accelerating government response to a national biological crisis*. Johns Hopkins Center for Health Security. https://www.centerforhealthsecurity.org/our-work/2015%20Jump%20Start/Jump%20Start

National Academies of Sciences, Engineering, and Medicine. (2016b). *Global health risk framework: Pandemic financing: Workshop summary*. The National Academies Press. https://doi.org/10.17226/21855

National Academies of Sciences, Engineering, and Medicine. (2018a). *Biodefense in the age of synthetic biology*. The National Academies Press. https://doi.org/10.17226/24890

Sirleaf, M. (2018a). Ebola does not fall from the sky: Structural violence & international responsibility. *Vanderbilt Journal of Transnational Law*, *51*(2), 477–554.

Sirleaf, M. (2018b). Responsibility for epidemics. *Texas Law Review*, *97*(2), 285–351. https://texaslawreview.org/responsibility-for-epidemics/

National Research Council (2004). *Biotechnology research in an age of terrorism*. The National Academies Press. https://doi.org/10.17226/10827

Taylor, P. (2008) Catastrophes and insurance. In N. Bostrom & M. M. Cirkovic (Eds.), *Global catastrophic risks* (pp. 164–183). Oxford University Press.

Zhang, J., Marris, C., & Rose, N. (2011, May). *The transnational governance of synthetic biology: Scientific uncertainty, cross-borderness and the 'art' of governance*. London: BIOS (Centre for the Study of Bioscience, Biomedicine, Biotechnology and Society). http://openaccess.city.ac.uk/16098/

EXISTING INFORMAL DISCUSSION

Anthony, G., & Neill, S. (2020 Jun. 5). *The International Underwriting Association backs proposals for "Pandemic Re"*. National Law Review. https://www.natlawreview.com/article/international-underwriting-association-backs-proposals-pandemic-re

Bruns, R. (2019). *Finance in a pandemic*. Event 201: A Global Pandemic Exercise. https://www.centerforhealthsecurity.org/event201/event201-resources/finance-fact-sheet-191009.pdf

Meenan, C. (2020 May 19). *The future of pandemic financing: Trigger design and 2020 hindsight*. Centre for Disaster Protection. https://www.disasterprotection.org/latest-news/the-future-of-pandemic-financing-trigger-design-and-2020-hindsight

5.3 Sharing the Benefits of Synthetic Biology

Similar to AI, synthetic biology could create vast potential for advancement and wealth across many industries and groups. Development could be directed and captured by a small set of actors, concentrating wealth and allocating benefits and risks to favor certain populations. If this distribution is suboptimal, humanity could permanently lose great potential (p-risk) or allow great suffering (s-risk) (Section 3.2.1). Therefore, it seems promising to investigate what legal mechanisms could be used to distribute benefits and risks, as well as how they ought to be distributed.

RESEARCH PROJECTS

5.3.1 Steering Research and Development

It may be possible and desirable to shape the direction of research and development to address near- and long-term global priorities. Synthetic biology is well-suited to address other cause areas. Climate change could be mitigated with biofuels, carbon capture, and sustainable production,[134] or global health and development aided through improved access to food,[135] clean water,[136] and healthcare.[137] Given the promise of synthetic biology, suboptimal development could represent permanent loss of great potential, constituting a p-risk. What legal tools could help steer such technological progress? How could intellectual property law, economic development law such as taxes and subsidies (cf. Posner, 2008), trade law, and other legal fields influence development of the synthetic biology market? What can we learn from other industries?

[134] Climate change issues could be mitigated by carbon capture by bioengineered plants (DeLisi, 2019), biofuels and biorefinery for alternative energy, optimizing carbon conversation or recapturing carbon in synthetic biology processes (François et al., 2020), more sustainable production methods (Le Feuvre & Scrutton, 2018), and engineering crops to withstand climate warming (Quint, et al. 2016).

[135] Access to food could be improved with increased yield, nutrition, and sustainability of crops and other agricultural products (Roell & Zurbriggen, 2020; Wurtzel et al., 2019), quality monitoring, processing, and storage (Aguilar et al., 2019; Tyagi et al., 2016).

[136] Synthetic biology has broad bioremediation applications, including microbial and plant-based solutions for cleaning up air, water, and soil pollution (Rylott & Bruce, 2020).

[137] Rooke (2013).

5.3.2 Access and Benefits-Sharing

As with other advancements, the allocation of potential benefits from synthetic biology could favor wealthier countries by default for at least two reasons: (a) firms are more likely to develop drugs and other products that will principally benefit those who can afford them, and (b) synthetic biology is complex and often capital intensive, meaning investors and workers in already-wealthy countries are more likely to capture the benefits to sellers, including intellectual property (Hollis, 2013). Some mechanisms exist for limited benefits-sharing; notably, the Convention on Biological Diversity and Nagoya Protocol on Access and Benefit-sharing aim, in part, to share benefits arising from genetic resources based on the geographic source, with varied national implementation of provider and user measures[138] (Sirakaya, 2019). However, there is no consensus on whether digital sequence information is within their scope, leading to ongoing discussion (see Ad Hoc Technical Expert Group on Synthetic Biology, 2015, para. 31; Bagley & Rai, 2013; Laird & Wynberg, 2018).[139] DIY bio, open access publishing, and "open source" biology could increase accessibility in low-income areas, but the wealthiest would still have earliest access and lesser risk from inadequate tools or expertise, such as for material storage or quality control (e.g., Foster, 2016). Other proposals and approaches for access and benefits-sharing include differential pricing, voluntary licensing models (Palfrey, 2017), compulsory licenses, payment mechanisms based on health impact (Hollis, 2013; WHO, 2013), allocation based on health access and risk factors,[140] and establishing rights and systems for accountability (Friedman & Gostin, 2015; Gostin & Friedman, 2020).

What institutions or legal instruments could equitably distribute wealth and resources produced by synthetic biology? Would their form vary geographically, at the national and international level, by nation, or by technology, and if so, how? How can they account for future development across all sectors, emergence of new technologies and resources, and means of bypassing such measures (United

[138] These provider and user measures enable enforcement of access and benefits-sharing requirements, often formalized in an agreement between the provider and user. Provider measures are established by a source country to ensure that its genetic resources are accessed based on mutually agreed-upon terms and with prior informed consent. User measures ensure that genetic resources are accessed according to these measures, for example through reporting requirements and compliance checkpoints.

[139] Several reports and decisions adopted by the Conference of the Parties to the Convention on Biological Diversity specifically discus synthetic biology, including Report of the Eleventh Meeting (2012 Dec. 5), Decision XII//24 (2014 Oct. 17), Decision XIII/17 (2016 Dec. 16), and Decision 14/19 (2018 Nov. 30).

[140] Most recently this type of framework was developed to plan for equitable vaccine allocation for COVID-19 (NASEM, 2020b, 2020c).

Nations Conference on Trade and Development, 2019 throughout & at p. 20)? What factors should be considered in distribution? How should they address changing circumstances over time? To what extent do DIY bio and open access distribute benefits optimally, weighed against the risks and distribution of risks, and what role might they play in an access and benefits-sharing regime?

5.3.3 Intellectual Property Regime

Intellectual property regimes may be important for synthetic biology, although in different ways than for AI (Section 4.3.5). In synthetic biology, the most relevant type of intellectual property is patents, with others used less frequently. Thus, patent law regimes in particular could help guide research and development toward desirable outcomes—influencing the rate of innovation, research directions, and magnitude and distribution of benefits (König et al., 2015). What intellectual property mechanisms have been used to steer innovation and public access in the past, and what were the consequences? For example, a patent law regime could permit compulsory licenses (Shore, 2020; but see Sirleaf, 2018b, p. 347, footnotes 346–347 and accompanying text), change patent eligibility for specific subject matter, tighten requirements for patentability, change exclusivity periods, or provide non-patent incentives.[141] What other mechanisms are conceivable (Douglas & Stemerding, 2014, Table 5 & pp. 14-15; Miguel Beriain, I. d., 2014)? Could human rights provide a basis for intellectual property law reform (Hale, 2018)? How might the law interact with soft governance and norms, for example around open source biology?

5.3.4 Distribution of Risks

Some risks from synthetic biology may be directed to certain populations or geographical locations, while universal risks may be readily avoided and mitigated locally by those with resources. Synthetic biology could replace the means of livelihood for people in developing countries (Kaebnick et al., 2014) or result in release of genetically engineered organisms that less wealthy countries do not have the resources to protect against (Hollis, 2013). This could have cascading effects, making it a risk factor. Clinical trials and experimental testing present varying and potentially great risks to humans and the environment, giving rise to questions of

[141] For example, the United States Orphan Drug Act of 1983 promotes development of treatments for rare diseases by offering incentives such as extended market exclusivity, reduced fees, and substantial tax credits for research and development. Others have adopted similar legislation, including Japan in 1993 and the European Union in 2000.

protection, informed consent, liability, and compensation.[142] What institutions or legal frameworks could equitably spread the distribution of risk? To what extent could and should they involve allocation of resources to protect against risk or liability and compensation schemes? Would their form vary geographically, or at the national and international level, and if so, how? What ethical criteria should research and clinical trials meet, and how can that change with circumstances? What questions should be answered in deciding whether to have human challenge trials, or other potentially great risks, during an emergency? Should requirements for informed consent (Kuiken, 2020, pp. 286–287; Sommers, 2020), compensation, and liability change during an emergency, and if so, how? Are there other great benefits or risks or extenuating circumstances that may warrant a different framework?

5.3.5 Human Enhancement and Beings Other than Humans

How should the law handle beings other than those we know today? With advancements in synthetic biology may come human enhancement beyond our limits today (Al-Rodhan, 2020; Gaspar et al., 2019; Masci, 2016), synthetic organisms with sentience, and animals that have been modified to have more human characteristics or contain human tissue, including brain tissue in the case of human-animal neurological chimeras (Crane et al., 2019; Kwisda et al., 2020; NASEM, 2020; Porsdam Mann et al., 2019). What can we learn from existing and proposed frameworks for legal personhood, citizenship, and rights and duties of humans and non-humans (Kurki, 2017)? Are these frameworks adequate for addressing potential ethical, legal, and societal issues that could arise with modified or synthetic beings (Emanuel et al., 2019, p.12–14; Wittes & Chong, 2014)? If not, what new or adapted framework could address these possibilities? What are the downstream legal and ethical implications of such a framework?

Given the vast potential of synthetic biology to positively (or negatively) shape the far future, how can the law consider animals and beings other than humans in

[142] Testing of particular concern may include (a) population testing, which presents a great burden in obtaining the informed consent of all potential participants and may not be as effective if the population is aware of being studied (DuBois, 2011; LaFreniere, 2019; Sutton, 2005) and (b) human challenge trials, in which participants are intentionally challenged with an infectious disease organism, for diseases that have high levels of morbidity and/or are poorly understood (Kolber, 2020). For a discussion of liability and compensation plans in the United States and possible alternatives, see Chapman et al., 2020 and Thomas, 2011. The World Health Organization (WHO), Expert Committee on Biological Standardization has published reports on regulatory considerations for human challenge trials (WHO, 2016; WHO, 2017), and the Working Group for Guidance on Human Challenge Studies in COVID-19 has published key criteria for ethical acceptability of such trials for COVID-19 (WHO, 2020).

distributing the benefits and risks it entails? Measures to prevent, detect, and respond to risk are attuned to humanity, while failing to address the welfare of vast numbers of animals. This oversight allows suffering and existential risks for non-human species.[143] How could legal mechanisms or proposals from other research questions in this Section be adapted to address these risks? Is an entirely separate institute or legal instrument warranted?

What can we learn from the broader discussions on non-human sentience in animal law (Section 9.2), artificial intelligence (Section 4.2.2), extraterrestrial intelligence (Section 8.2.3), sentience-sensitive institutions (Section 6.1.10), and moral circle expansion in judicial decision-making (Section 6.2.4)?

EXISTING ACADEMIC LITERATURE

Badley, M. A., & Rai, A. K. (2013). *The Nagoya Protocol and synthetic biology research: A look at the potential impacts*. Woodrow Wilson International Center for Scholars. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=5916&context=faculty_scholarship

Douglas, C. M., & Stemerding, D. (2014). Challenges for the European governance of synthetic biology for human health. *Life Sciences, Society and Policy*, *10*(6), 1–18. https://doi.org/10.1186/s40504-014-0006-7

Emanuel, P., Walper, S., DiEuliis, D., Klein, N., Petro, J. B., & Giordano, J. (2019, October). *CCDC CBC-TR-1599, Cyborg Soldier 2050: Human/Machine Fusion and the Implications for the Future of the DOD* (CCDC CBC-TR-1599). U.S. Army Combat Capabilities Development Command, Chemical Biological Center. https://community.apan.org/wg/tradoc-g2/mad-scientist/m/articles-of-interest/300458

Friedman, E. A., & Gostin, L. O. (2015). Imagining global health with justice: In defense of the right to health. *Health Care Analysis*, *23*(4), 308–329. https://doi.org/10.1007/s10728-015-0307-x

Gaspar, R., Rohde, P., & Giger, J. (2019). Unconventional settings and uses of human enhancement technologies: A non-systematic review of public and experts' views on self-enhancement and DIY biology/biohacking risks. *Human Behavior and Emerging Technologies*, *1*(4), 295–305. https://doi.org/10.1002/hbe2.175

Gostin, L. O., & Friedman, E. A. (2020). Imagining global health with justice: Transformative ideas for health and well-being while leaving no one behind. *Georgetown Law Journal*, *108*(6), 1535–1606. https://www.law.georgetown.edu/georgetown-law-journal/wp-

[143] Measures to prevent, detect, and respond to potential pandemics and other existential risks for animals may also have benefits for humanity, although it is uncertain whether their absence constitutes a risk factor. For example, if a virus is transmissible between humans and animals, the ability to detect it in animals and respond could limit its spread before it reaches humans, or it could prevent a human virus from mutating in animals and later causing a repeat outbreak among humans. For informal discussion, see McKenna, 2020, Briggs, 2020, and Calma, 2020 ("Animal health and human health are 'tightly interconnected'").

content/uploads/sites/26/2020/06/Gostin-Friedman_Imagining-Global-Health-with-Justice-Transformative-Ideas-for-Health-and-Well-Being-While-Leaving-No-One-Behind.pdf

Hale, Z. A. (2018). *Patently unfair: The tensions between human rights and intellectual property protection*. The Arkansas Journal of Social Change and Public Service. https://ualr.edu/socialchange/2018/04/04/patently-unfair/

Hollis, A. (2013). Synthetic biology: Ensuring the greatest global value. *Systems and Synthetic Biology*, *7*, 99–105. https://doi.org/10.1007/s11693-013-9115-5

Kaebnick, G. E., Gusmano, M. K., & Murray, T. H. (2014). The ethics of synthetic biology: Next steps and prior questions. *Synthetic Future*, *44*(S5), S4–S26. https://doi.org/10.1002/hast.392

König, H., Dorado-Morales, P., & Porcar, M. (2015). Responsibility and intellectual property in synthetic biology: A proposal for using Responsible Research and Innovation as a basic framework for intellectual property decisions in synthetic biology. *EMBO reports*, *16*(9), 1055–1059. https://doi.org/10.15252/embr.201541048

Kuiken, T. (2020) Biology without borders: Need for collective governance? In B. D. Trump, C. L. Cummings, J. Kuzma, & I. Linkov (Eds.), *Synthetic biology 2020: Frontiers in risk analysis and governance* (pp. 269–295). Springer, Cham. https://dx.doi.org/10.1007/978-3-030-27264-7_12

Kurki, V. (2017). Why things can hold rights: Reconceptualizing the legal person. In V. Kurki, & T. Pietrzykowski (Eds.), *Legal personhood: Animals, artificial intelligence and the unborn* (pp. 69–89). Springer, Cham. https://doi.org/10.1007/978-3-319-53462-6

Kuzma, J., & Rawls, L. (2016). Engineering in the wild: Gene drives and intergenerational equity, *Jurimetrics Journal*, *56*, 279–296. https://research.ncsu.edu/ges/files/2014/02/engineering_the_wild.authcheckdam.pdf

Kwisda, K., White, L., & Hübner, D. (2020). Ethical arguments concerning human-animal chimera research: A systematic review. *BMC Medical Ethics*, *21*. https://dx.doi.org/10.1186/s12910-020-00465-7

Laird S. A., & Wynberg R. P. (2018, January 10). *A fact finding and scoping study on digital sequence information on genetic resources in the context of the convention on biological diversity and Nagoya Protocol* (CBD/DSI/AHTEG/2018/1/3). Convention on Biological Diversity, Ad Hoc Technical Expert Group on Synthetic Biology. https://www.cbd.int/doc/c/e95a/4ddd/4baea2ec772be28edcd10358/dsi-ahteg-2018-01-03-en.pdf

Miguel Beriain, I. d. (2014). Synthetic biology and IP rights: Looking for an adequate balance between private ownership and public interest. In J. Boldt (Ed.), *Synthetic biology: Metaphors, worldviews, ethics, and law*. Springer VS. http://doi.org/10.1007/978-3-658-10988-2

Nielsen, M. E. J., Kongsholm, N. C. H., & Schovsbo, J. (2019). Property and human genetic information. *Journal of Community Genetics*, *10*, 95–107. https://doi.org/10.1007/s12687-018-0366-4

Palfrey, Q. A. (2017). Expanding access to medicines and promoting innovation: A practical approach. Georgetown Journal on Poverty Law and Policy, 24(2), 161–203.

Porsdam Mann, S., Sun, R., & Hermerén, G. (2019) A framework for the ethical assessment of chimeric animal research involving human neural tissue. *BMC Medical Ethics*, *20*. https://doi.org/10.1186/s12910-019-0345-2

Posner, R. A. (2008) Public policy towards catastrophe. In N. Bostrom & M. M. Cirkovic (Eds.), *Global catastrophic risks* (pp. 184–201). Oxford University Press.

Sirakaya, A. (2019). Balanced options for access and benefit-sharing: Stakeholder insights on provider country legislation. *Frontiers in Plant Science*, *10*, 1175. https://doi.org/10.3389/fpls.2019.01175

Sirleaf, M. (2018b). Responsibility for epidemics. *Texas Law Review*, *97*(2), 285–351. https://texaslawreview.org/responsibility-for-epidemics/

Sommers, R. (2020). Commonsense consent. Yale Law Journal, 129(8), 2232–2324. https://www.yalelawjournal.org/article/commonsense-consent

United Nations Conference on Trade and Development (2019). *Synthetic biology and its potential implications for biotrade and access and benefit-sharing* (UNCTAD/DITC/TED/INF/2019/12). https://unctad.org/system/files/official-document/ditctedinf2019d12_en.pdf

Wittes, B., & Chong, J. (2014, September). *Our cyborg future: Law and policy implications*. Brookings Institution. https://www.brookings.edu/research/our-cyborg-future-law-and-policy-implications/

World Health Organization, World Intellectual Property Organization, & World Trade Organization (2013). *Promoting access to medical technologies and innovation: Intersections between public health, intellectual property and trade*. https://www.who.int/phi/promoting_access_medical_innovation/en/

EXISTING INFORMAL DISCUSSION

Al-Rodhan, N. (2020, June 29). *A neurophilosophy of two technological game-changers: Synthetic biology & superintelligence*. Blog of the American Philosophical Association. https://blog.apaonline.org/2020/06/29/a-neurophilosophy-of-two-technological-game-changers-synthetic-biology-superintelligence/

Cotton-Barratt, C. (2014, October 1). *Effective policy? Requiring liability insurance for dual-use research* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/zvRerivrWdZ5J5rD9/effective-policy-requiring-liability-insurance-for-dual-use

Masci, D. (2016, July 26). *Human enhancement: The scientific and ethical dimensions of striving for perfection*. Pew Research Center. https://www.pewresearch.org/science/2016/07/26/human-enhancement-the-scientific-and-ethical-dimensions-of-striving-for-perfection/

National Academies of Sciences, Engineering, and Medicine (2020). *Ethical, legal, and regulatory issues associated with neural chimeras and organoids*. https://www.nationalacademies.org/our-work/ethical-legal-and-regulatory-issues-associated-with-neural-chimeras-and-organoids

6 Institutional Design

While previous Sections on artificial intelligence and synthetic biology focused on how the law might effectively address issues within specific cause areas, the following Section focuses on how the law might be utilized and improved to positively affect the long-term future more generally. This approach allows us to account for methodological and other uncertainties related to our existing cause areas (see Section 3), such as by addressing issues that might be more tractable from a legal standpoint. Furthermore, it allows us to tackle risks that are yet unknown but might arise in the future. The Section is divided into three parts and focuses on the design of legal institutions (6.1), judicial decision-making (6.2), and the impact, evaluation, and uncertainty in law (6.3).

6.1 Design of Legal Institutions

Perhaps the most obvious method of improving law is designing or changing written laws (so-called "law on the books") and legal institutions themselves. What are the most effective institutional design mechanisms for positively influencing the long-term future?

Research Projects

6.1.1 Protection of Intergenerational Global Public Goods

Because protecting humanity from existential threats would benefit all of humanity (non-excludable), and the protection of one does not come at the expense of that of any other individual (non-rivalrous), protection from existential risk is a global public good. Additionally, the beneficiaries are not merely global, but intergenerational—all the people who would ever live. Protection from existential risk is therefore an *intergenerational global public good* (Ord, 2020). This means that we can expect existential risk to be neglected by nation states and markets and, hence, properly avoiding existential threats is likely to necessitate strong cooperative efforts among transnational legal actors (Bostrom, 2002; Bostrom, 2013; Farquhar et al., 2017, p. 6). Beyond this, it requires intergenerational coordination.[144] How can we ensure that such efforts are effectively implemented and adhered to, and not easily flouted? Would it be more effective to approach the issue in an incremental way, or do we need truly transformative institutions designed to protect intergenerational global public goods?

6.1.2 Mechanisms of Global Cooperation & World Governance

Truly transformative ways to protect humanity from existential threats involve major changes to existing international bodies like the World Health Organization, the Biological Weapons Convention and the UN Security Council, or an overhaul, with entirely new institutions for governing and ensuring the safety of humanity fomented instead. While such major shifts in international relations are unlikely, for the time being, this may change in the future, for instance, as a result of "warning shots" such as the ongoing COVID-19 pandemic,[145] or a reordering of World Affairs after a major global catastrophe, such as a great power conflict (cf. Ord, 2020). What kind of institutions would most effectively decrease x-, s-, and p-risks, and how can humanity achieve the establishment of such institutions? How can we ensure that such powerful institutions do not add to the risk of a global authoritarian lock-in?[146]

6.1.3 Flexible Constitutions in a Vulnerable World

Scientific and technological progress might destabilize civilization. Novel military technologies, such as lethal autonomous weapons systems, bioweapons, and (perhaps most importantly) the development of yet unknown technologies, may pose risks which put the world in a "vulnerable position" (Bostrom, 2019). Constitutions need to be designed in a way that increases the ability of nation states and supranational and international organizations to mitigate such risks. Given the uncertainty regarding the characteristics of future technologies, should constitutions become more flexible, and if so, how might this be achieved? How can we measure the flexibility of constitutional provisions? How might potentially harmful

[144] One example of intergenerational cooperation may be the establishment of a "World Climate Bank" (see Broome & Foley, 2016).

[145] Having said this, one should not necessarily rely on the effectiveness of such warning shots (see Section 3.2.2).

[146] Case in point, while "world government" may in theory be best equipped to reduce existential threats (cf. Einstein, 1948), the risk of a totalitarian regime backed up by such a powerful institution, combined with increasing technological capabilities, would in practice increase rather than reduce the total amount of risk. See also Ord (2020) with further references.

constitutional provisions be flexibly interpreted and/or amended? How can future constitutions be designed to maximize flexibility, if that is desired?

6.1.4 (Constitutional) Mechanisms to Protect Future Generations

Legal institutions have been and continue to be very short-term oriented, with policy making geared towards solving contemporary issues and the democratic process reserved exclusively for the current generation (see Ace Project, 2020). What legal mechanisms are available to better protect future generations and prevent existential threats, and how might they be implemented?[147] Given that rigid constitutional provisions can create strong lock-in effects,[148] aligning constitutions with longtermist values may be of particularly high importance. What are some of the most effective constitutional provisions that aim at the protection of future generations? And how can future constitutions be designed to be more favorable to the interests of future generations?

6.1.5 Preventing a Permanent and Global (Digital) Authoritarian Lock-in

Legal policy and institutions are often resistant to change and have become more stable over time,[149] which creates the tendency to produce strong negative "lock-in" effects if such policies turn harmful. At the same time, effectively preventing existential risks, such as those arising from pandemics or modern weaponry, may lead or even require (Bostrom, 2019) governments to adopt strong surveillance systems and/or the establishment of entirely new and very powerful institutions. Such a development would increase the risk of digital authoritarianism—the use of technology by authoritarian regimes to surveil, repress, and manipulate domestic or foreign populations (Polyaka & Meserole, 2019). What are the potential risks of such authoritarian policies becoming "locked in" on a global scale (Caplan, 2011), and how might they be avoided?[150] What is the role of judicial independence, and

[147] For an overview of present mechanisms to protect future generations, see John (2020a); see also John and MacAskill (2020) and Gonzalez-Ricoy and Axel Gosseries (2016).

[148] But see also Elkins et al. (2007), which found that (a) the average lifespan of a constitution is just 17 years, and (b) the probability of a constitution lasting at least 50 years is just 19%.

[149] Rigid constitutional provisions (cf. Gosseries & Meyer, 2009), the role of precedent (Gerhardt, 1991), and path dependent effects in law (Hathaway, 2003; Lindquist & Cross, 2008), to name a few examples, all contribute to greater stability over time. See also Crootof (2019) for further discussion on path dependence and lock-in. See Section 7 on meta-research questions for long-term effects of law more generally.

[150] Caplan (2011, p. 516) assigns the risk of a world totalitarian government emerging during the next 1000 years and lasting for at least 1000 years a 5% probability. Given

liberal democratic constitutionalism in this regard (Winter, 2021a)? Can specific constitutional provisions, such as the eternity clause adopted by the German Basic Law which does not allow any changes with regard to its core principles including those of human dignity, democracy and the rule of law (German Basic Law Article 79 III) serve as effective safeguards against authoritarianism in the digital age? What are the accompanying risks of such constitutional value lock-ins?

6.1.6 Encouraging a Morally Exploratory and Reflective Society

Given our uncertainty about what the ideal future of humanity might look like, different researchers have suggested that we need a "period of long reflection" before values are locked in (Greaves et al., 2020; Lewis, 2018a; MacAskill, 2018b; Ord, 2020). During this period—which could perhaps last for tens of thousands of years—human civilization would dedicate itself to working out what is ultimately of value (MacAskill 2018b). Assuming this period is desirable, how might we design the legal system to increase the odds of a sustainable long reflection? Do we need to improve freedom of speech laws to ensure the continuous exploration of controversial ideas, and if so, under what circumstances?[151] What is the role of liberal democratic principles in ensuring such a long-lasting debate? If we consider that such a long lasting process of reflection will be impacted by (artificially) improved cognitive ability, how might this be regulated? What role might freedom of thought play (see Bublitz, 2014; Bublitz & Merkel, 2014; Bublitz, 2015; McCarthy-Jones, 2019)?[152] What (other) mechanisms are available to (a) ensure that the period will take place and (b) increase the likelihood of a maximally beneficial outcome of said period?

6.1.7 Increasing Budgets for X-Risk Prevention

While it is difficult to precisely measure global spending on existential risk, Ord (2020) points out that we can state with confidence that humanity spends more on

that authoritarianism has historically been more durable than totalitarianism (Caplan, 2011, p. 507), it seems reasonable to assume that the probability of a global authoritarian regime emerging in and lasting for the above-mentioned period of time is significantly higher than 5%.

[151] While one should be aware that "improve" can also imply to restrict freedom of speech under specific circumstances when it conflicts with other fundamental rights, in most jurisdictions, "improve" would very likely imply a "strengthening" of such laws in order to allow for a reflective and exploratory discourse.

[152] Given the neglectedness of "freedom of thought" on the one hand, and the wide-reaching and intense discussion surrounding "freedom of speech" on the other hand, one might prioritize the former.

ice cream every year than on ensuring that the technologies we develop do not destroy us. Dealing with existential risks is likely to necessitate a degree of sustained government spending (Bostrom, 2013; Farquhar et al., 2017, p. 6). Given that the UN Office of Disaster Risk Reduction reports that "an investment of $6 billion annually in disaster risk management would result in avoided losses of $360 billion over the next 15 years," it seems fair to assume that the prevention of catastrophic and existential risks is heavily underinvested.[153] What are the possibilities of introducing mandatory minimum budgets for the reduction of existential risks? Should one, all things considered, primarily aim at national legislation or international agreements in this regard? What are possible alternative mechanisms in case implementing mandatory budgets is infeasible?

6.1.8 Patient vs Urgent Legal Longtermism

Longtermist views often differ with respect to the urgency they ascribe to attempting to directly influence the long-term future. For example, whereas urgent longtermism (e.g., Ord, 2020) holds that we should attempt to address long-term threats right now, patient longtermism (cf. Todd, 2020a) suggests that we should focus on preparing ourselves to be ready for long-term threats when they become more threatening. How might the ideal legal system look under varying accounts of longtermism, and how might we account for the uncertainty with regard to the urgency of long-term threats in designing our legal institutions? How can legal systems invest or build capacity so as to effectively prepare for future risks? How might we identify nearsightedness, course setting, self-improvement, growth, and changing opportunities that influence the timing of direct work (MacAskill, 2019b; Ord, 2014) in legal systems?

6.1.9 Criminal Laws Against Increasing Existential Risk

Every year as we invent new technologies, we may have a chance of stumbling across something that offers the destructive power of the atomic bomb or a deadly pandemic, but which turns out to be easy to produce from everyday materials (Bostrom, 2019; Ord, 2020). In addition to necessitating coordination and cooperation among the World's most powerful nations, such unforeseen risks may also require the use of the intervention-intensive means of criminal law. To what extent may current (national) criminal laws already apply in such context? Do we need

[153] See also John (2020b).

new national and/or international criminal laws against the endangerment of humanity?[154]

6.1.10 Building Sentience-Sensitive Institutions

Given that s-risks and p-opportunities both depend on sentience (i.e., the capacity to have positive and negative experiences, usually thought of as happiness and suffering), and it is likely that a number of p-opportunities and s-risks are still unknown, it seems reasonable to aim at designing institutions in a sentient-sensitive manner more generally.[155] This may be especially important, given that the difference between the two extremes of p-opportunities and s-risks is very large once we consider the potential number of sentient beings whose experiences are already at stake, in particular non-human animals,[156] and those whose experiences may be at stake in the future, including artificial sentience (Tomasik, 2017). What are the most effective ways to protect sentience and design institutions accordingly? Is a "Universal Declaration of Sentient Rights" feasible, and what would it look like (see Woodhouse, 2019)? What mechanisms are available to represent non-participatory stakeholders (see Kurki & Pietrzykowski, 2017)? Is the traditional legal bifurcation between "persons" and "things" capable of protecting all sentient beings (Kurki & Pietrzykowski, 2017)? How might institutions resolve tradeoffs between very different kinds of interests on behalf of very different kinds of sentient beings (Stawasz, 2020)? How should legal institutions deal with uncertainty regarding what constitutes consciousness (Bourget & Chalmers, 2013), and what entities can be considered as sentient (cf. Sebo, 2018)? What can we learn from the

[154] Farquhar et al. (2017) remain cautious with regards to prioritizing efforts to include existential risk negligence as a crime against humanity within the Rome Statute. But see also Torres (2020a), McKinnon (2017), and Binder (2018).

[155] Philosophers and legal theorists of different schools of thought also often hold that sentience is sufficient for the possession of interests (e.g., Bentham, 1823; Feinberg, 1974; Korsgaard, 2018; Regan, 2001; Singer, 1975). Views emphasizing the impartial advancement of interests are therefore equally eager to enshrine within legal insitutitions protections for the interests of all sentient beings.

[156] For instance, just considering currently existing animals, Tomasik (2019b) estimates that there are upwards of 10^{14} land vertebrates, and at least 10^{13} marine vertebrates. In contrast, there are less than 10^{10} humans. Even ignoring invertebrates (of whom there are at least 10^{18}), and further supposing that only a fraction of these non-human vertebrates are sentient, it seems likely that nonhumans comprise the vast majority of plausibly sentient beings in existence.

field of animal law, where definitions and attributions of sentience have occasionally been incorporated within laws?[157]

6.1.11 Artificial Intelligence and the Executive

While recent work (e.g., Winter, 2021a) has explored the potential costs, benefits, and other considerations associated with artificially intelligent decision makers in the judiciary, similar questions remain open with regard to the executive (for example, in the context of federal agencies). What is the potential role of artificially intelligent systems, if any, in designing or promulgating regulations to protect the interests of future generations? What are the risks of using AI in a decision-making context in the executive branch of government, and how do those compare to those in the judiciary?

EXISTING ACADEMIC LITERATURE

Allen, G., & Chan, T. (2017, July). *Artificial intelligence and national security*. Belfer Center for Science and International Affairs, Harvard Kennedy School. https://www.belfercenter.org/publication/artificial-intelligence-and-national-security

Aschenbrenner, L. (2020). *Existential risk and growth* [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/Leopold-Aschenbrenner_Existential-risk-and-growth_.pdf

Blattner, C. (2019). The recognition of animal sentience by the law. *Journal of Animal Ethics*, *9*(2), 121–136. https://doi.org/10.5406/janimalethics.9.2.0121

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Broome, J., & Foley, D. (2016). A world climate bank. In A. Gosseries & I. González-Ricoy (Eds.), *Institutions for Future Generations* (pp. 156–169). Oxford University Press.

Bublitz, J. C. (2014). Freedom of thought in the age of neuroscience. *Archiv für Rechts-und Sozialphilosphie*, *100*(1), 1–25. https://www.researchgate.net/publication/261950057_Freedom_of_Thought_in_the_Age_of_Neuroscience

Bublitz, J. C., & Merkel, R. (2014). Crimes against minds: on mental manipulations, harms and a human right to mental self-determination. *Criminal Law and Philosophy*, *8*(1), 51–77. https://doi.org/10.1007/s11572-012-9172-y

Bublitz, C. (2015). Cognitive Liberty and the International Right to Freedom of Thought. In J. Clausen & N. Levy (Eds.), *Springer Handbook of Neuroethics*. Springer. https://doi.org/10.1007/978-94-007-4707-4_166

Cochrane, A. (2012). From human rights to sentient rights. *Critical Review of International Social and Political Philosophy*, *16*(5), 655–675.

[157] See, e.g., Art. 13 TFEU; Chessman, 2018; Reddy, 2020; World Animal Protection, 2014; see also Section 9.

Cochrane, A. (2018). *Sentientist politics: a theory of global inter-species justice*. Oxford University Press.Cotton-Barratt, O. (2015). *Allocating risk mitigation across time* [Technical report]. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/Allocating-risk-mitigation.pdf

Elkins, Z., Ginsburg, T., & Melton, J. (2007, July). The lifespan of written constitutions. In *American Political Science Association Meeting, Chicago*. https://www.researchgate.net/publication/228813917_The_Lifespan_of_Written_Constitutions

Farquhar, S., Halstead, J., Cotton-Barratt, O., Schubert, S., Belfield, H., & Snyder-Beattie, A. (2017). *Existential risk: diplomacy and governance*. Global Priorities Project. https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf

Greaves, H., MacAskill, W., O'Keeffe-O'Donovan, R., & Trammell, P. (2019). *Research Agenda for the Global Priorities Institute*. Global Priorities Institute. https://globalprioritiesinstitute.org/wp-content/uploads/GPI-research-agenda-version-2.1.pdf

Gunkel, D. J. (2018). *Robot Rights*. MIT Press.

Hilton, S. & Baylon, C. (2020). *Risk management in the UK: What can we learn from COVID-19 and are we prepared for the next disaster?*. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/resources/risk-management-uk

John, T. M., & MacAskill, W. (2021). *Longtermist institutional reform* [Forthcoming]. In N. Cargill & T. M. John (Eds.), *The Long View*. https://philpapers.org/rec/JOHLIR

Jones, N., O'Brien, M., & Ryan, T. (2018). Representation of future generations in United Kingdom policy-making. *Futures*, *102*, 153–163. https://doi.org/10.1016/j.futures.2018.01.007

MacAskill, W. (2020b). *Are we living at the hinge of history?* [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/william-macaskill-are-we-living-at-the-hinge-of-history/

McKinnon, C. (2017). *Endangering humanity: an international crime? Canadian Journal of Philosophy*, *47*(2-3), 395–415. https://doi.org/10.1080/00455091.2017.1280381

Muñiz-Fraticelli, V. (2009). The problem of a perpetual constitution. In A. Gosseries & L. Meyer (Eds.), *Intergenerational Justice*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199282951.003.0015

Ord, T. (2020). *The precipice: existential risk and the future of humanity*. Hachette Books.

Torres, P. (2020a). *International criminal law and the future of humanity: A theory of the crime of omnicide* [Unpublished manuscript]. https://www.xriskology.com/publications

Torres, P. (2020b). *Crimes without a name: On global governance and existential risks* [Unpublished manuscript]. https://www.xriskology.com/publications

Trammell, P. (2020). *Patience and philanthropy*. Global Priorities Institute, University of Oxford. https://philiptrammell.com/static/PatienceAndPhilanthropy.pdf

Weitzdörfer, J. & Beard, S. (2019). Law and policy responses to disaster-induced financial distress. In A. Kamesaka & F. Waldenberger (Eds.), *Governance, Risk and Financial Impact of Mega Disasters: Lessons from Japan*. https://www.cser.ac.uk/resources/law-and-policy-responses-financial-distress/

Wilson, G. (2013). Minimizing global catastrophic and existential risks from emerging technologies through international law. *Virginia Environmental Law Journal*, *31*(2), 307–364. https://www.jstor.org/stable/44679544

EXISTING INFORMAL DISCUSSION

Danaher, J. (2018, September 15). *The robot rights debate (index)* [Blog post.] Philosophical Disquisitions. https://philosophicaldisquisitions.blogspot.com/2018/09/the-robot-rights-debate-index.html

Todd, B. (2020a, August). *The emerging school of patient longtermism.* 80,000 Hours. https://80000hours.org/2020/08/the-emerging-school-of-patient-longtermism/

Tomasik, B. (2017). *Machine sentience and robot rights* [Blog post]. Essays on Reducing Suffering. https://reducing-suffering.org/machine-sentience-and-robot-rights/

Wiblin, R. & Harris, K. (2018b). *Our descendants will probably see us as moral monsters. What should we do about that?* [Podcast]. 80,000 Hours. https://80000hours.org/podcast/episodes/will-macaskill-moral-philosophy/

6.2 Judicial Decision-Making

Just as important as the written laws and legal institutions themselves is their application, typically (though perhaps not eternally) by a human decision maker. This Section focuses on one avenue of institutional decision-making: judicial decision-making. Although other forms of institutional decision-making (for example, executive or legislative) are also relevant from a longtermist perspective (see Section 7), legal research seems particularly well-suited to address judicial decision-making as opposed to other disciplines due to its specialization. Aside from that, judicial decision-making is a less explored field in comparison with other areas of institutional decision-making, which typically focus on the executive branch (Stauffer, 2019; Whittlestone, 2017a). Below, we list research projects for improving judicial decision-making to positively shape the far future.

RESEARCH PROJECTS

6.2.1 Cognitive Biases and Non-Quantitative Legal Standards

Much of law contains vague, undefined legal standards, particularly in the context of judicial decision-making. Some of these standards are ubiquitous, recur regularly and have particularly high stakes, such as "beyond a reasonable doubt,"[158] "probable cause,"[159] and the "proportionality test" in the European Union[160] and "balancing tests" in the United States Supreme Court.[161] What sorts of cognitive

[158] See *Patterson v. New York*, 432 U.S. 197 (1977).

[159] See U.S. Const. amend. IV, in the United States.

[160] See Case C-55/94, *Gebhard v. Consiglio dell'Ordine degli Avvocati e Procuratori di Milano* (1995) E.C.R. I-4165, para. 57.

[161] See *Wilkinson v. Austin*, 544 U.S. 74 (2005).

biases manifest themselves when applying these standards, particularly with regard to existential threats and issues of the long-term future when balancing conflicting rights, and what can be done to address/counteract them? What does the existence of cognitive biases suggest for if/when non-quantitative legal standards ought to be used in place of more rigid legal rules (cf. Kaplow, 1992)?

6.2.2 Judicial Innumeracy

In the United States, judges are often tasked with making and evaluating quantitative judgments.[162] In the European Union, judges are rarely tasked with such analyses and, hence, limit their engagement, even if it would be relevant for the case in question and the stakes are extremely high (Stucki & Winter, 2019; Winter 2020a). At the same time, many longtermist issues involve a sophisticated understanding of probability and decision theory. Since judges often receive no formal training in quantitative subjects (see discussion in Section 1) and are likely prone to statistical biases (see, e.g., Alexander & Weinberg, 2014; Gilovich et al., 2002; Kahneman & Tversky, 1982), including scope insensitivity (Baron & Greene, 1996; Greene & Baron, 2001), they may be ill-equipped to properly make such judgments. How can we address and/or mitigate judicial innumeracy so as to improve decision-making and positively shape the long-term future? How might both the severity of this issue and potential solutions vary based on how judges are selected (for example, elected, appointed, or by exam)?

6.2.3 Legal Mechanisms to Increase Evidence-Based Judicial Decision-Making

Proper legal decision-making often requires careful consideration of relevant facts. Many interventions have aimed to improve evidence-based judicial decision-making such as the use of curriculums (National Center for State Courts, 2018), checklists (Guthrie et al., 2007), and various legal reforms (Casey et al., 2013; Guthrie, Rachlinski, & Wistrich, 2001).[163] Which of these interventions are among the most effective to ensure that institutional actors are informed of and rely on evidence as opposed to external pressures or competing incentives when making frequent

[162] See *Gill v. Whitford*, 585 U.S. (2018) and Stephanopoulos and McGhee (2014) for the use of the "efficiency gap" to determine partisan gerrymandering. See *Elkins v. United States*, 364 U.S. 206 (1960) and Enos et al. (2017) for discussion on the "Negative Effect Fallacy." For more general discussion of the use of quantitative social science evidence, see Faigman (1989), Ryan (2003), and Mody (2002). For informal discussion, see Roeder (2017) and Fowler (2017).

[163] For greater discussion of Evidence-based Practices (EBP) in judicial decision-making, see Center for Effective Public Policy (2017) and U.S. Department of Justice, National Institute of Corrections (2013).

and/or high-stakes decisions? How can these interventions be aligned with issues relevant to positively shaping the long-term future?[164]

6.2.4 Moral Circle Expansion in Judicial Decision-Making

The interests of populations other than extant humans, such as non-human animals (see Section 9),[165] future generations, and artificial sentience (see Section 4.2.2), are reliably neglected in judicial decision-making (Winter, 2021b). The reasons for this are often psychological (Winter, 2021b), relating to speciesism (Caviola et al., 2019), cognitive biases (Stucki & Winter, 2019; Yudkowsky, 2008b), and short-term thinking (John & MacAskill, 2021). One solution is to expand the moral circle (Singer, 2011) of judicial decision-making—the "judicial moral circle"—by considering the interests of future generations, animals, and other sentient beings. What are the most effective ways to expand the judicial moral circle to include all sentient beings for the long-term future?

6.2.5 Accounting for Uncertainty in Judicial Decision-Making Interventions

Both empirical and normative uncertainties arise in discussions on improving judicial decision-making. For example, uncertainties pertaining to the timing and content of interventions, the underlying psychological biases (Guthrie et al., 2001), and the causes to focus on contribute to the limited knowledge of this research area.[166] How should this uncertainty be optimally accounted for in decision-making interventions? Should we favor interventions with lower evidence bases, all else equal (Askell, 2019)? Should we favor general or cause-specific interventions (Stauffer, 2019)?

6.2.6 Long-Term Challenges of Implementing AI into Judicial Decision-Making

The use of AI to assist and even replace judicial decision-making has occurred in many different capacities across many jurisdictions.[167] Some challenges of this

[164] See, for example, Jones et al. (2018) for discussion of representing future generations in policy-making in a number of countries including Finland, Hungary, Singapore, Israel, Scotland, Wales, and England.

[165] Cf. litigation projects by the Non-human Rights Project.

[166] For discussion of some of the uncertainties regarding improving institutional decisions-making in general, see Meichenbaum, 2009, Stauffer, 2019, and Whittlestone, 2017a.

[167] AI may *assist* judicial decision-making, for example with risk assessment algorithms (Coglianese & Ben Dor, 2019) including COMPAS in bail decisions (Northpointe, 2015). AI may also *replace* judicial decision-making, as with AI judges in China (Pillai, 2019),

transition to automated decision-making in the judiciary have already been explored, namely by identifying limitations of using AI to replace humans in this context and various threats to judicial legitimacy.[168] However, research thus far has neglected long-term challenges. What kinds of risks or challenges might emerge as AI increases in capability and assumes a more influential role in the judiciary? Is judicial decision-making entirely computable, and is artificial general intelligence needed to create an advanced artificial judicial intelligence (Moses, 2020; Winter, 2021a)? How might legal values lock-in to AI in this setting from both technological and institutional factors (Crootof, 2019)? How might AI influence structural features of the legal system, such as the separation of powers and judicial independence (Michaels, 2020; Winter, 2021a)?

EXISTING ACADEMIC LITERATURE

Bennett Moses, L. (2020). Not a single singularity. In S. Deakin and C. Markou (Eds.), *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence* (pp. 20–36). Hart Publishing. https://www.bloomsburyprofessional.com/uk/is-law-computable-9781509937066/

Chen, D. L., Moskowitz, T. J., & Shue, K. (2016). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, *131*(3), 1181–1242.

Epstein, L., Landes, W. M., & Posner, R. A. (2013). *The behavior of federal judges: A theoretical and empirical study of rational choice*. Harvard University Press.

Faigman, D. L. (1989). To have and have not: Assessing the value of social science to the law as science and policy. *Emory Law Journal*, *38*, 1005–1095. https://repository.uchastings.edu/faculty_scholarship/140/

Gatowski, S. I., Dobbin, S. A., Richardson, J. T., Ginsburg, G. P., Merlino, M. L., & Dahir, V. (2001). Asking the gatekeepers: A national survey of judges on judging expert evidence in a post-Daubert world. *Law and Human Behavior*, *25*(5), 433–458. https://doi.org/10.1023/A:1012899030937

Hans, V. P., Rachlinski, J. J., & Owens, E. G. (2011). Editors' Introduction to Judgment by the Numbers: Converting Qualitative to Quantitative Judgments in Law. *Journal of Empirical Legal Studies*, *8*, 1–5. https://doi.org/10.1111/j.1740-1461.2011.01222.x

Kaplow, L. (1992). Rules versus standards: An economic analysis. *Duke Law Journal*, *42*, 557–629. https://scholarship.law.duke.edu/dlj/vol42/iss3/2/

Langer, L., Tripney, J., & Gough, D. A. (2016). *The science of using science: Researching the use of research evidence in decision-making*. London: EPPI-Centre, Social Science

AI judges in Estonia (Niiler, 2019), and online dispute resolution (Carneiro et al., 2014).

[168] See Bennet Moses, 2020; Huq, 2020; Pasquale, 2019. For extensive discussion, see Deakin and Markou, 2020.

Research Unit, UCL Institute of Education, University College London. https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3504

Michaels, A. C. (2020). Artificial intelligence, legal change, and separation of powers. *University of Cincinnati Law Review*, 1083–1103. https://scholarship.law.uc.edu/cgi/viewcontent.cgi?article=1366&context=uclr

Mody, S. (2002). Brown Footnote Eleven in Historical Context: Social Science and the Supreme Court's Quest for Legitimacy. *Stanford Law Review, 54*(4), 793–829. https://doi.org/10.2307/1229579

Ryan, J. E. (2002). The limited influence of social science evidence in modern *desegregation* cases. *North Carolina Law Review*, *81*, 1659–1702. https://scholarship.law.unc.edu/nclr/vol81/iss4/8/

Stephanopoulos, N. O., & McGhee, E. M. (2014). Partisan gerrymandering and the efficiency gap. *University of Chicago Law Review*, *82*, 831–900.

Volokh, E. (2018). Chief Justice Robots. *Duke Law Journal, 68*, 1135–1192. https://dlj.law.duke.edu/article/chief-justice-robots-volokh-vol68-iss6/

Winter, C. K. (2020a). The value of behavioral economics for EU judicial decision-making. *German Law Journal*, *21*(2), 240–264. https://doi.org/10.1017/glj.2020.3

Winter, C. K. (2021a). Exploring the challenges of artificial judicial decision-making for liberal democracy. In P. Bystranowski, P. Janik, & M. Próchnicki (Eds.), *Judicial decision-making: Integrating empirical and theoretical perspectives*. https://www.christophwinter.net/s/AI-Judiciary.pdf

EXISTING INFORMAL DISCUSSION

Roeder, O. (2017, October 17). *The supreme court is allergic to math*. FiveThirtyEight. https://fivethirtyeight.com/features/the-supreme-court-is-allergic-to-math/

Stucki, S. & Winter, C. K. (2019, June). *Of Chicks and Men: Anmerkungen zum BVerwG-Urteil über die Tötung männlicher Küken*. Verfassungsblog. https://verfassungsblog.de/of-chicks-and-men/

6.3. Impact, Evaluation, and Uncertainty in Law

Jurisdictions vary greatly with respect to their legal institutions, institutional decision makers and individual legal policies. What mechanisms are available to evaluate these institutional features with regard to their impact on the long-term future, and how should the related uncertainties be addressed?

RESEARCH PROJECTS

6.3.1 Cost-Benefit vs Well-Being Analysis of Law

Over the last few decades, executive bodies in the United States have used various forms of cost-benefit analysis to evaluate the efficacy of certain proposed policy changes (see, e.g., Adler & Posner, 2000; Carey, 2014; Executive Order No. 12291, 1981; Executive Order No. 12866, 1993; Sen, 2000). Meanwhile, judges and legal academics have engaged in economic analyses of law to inform, influence and explain judicial outcomes (Adler, 2019; Jolls et al., 1998; Kaplow & Shavell, 2002; Posner, 1973; Shavell, 2009). In the European Union and elsewhere, cost-benefit analysis and other forms of economic prioritization are also used in various sectors (see, e.g., Andersson, 2018; Livermore & Revesz, 2013; McCabe et al., 2008). To what degree does the output of these models serve as an accurate proxy for welfare (Sunstein, 2019), particularly with regard to the long term? How have they improved upon previous methodologies, and how can these models improve upon themselves (see Stawasz, 2020)? At what point can/should "well-being analysis of law" replace or supplement "cost-benefit analysis" and "economic analysis of law" (cf. Adler, 2019; Bronstein et al., 2013; Foglia & Jennings, 2013; Sunstein, 2019)? How could existing institutions and instruments be adapted to use welfare-based analysis?

6.3.2 Theories of Legal Change and Change Through the Law

Sunstein (2019) and others (e.g., Anleu, 2009; Dror, 1958; Merryman, 1977) have proposed various theories and accounts of legal change and change through the law. How can insights from this literature better inform how to effectively enact legal mechanisms most likely to positively influence the distant future (specifically with regard to x-risks, s-risks, and p-risks)? In particular, how can these theories be used to evaluate the effectiveness of design and decision-making interventions within legal institutions for protecting future generations? What other methods of analysis ultimately inform our understanding of the potential for our projects to create positive lasting legal change and change through the law?

6.3.3 Behavioral Analysis of Law

Behavioral economics and cognitive psychology have shown the unreliability of statistical and moral intuitions, particularly with respect to events with small probabilities, exponential growth, and large numbers (see, e.g., Dickert et al., 2015; Greene & Baron, 2001; Slovic, 2010; Slovic et al., 2013). How might these intuitions systematically bias our laws and legal decision-making with respect to existential, suffering, and pleasure risks, many of which appear to involve small probabilities, exponential growth, and/or large numbers (cf. Schubert et al., 2019), and what can be done to address such biases? What are the normative legal implications of moral psychology with respect to the above-mentioned risks and protecting the far future more generally?

6.3.4 Jurisprudential Uncertainty

Recent philosophical literature has investigated how our normative uncertainty ought to influence moral decision-making (MacAskill et al., 2020). What are the implications of normative uncertainty with respect to legal theory, i.e., "jurisprudential uncertainty" (Winter, 2021b)? What are the effects of jurisprudential uncertainty on the evaluation of legal norms protecting the long-term future, and related risks in particular? What implications does jurisprudential uncertainty have regarding the justification of relevant criminal laws (Berry & Tomlin, 2020; Winter, 2021b)?

6.3.5 Comparative Legal Longtermism

Modern legal systems vary greatly in a variety of ways, including with respect to (a) primary source of law (cases vs. statutes/code), (b) court system (inquisitorial vs. adversarial), (c) trier of fact (judge vs. jury), and (d) role of past judgments in constraining future ones (see Dainow, 1966; Merryman, 1981; Merryman & Pérez-Perdomo, 2018; Pejovic, 2001; Tetley, 1999). What mechanisms might be used to determine which of these features are more likely to be beneficial or harmful in protecting future generations, and how might this vary by context?

Existing Academic Literature

Acemoglu, D., Johnson, S., & Robinson, J. A. (2005). Institutions as a fundamental cause of long-run growth. In P. Aghion & S. Durlauf (Eds.), *Handbook of Economic Growth, Volume 1A* (pp. 385–472). Elsevier North-Holland.

Adler, M. D. (2019). *Measuring social welfare: An introduction*. Oxford University Press. http://doi.org/10.1093/oso/9780190643027.001.0001

Adler, M. D., & Posner, E. A. (2000). *Cost-benefit analysis: Legal, economic and philosophical perspectives*. University of Pennsylvania, Institute for Law & Economics Research Paper 01-22. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=294422

Barry, C., & Tomlin, P. (2019). Moral Uncertainty and the Criminal Law. In L. Alexander & K. K. Ferzan (Eds.), *The Palgrave handbook of applied ethics and the criminal law* (pp. 445–467). New York: Palgrave. https://www.palgrave.com/gp/book/9783030228101

Bronsteen, J., Buccafusco, C., & Masur, J. (2013). Well-being analysis vs. cost-benefit analysis. *Duke Law Journal*, *62*, 1603–1689. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=3389&context=dlj

Dickert, S., Västfjäll, D., Kleber, J., & Slovic, P. (2015). Scope insensitivity: The limits of intuitive valuation of human lives in public policy. *Journal of Applied Research in Memory and Cognition*, *4*(3), 248-255. https://doi.org/10.1016/j.jarmac.2014.09.002

Epstein, L., Landes, W. M., & Posner, R. A. (2013). *The behavior of federal judges: a theoretical and empirical study of rational choice*. Harvard University Press.

Greene, J., & Baron, J. (2001). Intuitions about declining marginal utility, *Journal of Behavioral Decision Making*, *14*, 243–255. https://doi.org/10.1002/bdm.375

Kahan, D. M. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks, *Nature Climate Change*, *2*, 732–735. https://doi.org/10.1038/nclimate1547

Kaplow, L., & Shavell, S. (2002). *Fairness versus welfare*. Cambridge, Mass: Harvard University Press. (version previously published as Kaplow, L., & Shavell, S. (2001). Fairness versus welfare. *Harvard Law Review*, *114*, 961–1388)

Kaplow, L., & Shavell, S. (2003). Fairness versus welfare: Notes on the Pareto principle, preferences, and distributive justice, *Journal of Legal Studies*, *32*(1), 331–362. http://dx.doi.org/10.1086/345679

MacAskill, W. (2014). *Normative uncertainty* [Doctoral dissertation]. University of Oxford. http://www.academia.edu/download/34857095/Normative_Uncertainty__Complete.pdf

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American Economic Review*, *89*, 103–124. https://doi.org/10.1257/aer.89.1.103

Sen, A. (2000). The discipline of cost-benefit analysis. *The Journal of Legal Studies*, *29*(S2), 931–952. https://doi.org/10.1086/468100

Stawasz, A. (2020, July 4). *Why and How to Value Nonhuman Animals in Cost-Benefit Analyses*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3643473http://dx.doi.org/10.2139/ssrn.3643473

Sunstein, C. R. (2019). *How change happens*. MIT Press.

Sunstein, C. R. (2018). *The cost-benefit revolution*. MIT Press.

Wilson, G. (2013). Minimizing global catastrophic and existential risks from emerging technologies through international law, *Virginia Environmental Law Journal*, *31*(2), 307–364. https://www.jstor.org/stable/44679544

Winter, C. K. (2021b). *Metamoralisches Strafrecht* [Unpublished manuscript].

7 Meta-Research

Meta-research is the study of the methods, reporting, reproducibility, evaluation, and incentives of research (Ioannidis, 2018; Ioannidis et al, 2015). In the context of legal priorities research, the projects in this Section focus on empirical and methodological questions that are relevant to prioritizing causes and specific projects, or that otherwise have implications across many of our cause areas (Sections 4–6, 8, and 9). For instance, below we address what kind of law (comparative, international, or national) or legal actors (judicial, executive, or legislative) legal research ought to prioritize. Projects of this kind allow us to address many important uncertainties that relate to our methodology and thus shape our research approach.

Research Projects

7.1 Comparative vs International vs National Law

As outlined in Section 1.3, legal prioritization favors questions that are not specific to a particular jurisdiction, but that would contribute to the solution of cross-jurisdictional global problems, all else equal. However, research relevant to specific jurisdictions, such as the United States, China or the European Union, may sometimes be prioritized, given the disproportionate impact that these jurisdictions may have on specific risks. What framework allows us to identify whether a specific risk should be tackled by the means of national, international, or comparative law?

7.2 Legislative vs Executive vs Judiciary

Legal research and interventions to improve the long-term future may target specific legal actors or branches of government. Just as we may end up favoring national law over international law for some issues, we may also favor influencing some legal actors or areas of government over others. Is it better to focus on improving, for example, the decision-making of legal actors in the judicial branch than in the executive branch? What are the relevant variables to consider when deciding to target different kinds of legal actors? Under what circumstances should we favor research projects and interventions that target many different kinds of legal actors and multiple branches of government?

7.3 Long-Term Effects of Laws and Legal Institutions

Historically, various laws and legal systems have persisted for hundreds or even thousands of years.[169] Theories of legal change, including path dependence (Hathaway, 2003; Bell, 2012) and the role of precedent (Gerhardt, 1991), account for some of these effects. How do laws leave long-lasting effects on individual attitudes according to current theories of law and social change (Bilz & Nadler, 2014; McAdams, 1997; McAdams, 2000; Sunstein, 1996; Tankard & Paluck, 2016)? How can we identify relationships of influence between the law, long-term effects, and the many other variables at play (for example, other cultural institutions)? How can these theories help legal priorities research understand the scope and longevity of legal efforts to protect future generations? Relatedly, how can we ensure that the long-term effects of law are aligned with protecting future generations?

7.4 Sources of Bias in Legal Priorities Research

In the process of establishing legal priorities research as a new research area, it is important to be aware of sources of bias. Legal priorities research deals with a number of topics that are known to be affected by cognitive biases. For instance, evaluating existential risks (Bostrom, 2002; Schubert et al., 2019; Yudkowsky, 2008b) and forming beliefs about preferred policy solutions (Baron, 2009; Cohen, 2003; Kaplan et al., 2016) are subject to errors in judgements and decision-making. How can legal priorities research most effectively mitigate the effects of cognitive biases? How can we best identify additional biases in cause prioritization and the notion of longtermism, where psychological research has not previously been conducted? What other sources of bias are present in legal priorities research?[170]

7.5 Cross-Cause Prioritization

As stated in Section 3.1, we currently do not conduct our own cause prioritization research, but rely on existing organizations to derive our priorities. However, there may be cases where priorities in law deviate from global priorities. For instance,

[169] Long-term trends in international law (Croxton, 2010), criminal law (Eisner, 2003; Jefferey, 1957; Mueller, 1961), and private law (Baker & Milsom, 2010), among other areas, have left a lasting impact on the law for centuries. These effects also source from landmark court decisions (Hartman et al., 2014). Furthermore, legal systems such as Common law and Roman law both have had persistent effects spanning centuries (Berman, 1985; Watson, 1991), in addition to Eastern legal institutions (Chen et al., 2003; Kuran, 2011). See also lasting legislation (Kysar, 2011) vs temporary legislation (Gersen, 2017).

[170] Cf. Fanelli et al., 2017.

some cause areas may prove to be significantly more neglected or tractable from a legal perspective and thus become a top priority for law. Climate change, although not listed as the "highest-priority area"[171] by some of the organizations engaged in cause prioritization, may be more neglected in law and legal research. What deviations from global priorities within the ITN framework exist for law and legal research?

7.6 Within-Cause Prioritization

In Section 3.2, we outline primary and secondary criteria for identifying research projects within our top cause areas. Conditional on research projects meeting our primary criterion, we apply our secondary criteria to further prioritize and help disentangle different concerns. Given that there is still uncertainty over the best secondary criteria to include, how can we evaluate the effectiveness of our existing criteria and identify new ones? Should we assign greater weight to some criteria over others? Relatedly, what other mechanisms exist for disentangling research and developing guidelines for identifying new projects relevant to legal priorities research? Should there be a general checklist that considers different areas of law, jurisdictions, legal actors, or sub-risks of a cause area, such as the accidents, misuse, and structural risks distinction for AI?

7.7 Broad vs Narrow Legal Approaches

The research projects mentioned in Sections 4 through 6 may vary based on how broad or narrow they are. By broad, we mean how projects or entire cause areas, such as institutional design, can address many kinds of risks and possible futures. Conversely, narrow projects may address one risk or a small class of futures.[172] In the legal context, the broad vs narrow distinction may relate to a number of variables including the type of legal doctrine, legal actor, or jurisdiction used to shape the long-term future. This distinction may be important in evaluating the overall impact of our projects in addition to their practical significance. How can we prioritize between these two kinds of projects to best improve the long-term future? What other factors are relevant to the broad/narrow distinction in the legal context?

[171] See Open Philanthropy's focus on global catastrophic risks and 80,000 Hours' (2020) list of top priorities.

[172] Beckstead (2013a) introduces the broad/narrow intervention distinction: "broad approaches focus on unforeseeable benefits from ripple effects, whereas targeted approaches aim for more specific effects on the far future, or aim at a relatively narrow class of positive ripple effects." See Beckstead (2013b) for informal discussion.

7.8 Risk Mitigation vs Other Trajectory Shifts

As part of our primary criterion (see Section 3.2.1), we argue for the importance of increasing the probability of entering a positive trajectory shift and decreasing the probability of entering a negative one. Existential risks threaten positive trajectories by creating civilizational lock-ins such as extinction or destroying our long-term potential. However, we are also concerned with interventions outside of existential risk mitigation that could positively influence the human trajectory, as opposed to mitigate risks, for example through our work on institutional design. How can we prioritize between existential risk mitigation strategies and other efforts to positively influence the human trajectory? What nuance is missing within this two-fold categorization? Are there other ways to influence the long-term future that are neglected? Are there alternate ways to evaluate existential risks (Avin et al., 2018; Cotton-Barratt et al., 2020; Tonn & Stiefel, 2013) or identify interventions to positively shape the human trajectory?

7.9 Survey of Attitudes of Legal Academics

Appropriately addressing many of the issues raised throughout this agenda may crucially depend on insights and input from legal academia. What are legal academics' views regarding the importance, neglectedness, and tractability of legal priorities research as a general practice? Which cause areas and research questions do legal academics believe to be the highest impact? How can we integrate insights from legal academics into future iterations of the research agenda?

EXISTING ACADEMIC LITERATURE:

Avin, S., Wintle, B. C., Weitzdörfer, J., Ó hÉigeartaigh, S. S., Sutherland, W. J., & Rees, M. J. (2018). Classifying global catastrophic risks. *Futures*, *102*, 20–26. https://doi.org/10.1016/j.futures.2018.02.001

Baron, J. (2009). *Belief overkill in political judgments*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1427862

Bell, J. (2012). Path dependence and legal development. *Tulane Law Review*, *87*, 787–810.

Berman, H. J. (1985). *Law and revolution: The formation of the western legal tradition*. Cambridge, Massachusetts: Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674517769

Bilz, K., & Nadler, J. (2014). Law, moral attitudes, and behavioral change. In E. Zamir & D. Teichman (Eds.), *The Oxford handbook of behavioral economics and the law* (pp. 241–267). Oxford University Press. https://www.law.northwestern.edu/faculty/fulltime/nadler/Bilz-Nadler-LawMoralAttitudesPageProofs.pdf

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Cheng, L., Rosett, A., & Woo, M. (Eds.) (2003). *East Asian law: universal norms and local cultures*. New York: Routledge.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, *85*(5), 808–822. https://doi.org/10.1037/0022-3514.85.5.808

Cotton-Barratt, O., Daniel, M., & Sandberg, A. (2020). Defence in depth against human extinction: Prevention, response, resilience, and why they all matter. *Global Policy*, *11*(3), 271–282. https://doi.org/10.1111/1758-5899.12786

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714–3719. https://doi.org/10.1073/pnas.1618569114

Gerhardt, M. J. (1991). The role of precedent in constitutional decisionmaking and theory. *George Washington Law Review*, *60*, 68–159. https://scholarship.law.wm.edu/facpubs/980

Hathaway, O. A. (2003). Path dependence in the law: The course and pattern of legal change in a common law system. *Iowa Law Review*, *86*, 601–665.

Ioannidis, J. P. (2018). Meta-research: Why research on research matters. *PLOS Biology*, *16*(3), e2005468. https://doi.org/10.1371/journal.pbio.2005468

Ioannidis, J. P., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: evaluation and improvement of research methods and practices. *PLOS Biology*, *13*(10), e1002264. https://doi.org/10.1371/journal.pbio.1002264

Kahan, D. M. (2000). Gentle nudges vs. hard shoves: Solving the sticky norms problem. *The University of Chicago Law Review*, *67*, 607–645. https://chicagounbound.uchicago.edu/uclrev/vol67/iss3/2/

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, *6*, 39589. https://doi.org/10.1038/srep39589

Kuran, T. (2012). *The long divergence: How Islamic law held back the Middle East*. Princeton University Press. https://www.jstor.org/stable/j.ctt7t73p

McAdams, R. H. (2000). A focal point theory of expressive law. *Virginia Law Review*, *86*(8), 1649–1729. https://doi.org/10.2307/1073827

McAdams, R. H. (1997). The origin, development, and regulation of norms. *Michigan Law Review*, *96*(2), 338–433. https://doi.org/10.2307/1290070

Schubert, S., Caviola, L., & Faber, N. S. (2019). The psychology of existential risk: Moral judgments about human extinction. *Scientific reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-50145-9

Sunstein, C. R. (1996). Social norms and social roles. *Columbia Law Review*, *96*(4), 903–968. https://doi.org/10.2307/1123430

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022

Tonn, B., & Stiefel, D. (2013). Evaluating methods for estimating existential risks. *Risk Analysis*, *33*(10), 1772–1787. https://doi.org/10.1111/risa.12039

Watson, A. (1991). *Roman law & comparative law*. Athens, Georgia: University of Georgia Press.

Yudkowsky, E. (2008b). Cognitive biases potentially affecting judgment of global risks. In N. Bostrom & M. M. Ćirković, *Global catastrophic risks* (pp. 91–119).

EXISTING INFORMAL DISCUSSION

Beckstead, N. (2013b, May 27). *A proposed adjustment to the astronomical waste argument* [Online forum post]. LessWrong. https://www.lesswrong.com/posts/5czcpvqZ4RH7orcAa/a-proposed-adjustment-to-the-astronomical-waste-argument#Broad_and_narrow_strategies_for_shaping_the_far_future

Beckstead, N. (2013c). *How to compare broad and targeted attempts to shape the far future* [Presentation]. Future of Humanity Institute, University of Oxford. http://intelligence.org/wp-content/uploads/2013/07/Beckstead-Evaluating-Options-Using-Far-Future-Standards.pdf

Part 3

Cause Areas for Further Engagement

In Part 2, we discussed cause areas that, to a first approximation, appear to best fit our methodology criteria. Here, we outline further cause areas that also fit our methodology criteria but for which further research is needed to more precisely compare them with other cause areas. The Part is split into two Sections and covers space governance (Section 8) and animal law (Section 9). Though we refer to these as cause areas for further engagement, we encourage interested researchers to pursue projects in these fields, both at the meta- and object-level, and may integrate them into our main cause areas in future iterations of this agenda.

8 Space Governance

Becoming a multi-planetary species has the potential to be one of the most crucial steps in the long-term future of humanity. It will be important to safeguard our existence by mitigating the existential risk inherent to depending on a single planet (see Section 2.1), and it could bring opportunities to increase our welfare to standards never seen before. It might also put humanity into a trajectory of great pleasure—or, if space governance is left unattended, put us into a trajectory of immense suffering. For this reason, space exploration has been identified as a major global priority in different stances of prioritization research (80,000 Hours, 2020; Baumann, 2020; Center on Long-term Risk, 2020).

Currently, humanity is experiencing a second space race, as the necessary technology becomes more accessible to an increasing and diverse number of public and private actors. According to Devezas et al. (2012, p. 983), "[t]he strongest feature of this new space race will be a multipolar struggle for dominance in the new external border of the planet Earth, the 4th frontier, whether for political, military or commercial purposes." In that context, developing a fair and efficient legal framework for space governance is increasingly urgent, especially given that certain problems must be resolved before space exploration breaks out. However, space law is still a neglected field in many regards. International law, with the main piece of legislation being the 1967 Outer Space Treaty (OST) (United Nations Office for Outer Space Affairs [UNOOSA], 2020a),[173] still does not satisfactorily encompass many contemporary concerns, leaving room for nations to push their agenda through domestic regulations and bilateral accords (Tronchetti, 2013). Likewise, there are still not as many legal researchers dedicated to studying space law as its complexity and importance requires.[174]

The following projects aim at addressing that gap by contributing to a legal foundation for the governance of outer space in the long-term.

8.1 Coordination and Peace in Outer Space

This category of projects is concerned with optimizing human coordination efforts towards an orderly, peaceful expansion to outer space. The main forum for promoting peaceful cooperation in outer space is the United Nations Committee on the Peaceful Uses of Outer Space (UNCOPUOS), a committee established in 1959 under the United Nations Office for Outer Space Affairs (UNOOSA) with the goal of governing the sustainable use of space for the benefit of all humanity. This international forum has a particularly relevant role regarding the cooperative use of space in issues like climate change monitoring, disaster management, and preserving outer space for future generations. However, competing interests among nations still limit cooperation in different ways. These roadblocks impede the development of more sophisticated legal frameworks and create an uncertainty that is prone to foster conflict.

RESEARCH PROJECTS

8.1.1 Fostering International Cooperation in Space Exploration

Developing strategies in international law to bring nations together is arguably the basis of any regulatory effort in space law at the moment (Bittencourt Neto et al., 2020). If the underlying political conditions of a given agreement or regulatory piece are not well understood and addressed, the project will be inherently ineffective, as many efforts have been in the last decades (Lyall & Larsen, 2018). What are the particular concerns of space-faring nations that impede cooperation? How will cooperation adapt to developing nations entering the space race? What are the

[173] For an overview of the history of space law and governance, see UNOOSA (2020b) and Lyall and Larsen (2018).

[174] Some of the main academic centers for space law include the University of Leiden, the University of Nebraska, McGill University, and the University of Mississippi. We believe there are not as many centers dedicated to space law as its importance makes necessary.

most effective legal mechanisms to materialize international cooperation (for example, bilateral agreements, such as the Artemis Accords, vs. multilateral instruments)?

8.1.2 Legal Interoperability of Actors and Equipment in Space

Jurisdictions often have different regulatory standards for equipment, licensing, crew management, and other sensitive processes in outer space. Making materials and processes interoperable means harmonizing the differences in such standards in order to allow them to communicate with each other more efficiently (National Aeronautics and Space Administration [NASA], 2018). This is a promising line of research, as it investigates more concrete, practical legal challenges in international cooperation, involving fields such as export regulations and competition law, with the potential to lay the ground for more sophisticated models of technological cooperation in outer space. Which opportunities are there to harmonize existing standards? What role will the emergence of private actors and the commercialization of space activities play in standardizing processes? How can we ensure flexibility in legal mechanisms that allows expanding the potential of interoperable systems?

8.1.3 Regulating the Use of Weapons in Outer Space

Despite the utmost relevance of preventing warfare in outer space to safeguard humanity, existing international law on this subject is still limited (Schrogl et al., 2020). The Outer Space Treaty contains vague and narrow provisions, limited to providing that celestial bodies must be used for exclusively peaceful purposes and prohibiting the placement of nuclear weapons and other kinds of weapons of mass destruction in outer space. The Moon Agreement, on the other hand, is comprehensive but has only 18 states parties and four signatories, none of which are among the major space-faring nations.[175] Besides that, clear definitions are lacking, such as what is a "weapon," and more specific, widely-accepted provisions about other types of weapons, dual-use equipment, and other potentially destructive spacecraft

[175] Other useful legal mechanisms are (a) the Convention on the Prohibition of the Use of Environmental Modification Techniques for Military or Any Other Hostile Purposes (ENMOD Convention), which entered into force in 1978 and concerned the deliberate manipulation of the natural process of the dynamics, composition, and structure of the Earth and outer space for hostile or military purposes; (b) the Treaty on the Prohibition of Nuclear Weapons, which will enter into force in 2021 and leads towards the total elimination of nuclear weapons; and finally (c) the general international law through the Charter of the United Nations that establishes limits to conflicts and the use of force.

are non-existent (Jinyuan, 2017; von der Dunk, 2009). Attempts to fill in these gaps in international law by the European Union, Russia, and China, for example,[176] have failed. As it currently stands, international law provides insufficient regulations to prevent an arms race in outer space. Alternatives to that problem include the adoption of (a) confidence-building and security-building measures (CSBMS), (b) politically binding codes of conduct for space activities, and (c) (most ambitiously) an international prohibition of weapons in space. Which of these options is most promising and/or viable given the constant evolution of technology, and what are the potential deadlocks that might prevent the realization of each of these options? How can related on-going projects about the regulation of military activities in outer space, such as the Woomera Manual and the MILAMOS Project, contribute to the long-term peace of space exploration?

8.1.4 Sharing the Benefits of Space-Related Technology and Space Resource Activities

The concentration of power in few nations or companies has the potential to disrupt economies on Earth. For example, water ice and lunar regolith already pose international legal challenges (De Man, 2016). As other mineral-rich celestial bodies, such as asteroids, are explored in the long-term future commodities markets might be affected to the extent of severely harming developing countries dependent on exports of these products (Jakhu et al., 2017; Pop, 2008; Tronchetti, 2009). What should be the legal status of space resources? How should ownership rights be applied to extra-terrestrial resources? How to avoid concentrating resources on few space-faring nations and companies, sharing its benefits?

8.1.5 Planning the Legal Governance of Different Possibilities of Space Settlement

We might conceive of space governance as the administration of concentrated human settlements that have outreach activities in their surroundings (for example, settlements on Mars or the Moon), similarly to societal organization on Earth. However, considering the particularities of outer space and the pace of technological advancement, settlements might take a completely different shape. Developments along the lines of von Neumann probes (Sagan & Newman, 1983) and O'Neill cylinders (O'Neill, 1974) might allow a much more spread-out exploration of space by making isolated, autonomous colonies and machines possible. Distinct models of society will facilitate or hinder specific types of governance, such as

[176] See the 2014 Russian and Chinese Treaty on the Prevention of the Placement of Weapons in Space, the Threat or Use of Force Against Space Objects (PPWT).

authoritarian rule in confined, isolated settlements or anarchic-like, -*laissez-faire* commercial societies (Cockell, 2016). How should institutions be designed to accommodate these possibilities? How would a legal system steer space colonization towards (or away from) a certain model of governance? Which model would suit best each possibility of space settlement? Which existing legal mechanisms and institutions might serve as an inspiration (for example, UN, EU, ISS, Antarctica, international waters)?

8.1.6 Governing International Cooperation for Planetary Defense

Asteroids greater than 1km across "threaten global catastrophe and may also be large enough to pose an existential risk" (Ord, 2020, p. 70). Luckily, the probability of a relevant asteroid impact on the Earth is dim, at around 1 in 120,000 (*idem*). Even if an asteroid heads towards our planet, humanity already has the necessary infrastructure and expertise to predict it and the required weapons to destroy or deflect it. The greater challenge, however, is consolidating a global cooperation strategy that allows us to confidently develop a procedure to tackle the risk from near-Earth objects collisions, such as the one outlined by Drube et al. (2020). Should there be an explicit protocol for the use of nuclear weapons in such conditions? To what extent should nations be held liable for failed deflection attempts? Should there be a collective obligation to protect a threatened nation? Is a decision-making body for planetary defense desirable and, if so, what could it look like?

8.2 Sustainability of the Long-Term Presence of Humanity in Outer Space

In this subsection, we list projects concerned with how humanity will interact with outer space. Ensuring that we act sustainably will be crucial to guarantee our longevity as a multi-planetary species. Some of the issues involved with our long-term presence are already troubling, such as managing space debris and large constellations of satellites. Others involve greater uncertainty, such as developing protocols for interacting with extraterrestrial life. The following projects approach these themes in different ways.

RESEARCH PROJECTS

8.2.1 Environmental Concerns and Sustainable Use of Outer Space

Any space activity constitutes an unavoidable risk of contaminating outer space with microorganisms, pollutants, and waste. Attempts to regulate that field, such as the 2019 Guidelines for the Long-term Sustainability of Outer Space Activities (UNOOSA, 2020c), adopted by the United Nations Committee on the Peaceful Uses of Outer Space (UNCOPUOS), have advanced the debate but failed to bind nations and private actors to its recommendations. Which are the most effective legal mechanisms to ensure compliance with guidelines such as these? What regulations should be put into place to protect our planet (and other human-inhabited places) from extraterrestrial life or bioactive molecules in returned samples (NASA, 2020)? What can we learn and adapt from environmental law on Earth? How the law can contribute to balancing the right to explore outer space with the need to preserve it for future generations?

8.2.2 Long-term Human Presence in Space and Human Enhancement

Outer space brings several challenges to human health that we have not yet tackled, such as exposure to high-energy ionizing radiation. One of the most promising avenues to address these issues and ensure humanity's long-term presence in space is genetic modification. For example, studies have recently investigated the possibility of combining human cells with those of other species resistant to extreme environments, such as tardigrades, by employing novel techniques such as CRISPR. However, genome editing also brings to the table ethical and legal concerns (for a more detailed discussion, see Section 5.3.5). Is it conceivable to think of distinct regulatory instruments for individuals on Earth and for those who will inhabit the space settlements, considering the disparate living conditions? If so, how can the law mitigate the risks involved with separately modifying the genome of isolated parts of the human population?

8.2.3 Protocol for Governing Interactions with Extraterrestrial Intelligence

Although foreseeing how extra-terrestrial life will take form is inherently uncertain, preparing beforehand for possible scenarios is crucial to avoid makeshift solutions—especially as some countries have already made efforts to contact extraterrestrial intelligent life, such as the Voyager mission. Developing such a protocol would contribute to a more coordinated and ethical interaction. Legal scholars have debated some first principles that might serve as a stepstone for this project. Haley coined the Interstellar Golden Rule (Haley, 1963), later developed by Fasan into universal rights such as the prohibition of damaging another race, the right of a race to self-defense, and the right to adequate living space (Fasan, 1990). This project fits into a broader discussion about how to deal with non-human sentient beings more broadly (see projects referring to sentience in Sections 5.3.5, 6.1.10, 6.2.4, and 9.2.1). How can we develop these principles further into an internationally accepted protocol? Which scenarios should the protocol be flexible enough to cover, but sufficiently rigid to prevent?

8.2.4 Legal Protection of Science and Astronomy

The pollution of the Earth's orbit might lead to irreversible damage to scientific endeavors, as visibility of objects in space is hindered and scientific equipment is put at risk by mega-constellations of satellites. This process might even be repeated in other celestial bodies, such as the Moon and Mars, if a sustainable process of development is not designed beforehand; this could prevent humanity from developing evidence-based strategies to occupy outer space. How can we ensure scientists have representation, voice, and power within international bodies in the long-term future? How can the law contribute to a healthy cooperation between scientists and other stakeholders?

8.2.5 Legal Regulation of Space Debris

Space debris orbiting the Earth is currently considered one of the most critical threats to space activities (Klinkrad, 2010). The growth of this issue might lead to irreversible problems, such as a Kessler syndrome, and might be a concern for human settlements on other celestial bodies if regulatory solutions are not developed in a timely manner. At this point, international law has not sufficiently addressed this problem, with only non-binding legal mechanisms, such as the 2007 UN-COPUOS Space Debris Mitigation Guidelines (UNOOSA, 2007). How do we ensure the accountability of space actors in mitigating the negative consequences of space debris? What could be the legal incentives for debris control, especially for commercial players? How can the financial load be distributed to treat a potentially enormous damage caused by a limited number of actors? What would be the better system to guarantee legal certainty and keep up with technological development?

EXISTING ACADEMIC LITERATURE

Banner, S. (2008). *Who owns the sky? The struggle to control airspace from the Wright brothers on*. Harvard University Press.

Bittencourt Neto, O. (2015). *Defining the limits of outer space for regulatory purposes*. Springer.

Bittencourt Neto, O., Hofmann, M., Masson-Zwaan, T., & Stefoudi, D. (2020). *Building blocks for the development of an international framework for the governance of space resources activities: A commentary*. Eleven International Publishing. https://www.boomdenhaag.nl/en/webshop/building-blocks-for-the-development-of-an-international-framework-for-the-governance-of-space-resource-activities

Brunner, C., & Soucek, A. (Eds). (2011). *Outer space in society, politics and law*. Springer.

Cheng, B. (1998). *Studies in international space law*. Clarendon Press, Oxford.

Cockell, C. (Ed.) (2016). *Dissent, revolution, and liberty beyond* Earth. Springer.

De Man, P. (2016). *Exclusive use in an inclusive environment: The meaning of the non-appropriation principle for space resource exploitation (Vol. 9).* Springer.

Drube, L. et al. (2020). Planetary defence: Legal overview and assessment. Space Mission Planning Advisory Group. https://www.cosmos.esa.int/documents/336356/336472/SMPAG-RP-004_1_0_SMPAG_legal_report_2020-04-08+%281%29.pdf/60df8a3a-b081-4533-6008-5b6da5ee2a98?t=1586443949723

Jakhu, R. S., Pelton, J. N., & Nyampong, Y. O. M. (2017). *Space mining and its regulation.* Springer. https://doi.org/10.1007/978-3-319-39246-2

Jinyuan, S. U. (2017). Space arms control: Lex lata and currently active proposals. *Asian Journal of International Law*, *7*(1), 61–93. https://doi.org/10.1017/S2044251315000223

Lyall, F., & Larsen, P. (2018). *Space law: a treatise*. Routledge.

Masson-Zwaan, T., & Hofmann, M. (2019). *An introduction to space law* (4th ed.). Kluwer.

Pop, V. (2008). *Who owns the Moon?* Springer.

Schrogl, K. U., Hays, P. L., Robinson, J., Moura, D., & Giannopapa, C. (Eds.). (2020). *Handbook of space security*. Springer.

Tronchetti, F. (2009). *The exploitation of natural resources of the Moon and other celestial bodies: a proposal for a legal regime* (Vol. 4). Martinus Nijhoff Publishers.

Tronchetti, F. (2013). *Fundamentals of space law and policy*. Springer.

Viikari, L. (2008). *The environmental element in space law*. Martinus Nijhoff Publishers.

von der Dunk, F. and Tronchetti, F. (2015). *Handbook of space law*. Edward Elgar Publishing.

Wolter, D. (2005). *Common security in outer space and international law*. United Nations Institute for Disarmament Research. https://www.unidir.org/files/publications/pdfs/common-security-in-outer-space-and-international-law-283.pdf

EXISTING INFORMAL DISCUSSION

Baumann, T. (2020). Space governance is important, tractable, and neglected [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/QkRq6aRA84vv4xsu9/space-governance-is-important-tractable-and-neglected

Gradoni, L. (2018). What on Earth is happening to space law? [Blog post]. Blog of the European Journal of International Law. https://www.ejiltalk.org/what-on-earth-is-happening-to-space-law-a-new-space-law-for-a-new-space-race

9 Animal Law

At present, very few people are working on animal law from a longtermist perspective, and very few people are working on longtermism from a multi-species perspective. This separation of animal law and longtermism is understandable. On the one hand, given that humans harm and kill trillions of animals per year (e.g., Humane Ventures), it can be easy for people working on animal law to prioritize short-term legal reforms for animals over long-term legal revolutions for animals. On the other hand, given that humans have the power to determine whether or not there is a long-term future for humans and nonhumans alike, it can be easy for people working in longtermism to prioritize humans over nonhumans. As a result, while there is a lot of mutual sympathy across these communities, there is not much collaboration.

We believe this separation between animal law and longtermism is a mistake, in both directions. First, we see longtermism as essential to animal law. When working within legal and political frameworks designed by and for humans, there is a limit to how much good can be done for nonhumans through legal reforms. At present, the law classifies humans as legal and political subjects and nonhumans as legal and political objects (Wise, 2000; Andrews et al., 2019), and it applies concepts such as personhood, citizenship, representation, fairness, justice, equality, capitalism, liberalism, democracy, and more accordingly. In order to bring about systemic change for animals, this approach to the law must be challenged, along with the impact that this approach has had on our institutions. This is necessarily a long-term project.

Second, we also see animal law as essential for longtermism. This is partly true for the sake of other animals. Nonhumans represent more than 99% of the world's population (Tomasik, 2019a), and we have a responsibility to reduce nonhuman suffering, on the grounds that nonhuman suffering is massive, neglected, and tractable (cf. Section 3; Duda, 2016). Furthermore, nonhuman suffering is increasingly human-caused (see Singer, 1975; Sebo, forthcoming). This is also partly true for the sake of humans. Our treatment of nonhumans is not only linked to pandemics (e.g., World Organisation for Animal Health, 2020), climate change (e.g., Goodland & Anhang, 2009), and other threats, but it is also linked to harmful beliefs and values that favor the needs of the privileged few over the oppressed many (e.g., Caviola, 2019). Thus, expanding our legal and political circle to include nonhumans is essential as a means to improving human and nonhuman lives alike in the long run.

As a preliminary matter, a question remains as to whether incremental or fundamental legal change would plausibly do the most good for animals. We believe that the answer to this question, which hinges on many difficult empirical and normative judgments, is highly uncertain. We also doubt that these strategies are mutually exclusive; for instance, some incremental changes for animals might also help make fundamental changes more feasible.[177] Thus, we believe the optimal approach will likely include a mixture of both strategies.

This Section discusses both incremental changes *within* existing legal frameworks as well as fundamental legal changes *to* existing legal frameworks. These range from relatively moderate changes, such as reforming anti-cruelty laws to protect farmed animals and wild animals more effectively, to relatively radical changes, such as extending personhood to animals, extending citizenship to animals, and creating new political institutions for representing animals. As this discussion will make clear, legal research to pursue many of these priorities simultaneously would do the most good possible in the long run.

9.1 Incremental Changes

Even the incremental gains that animal law can help secure can impact many animals. Current data suggests that many tens of billions of land animals are slaughtered for human consumption every year (Food and Agriculture Organization of the United Nations). The vast majority of these animals live in conditions that are far from optimal and that very often yield lives that are probably not worth living at all (e.g., Thompson, 2020). For example, many egg-laying hens live in battery cages that make extending wings or engaging in almost any natural foraging behaviors impossible, and many sows are confined in gestation crates that are too narrow to permit even turning around. On top of these structural features, land animals raised for food are often subject to routine animal abuse, resulting from a lack of regulation and/or enforcement of existing laws (Hodges, 2010). These examples are in addition to uncounted trillions of aquatic animals killed annually for human consumption, often using extremely painful fishing methods (see Braithwaite, 2010) or aquaculture facilities ("fish farms") with common pain-causing features like poor water quality and overcrowding (Cerqueira & Billington, 2020). Moreover, the number of wild animals, even of plausibly sentient wild animals, alive today outnumber the number of humans alive today by many orders of

[177] For example, interventions that reduce meat eating over the short term may lower the cognitive dissonance that serves as a barrier to appropriately nonspeciesist moral judgments and actions, which could make more fundamental gains more feasible.

magnitude, and the likelihood that many of them live grim or even net-negative-welfare existences is very high (Tomasik, 2015b). The following research topics relate to incremental changes to animal law.

Research Projects

9.1.1 International and Comparative Animal Law

What legal mechanisms, including regulations by government agencies, statutes, constitutional provisions, and treaties, have the biggest positive impact on animals and why? How can successes in these areas in certain countries be replicated in other countries (see, e.g., Stilt, 2018), and how can shortcomings be overcome (see, e.g., Blattner, 2019a)? How can existing international legal movements and trends be translated or adapted to account for animals' interests (see, e.g., Blattner, 2019b; Peters, 2020)? And how can local, national, and international efforts be assessed, coordinated, or integrated optimally?

9.1.2 Expanding on Wild Animal Laws

How can existing conservation laws, such as the American Endangered Species Act or the Convention on International Trade in Endangered Species of Wild Fauna and Flora ("CITES"), be used or amended in ways that account for not just population survival, but individual animals' welfare? How might such protections extend to wild animals that are not endangered or threatened (see Animal Ethics, 2020 for a survey of relevant legislation)?

9.1.3 Identifying Opportunities for Laws Benefiting Aquatic Animals

Given the sheer magnitude of suffering among aquatic animals (Balcombe, 2016; Braithwaite, 2010) and the comparative lack of legal protections for individual such animals (Levenda, 2013), it is crucial to design novel beneficial regulations from the beginning. What are the greatest unmet needs for such animals that the law could help meet? How could novel regulations maximize impact as this field draws progressively more attention from animal activists?

9.1.4 Optimizing Animal Cruelty Laws

Animal cruelty laws have existed for centuries (see, e.g., Massachusetts Body of Liberties, 1641) and continue to represent some of the animal-related laws with which the public is most familiar in many countries. Yet they tend to be enforced disproportionately in settings in which animal suffering is comparatively mild and in which costs for defendants are disproportionately high (e.g., Marceau, 2019). Theorizing about and designing more effective animal cruelty laws can help focus enforcement where the potential benefits are highest. For instance, such laws could criminalize particularly cruel food-production methods like gestation crates and battery cages.

9.1.5 Supporting Alternative Protein

Developing and promoting alternative proteins represents a promising avenue to preventing trillions of animals from suffering and dying each year in our global food system (e.g., Animal Charity Evaluators, 2020). How could the law optimally support adoption of plant-based and cell-based meat—for instance, by reducing subsidies for conventional meat, increasing subsidies for alternative proteins, banning misleading labels for conventional meat, and resisting efforts to impose unattractive labels on alternative proteins (cf. Negowetti, 2018)?

9.1.6 Supporting Animal-Friendly Education and Advocacy

Legislators and regulators have many other mechanisms at their disposal to benefit animals. For instance, governments can sponsor pro-animal education by improving humane education in public schools and by engaging in public outreach around animal welfare and rights. Governments can also ban "ag-gag" laws that hinder whistleblowing and undercover investigations on factory farms as well as other laws that hinder activists. Which legislative and administrative levers could plausibly benefit animals the most and help create better regulatory outcomes for animals?

9.1.7 Meta-Option: Supporting Academic Animal Law Programs

Relatively few tenure-stream professors focus on animal law, in part because relatively few law schools have animal law programs, or even more than a single animal law class. Moreover, while a small number of legal journals focus exclusively on animal law,[178] their reach tends to be limited. Since successful programs can plausibly have a sizable impact on scholars, students, and the general public, supporting the development of new programs and improvement of existing programs—in research, teaching, programming, and more—could have a substantial impact.

[178] These journals include the Animal Law Review, the Journal of Animal Law, the Animal Law eJournal, the Global Journal of Animal Law, and the Journal of Animal & Natural Resources Law.

9.2 Fundamental Changes

While progress for animals can be made within existing legal and political frameworks, there is a limit to how much progress can be made that way. Current legal and political frameworks were built by and for (some) humans, yet nonhumans constitute more than 99% of our community. Progress for animals in the long run necessitates fundamental legal and political change as well, either by making current frameworks much more inclusive or by replacing them with other, much more inclusive alternatives. This means questioning many basic conceptual, empirical, and normative assumptions that legal scholars currently make. Consider some examples.

RESEARCH PROJECTS

9.2.1 Personhood

What is legal personhood, and who can be a legal person? In our current legal system, an entity can either be a legal person, with the capacity for rights, or a legal object, without the capacity for rights. And while humans (and stand-ins for human interests such as corporations) are classified as legal persons, nonhumans are classified as legal objects. As a result, legal options for protecting humans are much more expansive than legal options for protecting nonhumans. Addressing this issue requires either (a) extending legal personhood to nonhuman animals or (b) creating a new middle-ground category, such as a category for "sentient beings," and placing nonhumans in this category instead. Legal research could develop these approaches, evaluate them, and pursue either or both (Andrews et al., 2019; Deckha 2020; Kurki & Pietrzykowski, 2017).

9.2.2 Citizenship

What is citizenship, and who can be a citizen? In our current system, citizenship involves a wide range of rights, including a right to political representation and a right to reside in and return to your country of residence. And while some humans are classified as citizens, all nonhumans are not. As a result, legal options for representing, and protecting, humans are once again more expansive than legal options for representing, and protecting, nonhumans. Addressing this issue requires either (a) extending citizenship to some nonhuman animals or (b) creating a new middle-ground category, such as a category for "nonhuman members of the state," and placing nonhumans in this category instead. Legal research could develop these approaches, evaluate them, and pursue either or both as well (Donaldson & Kymlicka 2011).

9.2.3 Representation

How can we increase representation for nonhuman animals in human-administered political systems? One option is to increase informal representation of animals, for instance by including animals in impact assessments (see Stawasz, 2020) or by creating public assemblies that can advise the state on matters concerning animals. Another option is to establish formal representation for animals, for instance by creating a legislative house to represent the interests of animals (as well as other non-voting stakeholders), and by creating mechanisms to ensure that this house does this work faithfully. Legal research could develop, evaluate, and, possibly, implement these or other options for representing animals (Cochrane, 2018).

9.2.4 Fairness, Justice, and Equality

How should we understand basic concepts like fairness, justice, and equality, which underpin many legal theories, in a multi-species political society? For instance, many political theorists believe that these values require distributing social benefits and burdens such that they will either do the most good possible in general, or do the most good possible for the worst-off among us in particular. Plausibly, in a multi-species society, either interpretation would require allocating as many social benefits to nonhumans as possible, all else being equal. Should the law accept this apparent implication of these values? If so, a lot of work will be required to achieve this goal. If not, what is the true nature of these basic political values (Nussbaum, 2006)?

9.2.5 Capitalism, Liberalism, and Democracy

How should we understand basic concepts like capitalism, liberalism, and democracy in a multi-species political society? For instance, many people assume that using the law to coercively reduce the use of animals and increase support for animals interferes with the free market, individual liberty, and collective self-determination. However, if animals should be legal and political subjects rather than legal and political objects, then these assumptions are called into question. For example: In such a regime, when, if ever, can animals be owned? What, if anything, can they own? How much weight should their welfare and liberty carry in policy decisions? How much weight should their (human-represented) voices carry in policy decisions (Smith, 2012)?

9.2.6 Implications

What might follow from these foundational changes for a wide range of legal issues? For example, what changes might follow in education policies, employment policies, social services policies, infrastructure policies, and more in a multi-species political society? Plausibly, all would change substantially. To take infrastructure as an example, there may be substantially reduced deforestation and increased reforestation, for the sake of humans as well as nonhumans. And, insofar as cities are built, they may consider the needs of human as well as nonhuman residents, perhaps resulting in more urban parks, bird-friendly windows on buildings and vehicles, animal overpasses and underpasses on roads, and feeding stations, water stations, and habitat for nonhumans throughout (Sebo, 2020).

9.2.7 Related and Meta-Questions

Alongside these basic legal and political questions are many related empirical and normative questions. What is the basis of well-being, and which animals have the capacity of well-being? Can some animals have more well-being than others? Which nonhumans can flourish more in relatively captive environments, and which can flourish more in relatively free environments? Which nonhumans can benefit from expanded populations, and which can benefit from contracted populations? How do current policies impact nonhuman populations, and how might alternative policies impact them? How can legal research answer these questions responsibly, given current limits on knowledge and power, as well as speciesism, self-interest, and group interest (Sebo, 2021a)?

9.2.8 Timing and Prioritization of Research

Beyond these incremental and fundamental goals, a further question is when to pursue them. On one hand, there is a strong case for pursuing them all now. Nonhuman suffering is massive, neglected, and tractable. Additionally, since human neglect, exploitation, and extermination of nonhuman animals is linked to other global threats, reducing our use of nonhuman animals and increasing our support for nonhuman animals will benefit humans too. Each year results in unnecessary harm and death for trillions (or more) of sentient nonhumans, as well as delay of essential work towards addressing multi-species threats and expanding moral, legal, and political circles.

On the other hand, there is also a strong case for pursuing at least some of these goals later. There are extreme limits on knowledge, power, and political will at present. So, even if we wanted to help animals, we would lack the ability to do so ethically and effectively, especially at scale. Additionally, an initial focus on humans might involve many indirect benefits that a focus on nonhumans might not. Since humans will be administering multi-species legal and political systems for the foreseeable future, the more human needs are addressed now, the more future generations will be able to address human and nonhuman needs alike in the future.

Thus, one might think that, even if nonhumans should be a priority in the long run, humans should be a priority in the short term in part as a means to this end.

These considerations are compelling and favor a balanced approach. In particular, legal research should consider both humans and nonhumans now, and it should favor human needs only insofar as doing so is necessary for securing a good long-term future for humans and nonhumans alike. This approach allows legal research to consider human and nonhuman needs holistically; to seek shared solutions to shared problems; and to build knowledge, power, and political will toward helping humans and nonhumans alike more effectively in the long run. At the same time, this approach also favors human needs to a degree in the short term, insofar as doing so is necessary for securing a positive future for humans and nonhumans alike in the long run.

Of course, even on this balanced approach, there may be disagreement about exactly how much to focus on human and nonhuman needs in the short to medium term. These are difficult questions that not resolved here. However these questions are answered, legal research will have to consider the needs of humans and nonhumans alike. After all, if human needs may be favored to a degree in the short term, the reason is that doing so is the most effective way to improve human and nonhuman lives in the long run. And of course, in order to evaluate this strategy (as well as other, alternative strategies), legal research must consider nonhuman animals in discussions about law and politics much more than it does today.

EXISTING ACADEMIC LITERATURE

Andrews, K. (2020). *The animal mind*. Routledge.

Andrews, K., Comstock, G. L., Crozier G. K. D., Donaldson, S., Fenton, A., John, T. M., Johnson, L. S. M., Jones, R. C., Kymlicka, W., Meynell, L., Nobis, N., Pena-Guzman, D., & Sebo, J. (2019). *Chimpanzee rights: The philosophers' brief*. Routledge. https://www.routledge.com/Chimpanzee-Rights-The-Philosophers-Brief/Andrews-Comstock-GKD-Donaldson-Fenton-John-Johnson-Jones-Kymlicka-Meynell-Nobis-Pena-Guzman-Sebo/p/book/9781138618664

Balcombe, J. (2016). *What a fish knows: The inner lives of our underwater cousins*. Oneworld Publications.

Blattner, C. E. (2019a). The recognition of animal sentience by the law. *Journal of Animal Ethics*, *9*(2), 121–136. https://doi.org/10.5406/janimalethics.9.2.0121

Blattner, C. E. (2019b). *Protecting animals within and across borders: Extraterritorial jurisdiction and the challenges of globalization*. Oxford University Press.

Braithwaite, V. (2010). *Do fish feel pain?* Oxford University Press.

Caviola, L. (2019). *How we value animals: The psychology of speciesism* [Doctoral dissertation]. https://ora.ox.ac.uk/objects/uuid:9000a8be-b6dc-4c63-88e8-974b9daaa83a

Cochrane, A. (2020). *Should animals have political rights?* Polity Press.

Cochrane, A. (2018). *Sentientist politics: a theory of global inter-species justice*. Oxford University Press.

Deckha, M. (2020). *Animals as legal beings: Contesting anthropocentric legal orders*. University of Toronto Press.

Donaldson, S., & Kymlicka, W. (2011). *Zoopolis: A political theory of animal rights*. Oxford University Press.

Gabardi, W. (2017) *The next social contract: Animals, the anthropocene, and biopolitics*. Temple University Press.

Goodland, R., & Anhang, J. (2009, November). *Livestock and climate change*. A Well-Fed World. https://awellfedworld.org/wp-content/uploads/Livestock-Climate-Change-Anhang-Goodland.pdf

Gruen, L. (2011). *Ethics and animals: An introduction*. Cambridge University Press.

Kagan, S. (2019). *How to count animals, more or less*. Oxford University Press.

Korsgaard, C. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Levenda, K. (2013). Legislation to protect the welfare of fish. *Animal Law*, *20*, 119–144. https://www.animallaw.info/sites/default/files/lralvol20_1_119.pdf

Marceau, J. (2019). *Beyond cages: Animal law and criminal punishment*. Cambridge University Press.

Negowetti, N. E. (2018). Establishing and enforcing animal welfare labeling claims: Improving transparency and ensuring accountability. *Journal of Animal & Natural Resource Law*, *14*, 131–158.

Nussbaum, M. C. (2009). *Frontiers of justice: Disability, nationality, species membership*. Harvard University Press.

Palmer, C. (2010). *Animal ethics in context*. Columbia University Press.

Peters, A. (2020). Toward international animal rights. In A. Peters (Ed.), *Studies in global animal law* (pp. 109–120). Springer. https://doi.org/10.1007/978-3-662-60756-5_10

Regan, T. (1983). *The case for animal rights*. University of California Press.

Schlottmann, C., & Sebo, J. (2018). *Food, animals, and the environment: An ethical approach*. Routledge.

Sebo, J. (2021a). Animals and climate change. In M. Budolfson, T. McPherson & D. Plunkett (Eds.), *Philosophy and climate change*. Oxford University Press.

Sebo, J. (2021b). *Animal ethics in a human world* [Forthcoming]. Oxford University Press.

Singer, P. (1975). *Animal liberation*. Random House.

Smith, K. (2012). *Governing animals: Animal welfare and the liberal state*. Oxford University Press.

Stawasz, A. (2020, July 4). *Why and how to value nonhuman animals in cost-benefit analyses*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3643473

Stilt, K. A. (2018). Constitutional innovation and animal protection in Egypt. *Law & Social Inquiry*, *43*(4), 1364–1390. https://doi.org/10.1111/lsi.12312

Sunstein, C., & Nussbaum, M. (Eds.) (2012). *Animal rights: Current debates and new directions*. Oxford University Press.

Thompson, P. B. (2020). Philosophical ethics and the improvement of farmed animal lives. *Animal Frontiers*, *10*(1), 21–28. https://doi.org/10.1093/af/vfz054

Wise, S. M. (2000). *Rattling the cage: Toward legal rights for animals*. Da Capo Press. https://dacapopress.com/titles/steven-wise/rattling-the-cage/9780306824005/

References

80,000 Hours (2020, May). *Our current view of the world's most pressing problems*. https://80000hours.org/problem-profiles/

Abecassis, A., Bullock, J. B., Himmelreich, J., Hudson, V. M., Loveridge, J., & Zhang, B. (2020, June 26). *Contribution to a European agenda for AI: Improving risk management, building strong governance, accelerating education and research*. Medium. https://medium.com/berkman-klein-center/contribution-to-a-european-agenda-for-ai-13593e71202f

Ace Project. (2020). *Electoral systems*. https://aceproject.org/epic-en/CDTable?question=ES005&view=country

Acevedo-Rocha, C. G. (2016). The synthetic nature of biology. In K. Hagen, M. Engelhard, & G. Toepfer (Eds.), *Ambivalences of creating life* (pp. 9–53). Springer, Cham. https://dx.doi.org/10.1007%2F978-3-319-21088-9_2

Ad Hoc Technical Expert Group on Synthetic Biology (2015, October 7). Report of the ad hoc technical expert group on synthetic biology (UNCEP/CBD/SYNBIO/AHTEG/2015/1/3). Convention on Biological Diversity. https://www.cbd.int/doc/meetings/synbio/synbioahteg-2015-01/official/synbioahteg-2015-01-03-en.pdf

Adamczewski, T. (2019, May 25). *A shift in arguments for AI risk*. Fragile Credences. https://fragile-credences.github.io/prioritising-ai

Adams, F. C., & Laughlin, G. (1997). A dying universe: The long-term fate and evolution of astrophysical objects. *Reviews of Modern Physics*, 69(2), 337. https://doi.org/10.1103/RevModPhys.69.337

Adams, F. C., & Laughlin, G. (1999). *The five ages of the universe: Inside the physics of eternity*. Simon & Schuster.

Adler, M. D. (2019). *Measuring social welfare: An introduction*. Oxford University Press. http://doi.org/10.1093/oso/9780190643027.001.0001

Adler, M. D., & Posner, E. A. (2000). *Cost-benefit analysis: Legal, economic and philosophical perspectives*. University of Pennsylvania, Institute for Law & Economics Research Paper 01-22. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=294422

Aguilar, C. N., Ruiz, H. A., Rubio Rios, A., Chávez-González, M., Sepúlveda, L., Rodríguez-Jasso, R. M., Loredo-Treviño, A., Flores-Gallegos, A. C., Govea-Salas, M., & Ascacio-Valdes, J. A. (2019). Emerging strategies for the development of food industries. *Bioengineered*, *10*(1), 522–537. https://doi.org/10.1080/21655979.2019.1682109

Aguirre, A. (2020, November 11). *Why those who care about catastrophic and existential risk should care about autonomous weapons* [Online forum post]. Effective Altruism

Forum. https://forum.effectivealtruism.org/posts/oR9tLNRSAep293rr5/why-those-who-care-about-catastrophic-and-existential-risk-2

Aird, M. (2020, July 15). *Venn diagrams of existential, global, and suffering catastrophes* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/AJbZ2hHR4bmeZKznG/venn-diagrams-of-existential-global-and-suffering

Alexander, J., & Weinberg, J. M. (2014). The 'unreliability' of epistemic intuitions. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp/ 128–145). Routledge. https://www.routledge.com/Current-Controversies-in-Experimental-Philosophy/Machery-ONeill/p/book/9780415519670

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*(3), 149–155. https://doi.org/10.1007/s10676-006-0004-4

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, *12*(3), 251–261. https://doi.org/10.1080/09528130050111428

Allen, G., & Chan, T. (2017, July). *Artificial intelligence and national security*. Belfer Center for Science and International Affairs, Harvard Kennedy School. https://www.belfercenter.org/publication/artificial-intelligence-and-national-security

Alley, E. C., Turpin, M., Liu, A. B., Kulp-McDowall, T., Swett, J., Edison, R., Von Stetina, S. E., Church, G. M., & Esvelt, K. M. (2020). A machine learning toolkit for genetic engineering attribution to facilitate biosecurity. *Nature Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-19612-0

Al-Rodhan, N. (2020, June 29). *A neurophilosophy of two technological game-changers: Synthetic biology & superintelligence*. Blog of the American Philosophical Association. https://blog.apaonline.org/2020/06/29/a-neurophilosophy-of-two-technological-game-changers-synthetic-biology-superintelligence/

Althaus, D., & Baumann, T. (2020, April 29). *Reducing long-term risks from malevolent actors* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/LpkXtFXdsRd4rG8Kb/reducing-long-term-risks-from-malevolent-actors

Althaus, D. & Gloor, L. (2019) *Reducing risks of astronomical suffering: A neglected priority*. Center on Long-Term Risk. https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/

Amodei, D., & Hernandez, D. (2018, May 16). *AI and compute* [Blog post]. OpenAI. https://openai.com/blog/ai-and-compute

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv. https://arxiv.org/abs/1606.06565

Andersson, H. (2018). Application of BCA in Europe—experiences and challenges. *Journal of Benefit-Cost Analysis*, *9*(1), 84–96. https://doi.org/10.1017/bca.2018.5

Andrews, K., Comstock, G. L., Crozier G. K. D., Donaldson, S., Fenton, A., John, T. M., Johnson, L. S. M., Jones, R. C., Kymlicka, W., Meynell, L., Nobis, N., Pena-Guzman, D., & Sebo, J. (2019). *Chimpanzee rights: The philosophers' brief*. Routledge. https://www.routledge.com/Chimpanzee-Rights-The-Philosophers-Brief/Andrews-Comstock-GKD-Donaldson-Fenton-John-Johnson-Jones-Kymlicka-Meynell-Nobis-Pena-Guzman-Sebo/p/book/9781138618664

Animal Charity Evaluators (2020, November). *The Good Food Institute*. https://animalcharityevaluators.org/charity-review/the-good-food-institute/

Animal Ethics. (2020, October 25). *Introduction to the legal consideration of wild animals in the United States*. Animal Ethics. https://www.animal-ethics.org/introduction-to-the-legal-consideration-of-wild-animals-in-the-united-states/

Anleu, S. L. R. (2009). *Law and social change*. London: Sage Publications.

Anthony, G., & Neill, S. (2020 Jun. 5). The International Underwriting Association backs proposals for "Pandemic Re". *National Law Review*. https://www.natlawreview.com/article/international-underwriting-association-backs-proposals-pandemic-re

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI and Society*, *31*, 201–206. https://doi.org/10.1007/s00146-015-0590-y

Armstrong, S., & Sotala, K. (2015). How we're predicting AI—or failing to. In J. Romportl, E. Zackova, J. Kelemen (Eds.), *Beyond Artificial Intelligence* (pp. 11–29). Springer. https://doi.org/10.1007/978-3-319-09668-1_2

Arrhenius, G., Ryberg, J., & Tännsjö, T. (2017, January 16). *The repugnant conclusion*. Stanford Encyclopedia of Philosophy Archive. https://plato.stanford.edu/archives/spr2017/entries/repugnant-conclusion

Aschenbrenner, L. (2020). *Existential risk and growth* [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/Leopold-Aschenbrenner_Existential-risk-and-growth_.pdf

Askell, A., Brundage, M., & Hadfield, G. (2019). *The role of cooperation in responsible AI development*. arXiv. https://arxiv.org/abs/1907.04534

Askell, A. (2019). Evidence neutrality and the moral value of information. In H. Greaves & T. Pummer (Eds.), *Effective Altruism: Philosophical Issues* (pp. 37-52). Oxford University Press. https://doi.org/10.1093/oso/9780198841364.003.0003

Avin, S., & Amadae, S. M. (2019). Autonomy and machine learning as risk factors at the interface of nuclear weapons, computers and people. In V. Boulanin, (Ed.), *The impact of artificial intelligence on strategic stability and nuclear risk* (pp. 105–118). Stockholm International Peace Research Institute. https://doi.org/10.17863/CAM.44758

Avin, S., & Belfield, H. (2019). *Advice to EU High-Level Expert Group on artificial intelligence*. Centre for the Study of Existential Risk. https://www.cser.ac.uk/news/advice-eu-high-level-expert-group-artificial-intel/

Avin, S., Wintle, B. C., Weitzdörfer, J., ÓhÉigeartaigh, S. S., Sutherland, W. J., & Rees, M. J. (2018). Classifying global catastrophic risks. *Futures*, *102*, 20–26. https://doi.org/10.1016/j.futures.2018.02.001

Bagley, M. A., & Rai, A. K. (2013). *The Nagoya Protocol and synthetic biology research: A look at the potential impacts*. Woodrow Wilson International Center for Scholars. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=5916&context=faculty_scholarship

Baier, A. (1980). The rights of past and future generations. In E. Partridge (Ed.), *Responsibilities to future generations: environmental ethics* (pp. 171–183). http://profs-polisci.mcgill.ca/muniz/intergen/Baier.pdf

Baker, J. H., & Milsom, S. F. C. (2010). *Sources of English legal history: private law to 1750*. Oxford University Press, USA. https://global.oup.com/academic/product/baker-and-milsoms-sources-of-english-legal-history-9780199546800

Bakerlee, C., Guerra, S., Parthemore, C. Soghoian, D., & Swett, J. (2020). *Common misconceptions about biological weapons*. Council on Strategic Risks. https://councilonstrategicrisks.org/2020/12/07/briefer-common-misconceptions-about-biological-weapons/

Barnes Jr, J. H. (1984). Cognitive biases and their impact on strategic planning. *Strategic Management Journal*, *5*(2), 129–137. https://doi.org/10.1002/smj.4250050204

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B., & Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived? *Nature*, *471*(7336), 51–57. https://doi.org/10.1038/nature09678

Baron, J. (2009). *Belief overkill in political judgments*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1427862

Baron, J., & Greene, J. (1996). Determinants of insensitivity to quantity in valuation of public goods: Contribution, warm glow, budget constraints, availability, and prominence. *Journal of Experimental Psychology: Applied*, *2*(2), 107–125. https://doi.org/10.1037/1076-898X.2.2.107

Barry, C., & Tomlin, P. (2019). Moral Uncertainty and the Criminal Law. In L. Alexander & K. K. Ferzan (Eds.), *The Palgrave handbook of applied ethics and the criminal law* (pp. 445–467). New York: Palgrave. https://www.palgrave.com/gp/book/9783030228101

Baum, S. D. (2016). On the promotion of safe and socially beneficial artificial intelligence. *AI and Society, 32*(4), 543–551. https://doi.org/10.1007/s00146-016-0677-0

Baum, S. D. (2017). Social choice ethics in artificial intelligence. *AI and Society*, *35*, 165–176. https://doi.org/10.1007/s00146-017-0760-1

Baum, S. D. (2018a). Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI and Society*, *33*(4), 565-572. https://doi.org/10.1007/s00146-017-0734-3

Baum, S. D. (2018b). Superintelligence skepticism as a political tool. *Information, 9*(9), Article 209. https://doi.org/10.3390/info9090209

Baum, S. D. (2020a). Medium-term artificial intelligence and society. *Information*, *11*(6), 290. https://doi.org/10.3390/info11060290

Baum, S. D. (2020b). Quantifying the probability of existential catastrophe: A reply to Beard et al. *Futures*, *123*, Article 102608. https://doi.org/10.1016/j.futures.2020.102608

Baum, S. D., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., … Sotala, K. (2019). Long-term trajectories of human civilization. *Foresight*, *21*(1), 53–83. https://dx.doi.org/10.1108/FS-04-2018-0037

Baum, S. D., Barrett, A. M., & Yampolskiy, R. V. (2017). Modeling and interpreting expert disagreement about artificial superintelligence, *Informatica, 41*(4), 419–427. http://www.si/index.php/informatica/article/view/1812

Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an expert assessment. *Technological Forecasting and Social Change, 78(1)*, 185–195. https://doi.org/10.1016/j.techfore.2010.09.006

Baumann, T. (2017a, September 16). *Focus areas of worst-case AI safety*. Reducing Risks of Future Suffering. https://s-risks.org/focus-areas-of-worst-case-ai-safety

Baumann, T. (2017b, December 15). *Using surrogate goals to deflect threats*. Reducing Risks of Future Suffering. https://s-risks.org/using-surrogate-goals-to-deflect-threats

Baumann, T. (2018a, July 5). *An introduction to worst-case AI safety*. Reducing Risks of Future Suffering. https://s-risks.org/an-introduction-to-worst-case-ai-safety

Baumann, T. (2018b). *A typology of s-risks*. Center for Reducing Suffering. http://centerforreducingsuffering.org/a-typology-of-s-risks

Baumann, T. (2019). *Risk factors for s-risks*. Center for Reducing Suffering. https://centerforreducingsuffering.org/risk-factors-for-s-risks/

Baumann, T. (2020). *Space governance is important, tractable, and neglected* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/QkRq6aRA84vv4xsu9/space-governance-is-important-tractable-and-neglected

Beard, S., Rowe, T., & Fox, J. (2020a). An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards. *Futures*, *115*, Article 102469. https://doi.org/10.1016/j.futures.2019.102469

Beard, S., Rowe, T., & Fox, J. (2020b). Existential risk assessment: A reply to Baum. *Futures*, *122*, 102606. https://doi.org/10.1016/j.futures.2020.102606

Beard, S. et al. (2017). *Written evidence to lords select committee on artificial intelligence* [Technical report]. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/resources/written-evidence-lords-select-committee-artificial-intelligence/

Becker, U., Müller, H., & Wunderlich, C. (2005). While waiting for the protocol. *The Nonproliferation Review*, *12*(3), 541–572. https://doi.org/10.1080/10736700600601194

Beckstead, N. (2013a). *On the overwhelming importance of shaping the far future* [Doctoral dissertation]. Rutgers University-Graduate School-New Brunswick. https://rucore.libraries.rutgers.edu/rutgers-lib/40469/

Beckstead, N. (2013b, May 27). *A proposed adjustment to the astronomical waste argument* [Online forum post]. LessWrong. https://www.lesswrong.com/posts/5czcpvqZ4RH7orcAa/a-proposed-adjustment-to-the-astronomical-waste-argument#Broad_and_narrow_strategies_for_shaping_the_far_future

Beckstead, N. (2013c). *How to compare broad and targeted attempts to shape the*
far future [Presentation]. Future of Humanity Institute, University of Oxford. http://intelligence.org/wp-content/uploads/2013/07/Beckstead-Evaluating-Options-Using-Far-Future-Standards.pdf

Beckstead, N. (2014, June). *Will we eventually be able to colonize other stars? Notes from a preliminary review*. Oxford University, Future of Humanity Institute. https://www.fhi.ox.ac.uk/will-we-eventually-be-able-to-colonize-other-stars-notes-from-a-preliminary-review/

Beckstead, N. (2019). A brief argument for the overwhelming importance of shaping the far future. In H. Greaves & T. Pummer, *Effective Altruism: Philosophical Issues*. Oxford University Press. http://doi.org/10.1093/oso/9780198841364.001.0001

Beckstead, N., & Thomas, T. (2020, September). *A paradox for tiny probabilities and enormous values*. Global Priorities Institute Working Paper Series. GPI Working Paper. https://globalprioritiesinstitute.org/wp-content/uploads/Nick-Beckstead-and-Teruji-Thomas_A-paradox-for-tiny-probabilities-and-enormous-values.pdf

Beeckman, D. S. A., & Rüdelsheim, P. (2020). Biosafety and biosecurity in containment: A regulatory overview. *Frontiers in Bioengineering and Biotechnology*, *8*(650), 1–7. https://doi.org/10.3389/fbioe.2020.00650

Belfield, H. (2019). How to respond to the potential malicious uses of artificial intelligence? *Journal of Unsolved Questions, 9*(2), 5–6. http://junq.info/wp-content/uploads/2019/09/JUNQ.pdf

Belfield, H. (2020a). Activism by the AI community: Analysing recent achievements and future prospects. In *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 15–21. https://doi.org/10.1145/3375627.3375814

Belfield, H. (2020b). *From tech giants to a tech colossus: Antitrust objections to the Windfall Clause* [Manuscript submitted for publication].

Belfield, H., & ÓhÉigeartaigh, S. (2017). *Long-term catastrophic risk from artificial Intelligence*. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/75539.html

Belfield, H., Hernández-Orallo, J., ÓhÉigeartaigh, S., Maas, M. M., Hagerty, A., & Whittlestone, J. (2020). *Consultation on the White Paper on AI—a European approach*. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/media/uploads/files/Consultation_Response_White_Paper_on_AI_-_Belfield_Hern%C3%A1ndez-Orallo_%C3%93_h%C3%89igeartaigh_Maas_Hagerty_Whittlestone.pdf

Belfield, H., Jayanti, A., & Avin, S. (2020, May 19). *Written evidence—Defence industrial policy: procurement and prosperity*. Centre for the Study of Existential Risk, University of Cambridge. https:// www.cser.ac.uk/resources/written-evidence-defence-industrial-policy-procurement-and-prosperity/

Bell, J. (2012). Path dependence and legal development. *Tulane Law Review*, *87*, 787–810.

Ben-Haim, Y. (2006). *Info-gap decision theory: Decisions under severe uncertainty*. Elsevier.

Benner, S. A., & Sismour, A. M. (2005). Synthetic biology. *Nature Reviews Genetics*, *6*(7), 533–543. https://dx.doi.org/10.1038%2Fnrg1637

Benner, S. A., Yang, Z., & Chen, F. (2011). Synthetic biology, tinkering biology, and artificial biology. What are we learning? *Comptes Rendus Chimie*, *14*(4), 372–387. https://doi.org/10.1016/j.crci.2010.06.013

Bennett Moses, L. (2020). Not a single singularity. In S. Deakin and C. Markou (Eds.), *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence* (pp. 20–36). Hart Publishing. https://www.bloomsburyprofessional.com/uk/is-law-computable-9781509937066/

Benson, J. F., Mahoney, P. J., Sikich, J. A., Serieys, L. E., Pollinger, J. P., Ernest, H. B., & Riley, S. P. (2016). Interactions between demography, genetics, and landscape connectivity increase extinction probability for a small population of large carnivores in a major metropolitan area. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1837), 20160957. https://doi.org/10.1098/rspb.2016.0957

Bentham, J. (1823). *An introduction to the principles of morals and legislation*. Oxford: Clarendon Press. https://www.econlib.org/library/Bentham/bnthPML.html

Berger, A. (2014, June 26). *Potential global catastrophic risk focus areas*. Open Philanthropy. https://www.openphilanthropy.org/blog/potential-global-catastrophic-risk-focus-areas

Berman, H. J. (1985). *Law and revolution: The formation of the western legal tradition.* Cambridge, Massachusetts: Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674517769

Bernstein, G. (2006). When new technologies are still new: Windows of opportunity for privacy protection. *Villanova Law Review*, 51, 921. https://digitalcommons.law.villanova.edu/vlr/vol51/iss4/8

Bidwell, C. A., & Bhatt, K. (2016, February). *Use of attribution and forensic science in addressing biological weapon threats: A multi-faceted study*. Federation of American Scientists. https://fas.org/pub-reports/biological-weapons-and-forensic-science/

Blattner, C. E. (2019a). The recognition of animal sentience by the law. *Journal of Animal Ethics*, *9*(2), 121–136. https://doi.org/10.5406/janimalethics.9.2.0121

Blattner, C. E. (2019b). *Protecting animals within and across borders: Extraterritorial jurisdiction and the challenges of globalization.* Oxford University Press. https://global.oup.com/academic/product/protecting-animals-within-and-across-borders-9780190948313

Bilz, K., & Nadler, J. (2014). Law, moral attitudes, and behavioral change. In E. Zamir & D. Teichman (Eds.), *The Oxford handbook of behavioral economics and the law* (pp. 241–267). Oxford University Press. https://www.law.northwestern.edu/faculty/fulltime/nadler/Bilz-Nadler-LawMoralAttitudesPageProofs.pdf

Binder, D. (2018). *The findings of an empirical study of the application of criminal law in non-terrorist disasters and tragedies*, *Futures*, *102*, 134–145. https://doi.org/10.1016/j.futures.2018.01.008

Bittencourt Neto, O., Hofmann, M., Masson-Zwaan, T., & Stefoudi, D. (2020). *Building blocks for the development of an international framework for the governance of space resources activities: A commentary*. Eleven International Publishing. https://www.boomdenhaag.nl/en/webshop/building-blocks-for-the-development-of-an-international-framework-for-the-governance-of-space-resource-activities

Bo, M. (2020, December 18). Meaningful human control over autonomous weapon systems: An (international) criminal law account. *Opinio Juris*. http://opiniojuris.org/2020/12/18/meaningful-human-control-over-autonomous-weapon-systems-an-international-criminal-law-account

Book, A., Visser, B. A., & Volk, A. A. (2015). Unpacking "evil": Claiming the core of the dark triad. *Personality and Individual Differences*, *73*, 29–38. https://doi.org/10.1016/j.paid.2014.09.016

Bostrom, N. (1998). *How long before superintelligence?* Future of Humanity Institute, University of Oxford. https://www.nickbostrom.com/superintelligence.html

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, *9*(1). https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c

Bostrom, N. (2003a). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, *15*(3), 308–314. https://www.nickbostrom.com/astronomical/waste.pdf

Bostrom, N. (2003b). *Ethical issues in advanced artificial intelligence.* Future of Humanity Institute, University of Oxford. https://nickbostrom.com/ethics/ai.html

Bostrom, N. (2009). Pascal's mugging. *Analysis*, *69*(3), 443–445. https://www.jstor.org/stable/40607655

Bostrom, N. (2011a). Infinite ethics. *Analysis and Metaphysics*, *10*, 9–59.

Bostrom, N. (2011b). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, *10*, 44–79. https://nickbostrom.com/information-hazards.pdf

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. https://global.oup.com/academic/product/superintelligence-9780199678112

Bostrom, N. (2017). Strategic implications of openness in AI development. *Global Policy, 8(2)*, 135–148. https://doi.org/10.1111/1758-5899.12403

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455–476. https://doi.org/10.1111/1758-5899.12718

Boulanin, V., Saalman, L., Topychkanov, P., Su, F., & Carlsson, M. P. (2020). *Artificial intelligence, strategic stability and nuclear risk*. Stockholm International Peace Research Institute. https://www.sipri.org/publications/2020/other-publications/artificial-intelligence-strategic-stability-and-nuclear-risk

Bourget, D., & Chalmers, D. J. (2013). What do philosophers believe? *Philosophical Studies*, *170*(3), 465–500. https://doi.org/10.1007/s11098-013-0259-7

Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. Oxford University Press. https://global.oup.com/academic/product/the-brussels-effect-9780190088583

Braithwaite, V. (2010). *Do fish feel pain?* Oxford University Press.

Brand, S. (2000, April 26). *Taking the long view*. Time. http://content.time.com/time/magazine/article/0,9171,996757,00.html

Bressler, D., & Bakerlee, C. (2018, December 6). *"Designer bugs": How the next pandemic might come from a lab*. Vox. https://www.vox.com/future-perfect/2018/12/6/18127430/superbugs-biotech-pathogens-biorisk-pandemic

Briggs, H. (2020, November 13). *Mutated coronavirus may 'jump back and forth' between animals*. BBC News. https://www.bbc.com/news/science-environment-54918267

Briggs, R. A. (2019). *Normative theories of rational choice: Expected utility*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2019/entries/rationality-normative-utility/

Brockman, G., Sutskever, I., & OpenAI (2019, March 11). *OpenAI LP* [Press release]. https://openai.com/blog/openai-lp

Bronsteen, J., Buccafusco, C., & Masur, J. (2013). Well-being analysis vs. cost-benefit analysis. *Duke Law Journal*, *62*, 1603–1689. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=3389&context=dlj

Broome, J. (2008). The ethics of climate change. *Scientific American*, *298*(6), 96–102. https://www.jstor.org/stable/26000646

Broome, J., & Foley, D. (2016). A world climate bank. In A. Gosseries & I. González-Ricoy (Eds.), *Institutions for Future Generations* (pp. 156–169). Oxford University Press. https://global.oup.com/academic/product/institutions-for-future-generations-9780198746959

Brundage, M. (2015). Taking superintelligence seriously: Superintelligence: Paths, dangers, strategies by Nick Bostrom (Oxford University Press, 2014). *Futures*, *72*, 32–35. https://doi.org/10.1016/j.futures.2015.07.009

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., ÓhÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., . . . Amodei, D. (2018). *Malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv. https://arxiv.org/abs/1802.07228

Bruns, R. (2019). *Finance in a pandemic*. Event 201: A Global Pandemic Exercise. https://www.centerforhealthsecurity.org/event201/event201-resources/finance-fact-sheet-191009.pdf

Bublitz, J. C. (2014). Freedom of thought in the age of neuroscience. *Archiv für Rechts-und Sozialphilosphie*, *100*(1), 1–25. https://www.researchgate.net/publication/261950057_Freedom_of_Thought_in_the_Age_of_Neuroscience

Bublitz, J. C. (2015). Cognitive liberty and the international right to freedom of thought. In J. Clausen & N. Levy (Eds.), *Springer Handbook of Neuroethics*. Springer. https://doi.org/10.1007/978-94-007-4707-4_166

Bublitz, J. C., & Merkel, R. (2014). Crimes against minds: on mental manipulations, harms and a human right to mental self-determination. *Criminal Law and Philosophy*, *8*(1), 51–77. https://doi.org/10.1007/s11572-012-9172-y

Buchholz, W., & Schumacher, J. (2010). Discounting and welfare analysis over time: Choosing the η. *European Journal of Political Economy*, *26*(3), 372–385. https://doi.org/10.1016/j.ejpoleco.2009.11.011

Burden, J., & Hernández-Orallo, J. (2020). Exploring AI Safety in Degrees: Generality, Capability and Control. In *AAAI workshop on artificial intelligence safety (SafeAI 2019).* (pp. 36–40). CEUR Workshop Proceedings.

Caldeira, K., & Kasting, J. F. (1992). The life span of the biosphere revisited. *Nature*, *360*(6406), 721–723. https://doi.org/10.1038/360721a0

Calma, J. (2020, April 10). *To prevent the next pandemic, scientists search for animal zero*. The Verge. https://www.theverge.com/2020/4/10/21216165/pandemic-prevention-sciencetist-animal-human-health-disease

Calo, R. (2014, September 15). *The case for a federal robotics commission*. Brookings Institution. https://www.brookings.edu/research/the-case-for-a-federal-robotics-commission

Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. *UC Davis Law Review, 51*(2), 399–435.

Calvin, N., & Leung, J. (2020). *Who owns artificial intelligence? A preliminary analysis of corporate intellectual property strategies and why they matter* (Working paper). Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-working-paper-Who-owns-AI-Apr2020.pdf

Calvo, R. A., Peters, D., & Cave, S. (2020). Advancing impact assessment for intelligent systems. *Nature Machine Intelligence*, *2*, 89–91. https://doi.org/10.1038/s42256-020-0151-z

Cameron, E. E., Carter, S. R., Jordan, J. L., World Economic Forum, & Morhard, R. (2020, January). *Biosecurity innovation and risk reduction: A global framework for accessible,* *safe and secure DNA synthesis*. Nuclear Threat Initiative & World Economic Forum. https://media.nti.org/documents/Biosecurity_Innovation_and_Risk_Reduction.pdf

Caplan, B. (2011). The totalitarian threat. In N. Bostrom & M. Cirkovic (Eds.). *Global catastrophic risks* (pp. 504–519). Oxford University Press.

Carey, M. P. (2014, December 9). *Cost-benefit and other analysis requirements in the rulemaking process*. Congressional Research Service. https://fas.org/sgp/crs/misc/R41974.pdf

Carlier, A., Clarke, S., & Schuett, J. (2020). *AI risk survey* [Manuscript in preparation].

Carlier, A., & Davidson, T. (2020, February 8). *What can the principal-agent literature tell us about AI risk?* [Online forum post]. AI Alignment forum. https://www.alignmentforum.org/posts/Z5ZBPEgufmDsm7LAv/what-can-the-principal-agent-literature-tell-us-about-ai

Carter, S. R., & Friedman, R. M. (2015, October). *DNA synthesis and biosecurity: Lessons learned and options for the future*. J. Craig Venter Institute. https://www.jcvi.org/research/dna-synthesis-and-biosecurity-lessons-learned-and-options-future

Casadevall, A., & Relman, D. A. (2010). Microbial threat lists: obstacles in the quest for biosecurity? *Nature Reviews Microbiology*, *8*(2), 149–154. https://doi.org/10.1038/nrmicro2299

Casadevall, A., Enquist, L., Imperiale, M. J., Keim, P., Osterholm, M. T., & Relman, D. A. (2013). Redaction of sensitive data in the publication of dual use research of concern. *mBio*, *5*(1), e00991–13. https://doi.org/10.1128/mBio.00991-13

Casey, P., Burke, K., & Leben, S. (2013). Minding the court: Enhancing the decision-making process. *International Journal for Court Administration*, *5*(1), 45–54. http://doi.org/10.18352/ijca.8

Castel, J.-G., & Castel, M. E. (2016). The road to artificial superintelligence: Has international law a role to play? *Canadian Journal of Law and Technology*, *14*(1), 1–15. https://ojs.library.dal.ca/CJLT/article/view/7211

Cave, S. (2017). *AI: Ethics and Governance: The Issues*. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69702.html

Cave, S., & ÓhÉigeartaigh, S. S. (2019). Bridging near-and long-term concerns about AI. *Nature Machine Intelligence*, *1*(1), 5–6. https://doi.org/10.1038/s42256-018-0003-2

Caviola, L. (2019). *How we value animals: the psychology of speciesism* [Doctoral dissertation]. University of Oxford. https://ora.ox.ac.uk/objects/uuid:9000a8be-b6dc-4c63-88e8-974b9daaa83a

Caviola, L., Everett, J. A., & Faber, N. S. (2019). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology*, *116*(6), 1011–1029. https://doi.org/10.1037/pspp0000182

Center for Effective Public Policy (2017, June). *A framework for evidence-based decision making in state and local criminal justice systems*. Center for Effective Public Policy. https://cepp.com/wp-content/uploads/2018/10/A-Framework-for-Evidence-Based-Decision-Making-in-State-and-Local-Criminal-Justice-Systems.pdf

Centers for Disease Control and Prevention (2020). *HHS and USDA Select Agents and Toxins*. U.S. Department of Health and Human Services. https://www.selectagents.gov/sat/list.htm

Center on Long-Term Risk (2020). *Priority Areas*. https://longtermrisk.org/priority-areas/

Centre for the Governance of AI. (2020). *Consultation on the European Commission's white paper on artificial intelligence: a European approach to excellence and trust*. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/EU-White-Paper-Consultation-Submission-GovAI-Oxford.pdf

Centre for the Study of Existential Risk (2020). *Global catastrophic biological risks*. University of Cambridge. https://www.cser.ac.uk/research/global-catastrophic-biological-risks

Cerqueira, M., & Billington, T. (2020). *Fish welfare improvements in aquaculture*. Fish Welfare Initiative. https://www.fishwelfareinitiative.org/fish-welfare-improvements

Chapman, C. R., Sukumaran, S., Tsegaye, G. T., Shevchenko, Y., & Caplan, A. L. (2019). The quest for compensation for research-related injury in the United States: A new proposal. *Journal of Law, Medicine and Ethics*, *47*(4), 732–747.

Cheng, L., Rosett, A., & Woo, M. (Eds.). (2003). *East Asian law: Universal norms and local cultures*. Routledge.

Chernov, D., & Sornette, D. (2016). *Man-made catastrophes and risk information concealment*. Springer.

Chessman, C. F. (2018). *Not quite human: Artificial intelligence, animals, and the regulation of sentient property*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3200802

Chesterman, S. (2020). Artificial intelligence and the limits of legal personality. *International & Comparative Law Quarterly, 69*(4), 819–844. https://doi.org/10.1017/S0020589320000366

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W. W. Norton & Company.

Christiano, P. (2018a, February 24). Takeoff speeds. *The sideways view*. https://sideways-view.com/2018/02/24/takeoff-speeds

Christiano, P. (2018b, April 7). Clarifying "AI alignment". *AI Alignment*. https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6

Christiano, P. (2019, March 17). *What failure looks like* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like

Christiano, P., Shlegeris, B., & Amodei, D. (2018). *Supervising strong learners by amplifying weak experts*. arXiv. https://arxiv.org/abs/1810.08575

Cihon, P. (2019, April). *Standards for AI governance: International standards to enable global cooperation in AI research & development* [Technical report]. Center for the Governance of AI, Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., & Baum, S. D. (2020). *AI certification: Advancing ethical practice by reducing information asymmetries* [Manuscript submitted for publication].

Cihon, P., Schuett, J., & Baum, S. D. (2020). *Corporate governance of artificial intelligence in the public interest* [Manuscript submitted for publication].

Cihon, P., Maas, M. M., & Kemp, L. (2020a). Fragmentation and the future: Investigating architectures for international AI governance. *Global Policy, 11*(5). https://doi.org/10.1111/1758-5899.12890

Cihon, P., Maas, M. M., & Kemp, L. (2020b). Should artificial intelligence governance be centralised? Design lessons from history. *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 228–234. https://doi.org/10.1145/3375627.3375857

Clark, J., & Amodei, D. (2016, December 21). *Faulty reward functions in the wild.* Open AI. https://openai.com/blog/faulty-reward-functions/

Clark, J., & Hadfield, G. K. (2018). *Regulatory markets for AI safety*. arXiv. https://arxiv.org/abs/2001.00078

Clifton, J. (2020). *Cooperation, conflict, and transformative artificial intelligence: A research agenda.* Center on Long-Term Risk. https://longtermrisk.org/research-agenda

Cline, W. R. (1992). *The economics of global warming.* Institute for International Economics.

Cochrane, A. (2018). *Sentientist politics: A theory of global inter-species justice.* Oxford University Press.

Cockell, C. S. (Ed.). (2016). *Dissent, revolution and liberty beyond earth.* Springer. https://doi.org/10.1007/978-3-319-29349-3

Coglianese, C., & Ben Dor, L. (2020). AI in adjudication and administration [Forthcoming]. *Brooklyn Law Review.* https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=3120&context=faculty_scholarship

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, *85*(5), 808–822. https://doi.org/10.1037/0022-3514.85.5.808

Conger, K., Fausset, R. & Kovaleski, S. F. (2019, May 14). *San Francisco bans facial recognition technology. New York Times*. https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html

Cotra, A. (2018, March). *Iterated distillation and amplification* [Online forum post]. AI Alignment Forum. https://ai-alignment.com/iterated-distillation-and-amplification-157debfd1616

Cottier, B., & Shah, R. (2019, August 15). *Clarifying some key hypotheses in AI alignment* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/mJ5oNYnkYrd4sD5uE/clarifying-some-key-hypotheses-in-ai-alignment

Cotton-Barratt, C. (2014, October 1). *Effective policy? Requiring liability insurance for dual-use research* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/zvRerivrWdZ5J5rD9/effective-policy-requiring-liability-insurance-for-dual-use

Cotton-Barratt, O., Daniel, M., & Sandberg, A. (2020). Defence in depth against human extinction: Prevention, response, resilience, and why they all matter. *Global Policy*, *11*(3), 271–282. https://doi.org/10.1111/1758-5899.12786

Cotton-Barratt, O., Farquhar, S., Halstead, J., Schubert, S., & Snyder-Beattie, A. (2016). *Global catastrophic risks 2016.* Global Challenges Foundation.

https://globalchallenges.org/wp-content/uploads/2019/07/Global-Catastrophic-Risk-Annual-Report-2016.pdf

Cotton-Barratt, O., & Ord, T. (2015). *Existential risk and existential hope: Definitions* [Technical report #2015-1]. Future of Humanity Institute, University of Oxford. http://amirrorclear.net/files/existential-risk-and-existential-hope.pdf

Cowen, T., & Parfit, D. (1992). Against the social discount rate. In P. Laslett & J. S. Fishkin (Eds.), *Philosophy, Politics, and Society: Volume 6, Justice Between Age Groups and Generations* (pp. 144–161). New Haven and London: Yale University Press.

Crane, A. T., Voth, J. P., Shen, F. X., & Low, W. C. (2019). Concise review: Human-animal neurological chimeras: Humanized animals or human cells in an animal? *Stem Cells, 37*(4), 444–452. https://doi.org/10.1002/stem.2971

Crawford, M., Adamson, F., & Ladish, J. (2019, September 16). *Bioinfohazards* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/ixeo9swGQTbYtLhji/bioinfohazards-1

Cremer, C. Z., & Whittlestone, J. (2020). Canaries in technology mines: Warning signs of discontinuous progress in AI. *Evaluating Progress in AI Workshop—ECAI 2020*. http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_4.pdf

Crootof, R. (2019). "Cyborg justice" and the risk of technological-legal lock-in. *Columbia Law Review*, *119*(7), 233–251. https://columbialawreview.org/content/cyborg-justice-and-the-risk-of-technological-legal-lock-in

Croxton, D. (1999). The Peace of Westphalia of 1648 and the Origins of Sovereignty. *The International History Review*, *21*(3), 569–591. https://doi.org/10.1080/07075332.1999.9640869

Dafoe, A. (2018). *AI governance: A research agenda*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https:// www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf

Dai, W. (2018, December 16). *Two neglected problems in human-AI safety* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/HTgakSs6JpnogD6c2/two-neglected-problems-in-human-ai-safety

Dai, W. (2019, February 10). *The argument from philosophical difficulty* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/w6d7XBCegc96kz4n3/the-argument-from-philosophical-difficulty

Dainow, J. (1966). The civil law and the common law: some points of comparison. *American Journal of Comparative Law*, *15*(3), 419–435.

Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics, 26*, 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Dando, M., Evans, N., Lentzos, F., Revill, J., & Sims, N. (2018, December 5). *Joint NGO statement to Biological Weapons Convention Meeting of States Parties*. https://media.nti.org/documents/MSP_2018_joint_NGO_statement_FINAL.pdf

Daniel, M. (2017). *S-risks: Why they are the worst existential risks, and how to prevent them (EAG Boston 2017)*. Center on Long-Term Risk. https://longtermrisk.org/s-risks-talk-eag-boston-2017/

Das, T. K., & Teng, B. S. (1999). Cognitive biases and strategic decision processes: An integrative perspective. *Journal of Management Studies*, *36*(6), 757–778. https://doi.org/10.1111/1467-6486.00157

Dasgupta, P. (2008). Discounting climate change. *Journal of Risk and Uncertainty*, 37(2–3), 141–169. https://doi.org/10.1007/s11166-008-9049-6

Davis, G., & Marcus, E. (2019). *Rebooting AI: Building artificial intelligence we can trust.* Pantheon.

Deakin, S., & Markou, C. (Eds.) (2020). *Is law computable?* Hart Publishing.

Deckha, M. (2020). *Animals as legal beings: Contesting anthropocentric legal orders*. University of Toronto Press.

DeLisi, C. (2019). The role of synthetic biology in climate change mitigation. *Biology Direct*, 14(1), 1-5. https://doi.org/10.1186/s13062-019-0247-8

de Lazari-Radek, K., & Singer, P. (2014). The point of view of the universe: Sidgwick and contemporary ethics. Oxford University Press.

De Man, P. (2016). *Exclusive use in an inclusive environment: The meaning of the non-appropriation principle for space resource exploitation (Vol. 9).* Springer.

Devezas, T., de Melo, F. C. L., Gregori, M. L., Salgado, M. C. V., Ribeiro, J. R., & Devezas, C. B. (2012). The struggle for space: Past and future of the space race. *Technological Forecasting and Social Change*, *79*(5), 963–985. https://doi.org/10.1016/j.techfore.2011.12.006

Dickens, M. (2016, October). *Evaluation Frameworks (or: When Importance / Neglectedness / Tractability Doesn't Apply)* [Blog post]. Philosophical Multicore. https://mdickens.me/2016/06/10/evaluation_frameworks_(or-_when_scale-neglectedness-tractability_doesn't_apply)/

Dickens, M. (2020, September). *"Disappointing futures" might be as important as existential risks* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/9AYmbh25eKLojeQGe/disappointing-futures-might-be-as-important-as-existential

Dickert, S., Västfjäll, D., Kleber, J., & Slovic, P. (2015). Scope insensitivity: The limits of intuitive valuation of human lives in public policy. *Journal of Applied Research in Memory and Cognition*, *4*(3), 248–255. https://doi.org/10.1016/j.jarmac.2014.09.002

Dietz, S., Hepburn, C, & Stern, N. (2018). Economics, ethics and climate change. In K. Basu & R. Kanbur (Eds.), *Arguments for a better world: Essays in honor of Amartya Sen.* Oxford University Press.

DiEuliis D., Carter S. R., & Gronvall G. K. (2017, August 23). Options for synthetic DNA order screening, revisited. *mSphere*, *2*(4), e00319–17. https://doi.org/10.1128/msphere.00319-17

Donaldson, S., & Kymlicka, W. (2011). *Zoopolis: A political theory of animal rights*. Oxford University Press.

Doremus, H. (2000). The rhetoric and reality of nature protection: Toward a new discourse. *Washington and Lee Law Review*, 57, 11–73.

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. https://arxiv.org/abs/1702.08608

Douglas, C. M., & Stemerding, D. (2014). Challenges for the European governance of synthetic biology for human health. *Life Sciences, Society and Policy*, *10*(6), 1–18. https://doi.org/10.1186/s40504-014-0006-7

Drexler, K. E. (2019). *Reframing superintelligence: Comprehensive AI services as general intelligence* [Technical report]. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf

Dror, Y. (1958). Law and social change. *Tulane Law Review*, *33*, 787–802.

Drube, L. et al. (2020). *Planetary defence: Legal overview and assessment*. Space Mission Planning Advisory Group. https://www.cosmos.esa.int/documents/336356/336472/SMPAG-RP-004_1_0_SMPAG_legal_report_2020-04-08+%281%29.pdf/60df8a3a-b081-4533-6008-5b6da5ee2a98?t=1586443949723

Drupp, M. A., Freeman, M., Groom, B., & Nesje, F. (2018). Discounting disentangled. *American Economic Journal: Economic Policy*, *10*(4), 109–34. https://doi.org/10.1257/pol.20160240

DuBois, K. (2011). The unknowing volunteers: Population testing in the United States. *Journal of Biosecurity, Biosafety and Biodefense Law*, *1*(1), 1–17. https://doi.org/10.2202/2154-3186.1004

Dubov, A. (2014). The concept of governance in dual-use research. *Medicine, Health Care and Philosophy*, *17*, 447–457. https://doi.org/10.1007/s11019-013-9542-9

Duda, R. (2016, April). *Factory farming*. 80,000 Hours. https://80000hours.org/problem-profiles/factory-farming/

Duda, R., & Koehler, A. (2016, April). *Climate change (extreme risks)*. 80,000 Hours. https://80000hours.org/problem-profiles/climate-change/

Duff, R. A., & Marshall, S. E. (2015). 'Abstract endangerment,' two harm principles, and two routes to criminalization, *Bergen Journal of Criminal Law and Criminal Justice*, *3*, 132–161.

Eckersley, P., & Sandberg, A. (2013). Is brain emulation dangerous? *Journal of Artificial General Intelligence*, *4*(3), 170–194.

Eckerström Liedholm, S. (2019, December 30). *Long-term design considerations for wild animal welfare interventions*. Wild Animal Initiative. https://www.wildanimalinitiative.org/blog/persistenceandreversibility

Einstein, A. (1948). A reply to the Soviet scientists. *Bulletin of the Atomic Scientists*, *4*(2), 35–38. https://doi.org/10.1080/00963402.1948.11460161

Eisenberg, T. (2011). The origins, nature, and promise of empirical legal studies and a response to concerns. *University of Illinois Law Review*, 1713–1738. https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=1760&context=facpub

Eisner, M. (2003). Long-term historical trends in violent crime. *Crime and Justice*, *30*, 83–142.

Elkins, Z., Ginsburg, T., & Melton, J. (2007, July). The lifespan of written constitutions. In *American Political Science Association Meeting, Chicago*. https://www.researchgate.net/publication/228813917_The_Lifespan_of_Written_Constitutions

Emanuel, P., Walper, S., DiEuliis, D., Klein, N., Petro, J. B., & Giordano, J. (2019, October). *Cyborg soldier 2050: Human/machine fusion and the implications for the future of the DOD*. U.S. Army Combat Capabilities Development Command, Chemical Biological

Center. https://community.apan.org/wg/tradoc-g2/mad-scientist/m/articles-of-interest/300458

Endy, D. (2005). Foundations for engineering biology. *Nature*, *438*(7067), 449–453. https://doi.org/10.1038/nature04342

Enemark, C. (2017). *Biosecurity dilemmas*. Washington, DC: Georgetown University Press. http://press.georgetown.edu/book/georgetown/biosecurity-dilemmas

Engineering Biology Research Consortium. (2020). *What is synthetic/engineering biology?* https://ebrc.org/what-is-synbio/

Enos, R. D., Fowler, A., & Havasy, C. S. (2017). The negative effect fallacy: A case study of incorrect statistical reasoning by federal courts. *Journal of Empirical Legal Studies*, *14*(3), 618–647. https://doi.org/10.1111/jels.12158

Erdélyi, O. J., & Goldsmith, J. A. (2018, February 2–3). Regulating artificial intelligence: Proposal for a global solution [Paper presentation]. *2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101. https://doi.org/10.1145/3278721.3278731

Etzioni, O. (2020, February 25). *How to know if artificial intelligence is about to destroy civilization*. MIT Technology Review. https://www.technologyreview.com/2020/02/25/906083/artificial-intelligence-destroy-civilization-canaries-robot-overlords-take-over-world-ai/

European Agency for Safety and Health at Work (2020). *Directive 2000/54/EC - Biological agents at work*. https://osha.europa.eu/en/legislation/directives/exposure-to-biological-agents/77

European Commision (2020). *On artificial intelligence - A European approach to excellence and trust* [White paper]. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission (2014). *Final opinion on synthetic biology*. https://ec.europa.eu/health/scientific_committees/consultations/public_consultations/scenihr_consultation_21_en

European Commission (2018). *Communication from the commission to the European parliament, the European council, the council, the European economic and social committee and the committee of the regions: Artificial intelligence for Europe*. https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF

Evans, N. G. (2014). Dual-use decision making: relational and positional issues. *Monash Bioethics Review*, *32*(3–4), 268–283. https://dx.doi.org/10.1007%2Fs40592-015-0026-y

Evans, N. G., & Commins, A. (2017, February 3). *Defining dual-use research: When scientific advances can both help and hurt humanity*. The Conversation. https://theconversation.com/defining-dual-use-research-when-scientific-advances-can-both-help-and-hurt-humanity-70333

Everitt, T., Ortega, P. A., Barnes, E., & Legg, S. (2019). *Understanding agent incentives using causal influence diagrams. Part I: Single action settings*. arXiv. https://arxiv.org/abs/1902.09980v6

Faigman, D. L. (1989). To have and have not: Assessing the value of social science to the law as science and policy. *Emory Law Journal*, *38*, 1005–1096. https://repository.uchastings.edu/faculty_scholarship/140/

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714–3719. https://doi.org/10.1073/pnas.1618569114

Farquhar, S., Cotton-Barratt, O., & Snyder-Beattie, A. (2019). Pricing externalities to balance public risks and benefits of research. *Health Security, 15*(4), 401–408. https://doi.org/10.1089/hs.2016.0118

Farquhar, S., Halstead, J., Cotton-Barratt, O., Schubert, S., Belfield, H., & Snyder-Beattie, A. (2017). *Existential risk: diplomacy and governance*. Global Priorities Project. https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf

Fasan, E. (1990). Discovery of ETI: Terrestrial and extraterrestrial legal implications. *Acta Astronautica*, *21*(2), 131–135. https://doi.org/10.1016/0094-5765(90)90140-G

Federal Government of Germany. (2019). *Germany: Artificial intelligence strategy* [Report of the European Commission]. https://knowledge4policy.ec.europa.eu/publication/germany-artificial-intelligence-strategy_en

Feinberg, J. (1974). The rights of animals and future generations. In W. Blackstone (Ed.), *Philosophy and environmental crisis* (pp. 43–68). Athens, Georgia: University of Georgia Press. http://www.animal-rights-library.com/texts-m/feinberg01.pdf

Feldman, F. (2006). Actual utility, the objection from impracticality, and the move to expected utility. *Philosophical Studies*, *129*(1), 49–79. https://www.jstor.org/stable/4321749

Feldman, N., Fischer, S.-C., Hadfield, G. GovAI Webinar #3 [Webinar]. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/webinar3/

Ferguson, A. G. (2019). Facial recognition and the fourth amendment [Forthcoming]. *Minnesota Law Review*, *105*. http://dx.doi.org/10.2139/ssrn.3473423

Floridi, L. (2020). Artificial intelligence as a public service: Learning from Amsterdam and Helsinki. *Philosophy and Technology, 33*, 541–546. https://doi.org/10.1007/s13347-020-00434-3

Flynn, C. (2017, September). *Personal thoughts on careers in AI policy and strategy* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/RCvetzfDnBNFX7pLH/personal-thoughts-on-careers-in-ai-policy-and-strategy

Flynn, C. (2020). *Recommendations on export controls for artificial intelligence*. Center for Security and Emerging Technology. https://cset.georgetown.edu/research/recommendations-on-export-controls-for-artificial-intelligence/

Fodor, J. (2018, December 13). *Critique of Superintelligence: Part 1* [Online forum post]. Effective Altruism. https://forum.effectivealtruism.org/posts/A8ndMGC4FTQq46RRX/critique-of-superintelligence-part-1

Foglia, A. T., & Jennings, A. K. (2013). A happiness approach to cost-benefit analysis: Foreword. *Duke Law Journal*, *62*(8), 1503–1508. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=3387&context=dlj

Food and Agriculture Organization of the United Nations. *FAOSTAT*. http://www.fao.org/faostat/en/#data/QL

Foote, M., & Raup, D. M. (1996). Fossil preservation and the stratigraphic ranges of taxa. *Paleobiology*, 121-140. https://www.jstor.org/stable/2401113

Foster, J. (2016). Do not hit print: The impact of 3D printing on distributive justice and why regulations are necessary to prevent consumer 3D vaccine printers. *Journal of*

Biosecurity, Biosafety and Biodefense Law, *7*(1), 25–45. https://doi.org/10.1515/jbbbl-2016-0007

Fowler, A. (2017, October 31). *Chief Justice Roberts and other judges have a hard time with statistics. That's a real problem*. The Washington Post. https://www.washingtonpost.com/news/monkey-cage/wp/2017/10/31/chief-justice-roberts-and-other-judges-have-a-hard-time-with-statistics-thats-a-real-problem/

François, J. M., Lachaux, C., & Morin, N. (2019). Synthetic biology applied to carbon conservative and carbon dioxide recycling pathways. *Frontiers in Bioengineering and Biotechnology*, 7, 446. https://doi.org/10.3389/fbioe.2019.00446

Friedman, E. A., & Gostin, L. O. (2015). Imagining global health with justice: In defense of the right to health. *Health Care Analysis*, *23*(4), 308–329. https://doi.org/10.1007/s10728-015-0307-x

Friesen, G., & Weller, P. A. (2006). Quantifying cognitive biases in analyst earnings forecasts. *Journal of Financial Markets*, *9*(4), 333–365. https://doi.org/10.1016/j.finmar.2006.07.001

Future of Life Institute. (2020). *Additional comments on the "White Paper: On Artificial Intelligence - A European approach to excellence and trust"*. https://futureoflife.org/wp-content/uploads/2020/10/Future-of-Life-Institute-_-Additional-Comments-on-European-Commision-White-Paper-on-AI-.pdf?x96845

Gaba, J. M. (1999). Environmental ethics and our moral relationship to future generations: Future rights and present virtue. *Columbia Journal of Environmental Law*, *24*(2), 249–288. https://core.ac.uk/download/pdf/323869626.pdf

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*, 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gallup, J. L., & Sachs, J. D. (2001). The economic burden of malaria. *American Journal of Tropical Medicine and Hygiene*, *64*(1-2 Suppl), 85–96.

Galway-Witham, J., & Stringer, C. (2018). How did Homo sapiens evolve? *Science*, *360*(6395), 1296–1298. https://doi.org/10.1126/science.aat6659

Gandhi, R., Sharma, A., Mahoney, W., Sousan, W., Zhu, Q., & Laplante, P. (2011). Dimensions of cyber-attacks: Cultural, social, economic, and political. *IEEE Technology and Society Magazine*, *30*(1), 28–38. https://doi.org/10.1109/MTS.2011.940293

García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research, 16*(1), 1437–1480. https://www.jmlr.org/papers/volume16/garcia15a/garcia15a.pdf

Garfinkel, B. (2018, October 12). *The future of surveillance*. Effective Altruism. https://www.effectivealtruism.org/articles/ea-global-2018-the-future-of-surveillance

Garfinkel, B. (2019, February 9). *How sure are we about this AI stuff?* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/9sBAW3qKppnoG3QPq/ben-garfinkel-how-sure-are-we-about-this-ai-stuff

Garfinkel, M. S., Endy, D., Epstein, G. L., & Friedman, R. M. (2007). Synthetic genomics: Options for governance. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, *5*(4), 359–362. https://doi.org/10.1089/bsp.2007.0923

Gaspar, R., Rohde, P., & Giger, J. (2019). Unconventional settings and uses of human enhancement technologies: A non-systematic review of public and experts' views on self-enhancement and DIY biology/biohacking risks. *Human Behavior and Emerging Technologies*, *1*(4), 295–305. https://doi.org/10.1002/hbe2.175

Gatowski, S. I., Dobbin, S. A., Richardson, J. T., Ginsburg, G. P., Merlino, M. L., & Dahir, V. (2001). Asking the gatekeepers: A national survey of judges on judging expert evidence in a post-Daubert world. *Law and Human Behavior*, *25*(5), 433–458. https://doi.org/10.1023/A:1012899030937

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

Gerhardt, M. J. (1991). The role of precedent in constitutional decisionmaking and theory. *George Washington Law Review*, 60, 68–159.

Gersen, J. E. (2007). Temporary legislation. *University of Chicago Law Review*, *74*, 247–298. https://chicagounbound.uchicago.edu/uclrev/vol74/iss1/12/

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press. https://doi.org/10.1017/CBO9780511808098

GiveWell. (2018, March). *Mass Distribution of Long-Lasting Insecticide-Treated Nets (LLINs)*. https://www.givewell.org/international/technical/programs/insecticide-treated-nets

Gloor, L. (2016a, August). *The case for suffering-focused ethics*. Center on Long-Term Risk. https://longtermrisk.org/the-case-for-suffering-focused-ethics/

Gloor, L. (2016b, November). *Altruists should prioritize artificial intelligence*. Center on Long-Term Risk. https://longtermrisk.org/altruists-should-prioritize-artificial-intelligence

Gloor, L. (2017, July). *Tranquilism*. Center on Long-Term Risk. https://longtermrisk.org/tranquilism/

Goertzel, B. (2015). Superintelligence: Fears, promises and potentials: Reflections on Bostrom's Superintelligence, Yudkowsky's *From AI to Zombies*, and Weaver and Veitas's "Open-Ended Intelligence". *Journal of Evolution and Technology, 24*(2), 55–87. https://jetpress.org/v25.2/goertzel.htm

Goertzel, B., & Pennachin, C. (2007). *Artificial general intelligence (Vol. 2)*. New York: Springer.

Gollier, C. (2013). *Pricing the planet's future: The economics of discounting in an uncertain world*. Princeton University Press.

González-Ricoy, I., & Gosseries, A. (Eds.) (2016). *Institutions for future generations*. Oxford University Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. arXiv. https://arxiv.org/abs/1412.6572

Goodland, R., & Anhang, J. (2009, November). *Livestock and climate change*. A Well-Fed World. https://awellfedworld.org/wp-content/uploads/Livestock-Climate-Change-Anhang-Goodland.pdf

Gosseries, A., & Lukas, M. (2009). *Intergenerational justice*. Oxford Scholarship Online. https://doi.org/10.1093/acprof:oso/9780199282951.001.0001

Goulas, E., & Zervoyianni, A. (2015). Economic growth and crime: Is there an asymmetric relationship? *Economic Modelling*, *49*, 286–295.

Government of Canada. (2019). *Canadian biosafety standards and guidelines.* https://www.canada.ca/en/public-health/services/canadian-biosafety-standards-guidelines.html

Grace, K. (2014, August 31). *Superintelligence reading group* [Online forum post]. LessWrong. https://www.lesswrong.com/posts/QDmzDZ9CEHrKQdvcn/superintelligence-reading-group

Grace et al., (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research, 62*, 729–754. https://doi.org/10.1613/jair.1.11222

Greaves, H. (2016). Cluelessness. *Proceedings of the Aristotelian Society*, *116*(3), 311–39. https://philpapers.org/rec/GREC-38

Greaves, H. (2017a). Discounting for public policy: A survey. *Economics and Philosophy*, *33*(3), 391–439. https://doi.org/10.1017/S0266267117000062

Greaves, H. (2017b). Population axiology. *Philosophy Compass*, *12*(11), e12442. https://doi.org/10.1111/phc3.12442

Greaves, H. (2020, November). *Evidence, cluelessness, and the long term - Hilary Greaves* [Talk transcript]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/LdZcit8zX89rofZf3/evidence-cluelessness-and-the-long-term-hilary-greaves

Greaves, H., & MacAskill, W. (2019). *The case for strong longtermism* [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/2020/Greaves_MacAskill_strong_longtermism.pdf

Greaves, H., MacAskill, W., O'Keeffe-O'Donovan, R., & Trammell, P. (2019). *Research agenda for the Global Priorities Institute.* Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/GPI-research-agenda-version-2.1.pdf

Greaves, H., & Pummer, T. (Eds.). (2019). *Effective altruism: philosophical issues.* Oxford University Press.

Greene, J., & Baron, J. (2001). Intuitions about declining marginal utility. *Journal of Behavioral Decision Making*, *14*(3), 243–255. https://doi.org/10.1002/bdm.375

Gronvall, G. K. (2015, February). *Mitigating the risks of synthetic biology.* Council on Foreign Relations: Center for Preventive Action. https://www.jstor.org/stable/resrep24166

Gronvall, G. K. (2016). *Synthetic biology: Safety, security, and promise.* Baltimore, MD: CreateSpace Independent Publishing Platform. https://www.centerforhealthsecurity.org/our-work/publications/synthetic-biology-safety-security-and-promise

Gronvall, G. K. (2017). Prevention of the development or use of biological weapons. *Health Security*, *15*(1), 36–37. https://dx.doi.org/10.1089%2Fhs.2016.0096

Gronvall, G. K., Bouri, N., Rambhia, K. J., Franco., C., & Watson, M. (2009). Prevention of biothreats: A look ahead. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, *7*(4), 433–442. https://doi.org/10.1089/bsp.2009.1112

Gruetzemacher, R. (2020). *Forecasting transformative AI* [Doctoral dissertation]. Auburn University. https://etd.auburn.edu/handle/10415/7338

Gruetzemacher, R., Dorner, F., Bernaola-Alvarez, N., Giattino, C., & Manheim, D. (2020). *Forecasting AI progress: A research agenda.* arXiv. https://arxiv.org/abs/2008.01848

Gruetzemacher, R., Paradice, D., & Lee, K. B. (2019). *Forecasting transformative AI: An expert survey.* arXiv. https://arxiv.org/abs/1901.08579

Gruetzemacher, R., & Whittlestone, J. (2019). *The transformative potential of artificial intelligence*. arXiv. https://arxiv.org/abs/1912.00747

Gunkel, D. J. (2018). *Robot rights*. MIT Press.

Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling artificial intelligence: Human perils. *Computer Law and Security Review, 32*(5), 749–758. https://doi.org/10.1016/j.clsr.2016.05.003

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2001). Inside the judicial mind. *Cornell Law Review*, *86*, 777–830. https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=1734&context=facpub

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2007). Blinking on the bench: How judges decide cases. *Cornell Law Review*, *93*, 1–43. https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=1707&context=facpub

Hale, Z. A. (2018, April 4). Patently unfair: The tensions between human rights and intellectual property protection. *The Arkansas Journal of Social Change and Public Service*. https://ualr.edu/socialchange/2018/04/04/patently-unfair/

Haley, A. G. (1956). *Space law and metalaw: A synoptic view*. Associazione Italiana Razzi.

Hanson, R. (2001). *Economic growth given machine intelligence* [Technical Report]. University of California, Berkeley. http://mason.gmu.edu/~rhanson/aigrow.pdf

Harrod, R. F. (1948). *Towards a Dynamic Economics: Some recent developments of economic theory and their application to policy*. MacMillan and Company, London. https://dspace.gipe.ac.in/xmlui/bitstream/handle/10973/29137/GIPE-017639.pdf?sequence=2

Hartman, G. R., Mersky, R. M., & Tate, C. L. (2014). *Landmark Supreme Court cases: The most influential decisions of the Supreme Court of the United States*. Infobase Publishing.

Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013). Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Global Environmental Change*, *23*(2), 485-498. https://doi.org/10.1016/j.gloenvcha.2012.12.006

Hathaway, O. A. (2003). Path dependence in the law: The course and pattern of legal change in a common law system. *Iowa Law Review*, *86*, 601–665.

Helgeson, C. (2020). Structuring decisions under deep uncertainty. *Topoi*, *39*, 257–269. https://doi.org/10.1007/s11245-018-9584-y

Hernández-Orallo, J., Martínez-Plumed, F., Avin, S., & Heigeartaigh, S. O. (2019, January). Surveying safety-relevant AI characteristics. In *AAAI Workshop on Artificial Intelligence Safety* (SafeAI 2019) (pp. 1–9). http://hdl.handle.net/10251/146561

Hewett, J. P., Wolfe, A. K., Bergmann, R. A., Stelling, S. C., & Davis, K. L. (2016). Human health and environmental risks posed by synthetic biology R&D for energy applications: A literature analysis. *Applied Biosafety*, *21*(4), 177–184. https://doi.org/10.1177/1535676016672377

Heyman, D., Epstein, G. L., & Moodie, M. (2009, December). *The Global Forum on Biorisks: Toward effective management and governance of biological risks*. Center for Strategic and International Studies. https://fas.org/programs/bio/resource/documents/The%20Global%20Forum%20on%20Biorisks.pdf

High-Level Expert Group on AI. (2019). *A definition of Artificial Intelligence: Main capabilities and scientific disciplines* [Technical report]. https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Hindin, D., Strosnider, K., & Trooboff, P. D. (2017, January 20). The role of export controls in regulating dual use research of concern: Striking a balance between freedom of fundamental research and national security. In National Academies of Sciences, Engineering, and Medicine, *Dual use research of concern in the life sciences*. The National Academies Press. https://doi.org/10.17226/24761

Hodges, S. (2010). *Detailed discussion of the humane methods of slaughter act*. Michigan State University College of Law. https://www.animallaw.info/article/detailed-discussion-humane-methods-slaughter-act

Hodgson, C. (2020, July 5). World Bank ditches second round of pandemic bonds. *Financial Times*. https://www.ft.com/content/949adc20-5303-494b-9cf1-4eb4c8b6aa6b

Hogarth, I. (2018, June 13). *AI nationalism*. https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism

Hollis, A. (2013). Synthetic biology: Ensuring the greatest global value. *Systems and Synthetic Biology*, *7*, 99–105. https://doi.org/10.1007/s11693-013-9115-5

Hollman, N., Winter, C. K., & Jauhar, A. (2021). *Long-term challenges of AI for the judiciary* [Manuscript in preparation].

Horowitz, M. C. (2019). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Journal of Strategic Studies*, *42*(6), 764–788. https://doi.org/10.1080/01402390.2019.1621174

Horowitz, M. C., Scharre, P., & Velez-Green, A. (2019). *A stable nuclear future? The impact of autonomous systems and artificial intelligence*. arXiv. https://arxiv.org/abs/1912.05291

Hottes, A. K. (2012). *Biosecurity challenges of the global expansion of high-containment biological laboratories*. The National Academies of Sciences, Engineering, Medicine. https://doi.org/10.17226/13315

Hsiang, S. M., Burke, M., & Miguel, E. (2013). Quantifying the influence of climate on human conflict. *Science*, *341*(6151). https://doi.org/10.1126/science.1235367

Hubbard, F. P. (2011). "Do Androids Dream?": Personhood and intelligent artifacts. *Temple Law Review, 83*(2), 405–474. https://www.templelawreview.org/article/83-2_hubbard

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019, June 5). *Deceptive alignment* [Online forum post]. AI Alignment forum. https://www.alignmentforum.org/posts/zthDPAjh9w6Ytbeks/deceptive-alignment

Humane Ventures. *2020 U.S. animal kill clock*. https://animalclock.org/

Huq, A. Z. (2020). A right to a human decision. *Virginia Law Review*, *106*, 611–688. https://www.virginialawreview.org/sites/virginialawreview.org/files/Huq_Book_0.pdf

Husbands, J. L. (2018). The challenge of framing for efforts to mitigate the risks of "dual use" research in the life sciences. *Futures*, *102*, 104–113. https://doi.org/10.1016/j.futures.2018.03.007

IARPA. (2020). *Finding engineering-linked indicators (FELIX)*. https://www.iarpa.gov/index.php/research-programs/felix

Ilchmann, K., & Revill, J. Chemical and biological weapons in the 'New Wars'. *Science and Engineering Ethics*, *20*, 753–767 (2014). https://doi.org/10.1007/s11948-013-9479-7

Ioannidis, J. P. (2018). Meta-research: Why research on research matters. *PLOS Biology*, *16*(3), e2005468. https://doi.org/10.1371/journal.pbio.2005468

Ioannidis, J. P., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLOS Biology*, *13*(10), e1002264. https://doi.org/10.1371/journal.pbio.1002264

Irving, G., Christiano, P., & Amodei, D. (2018). *AI safety via debate*. arXiv. https://arxiv.org/abs/1805.00899

Jakhu, R. S., Pelton, J. N., & Nyampong, Y. O. M. (2017). *Space mining and its regulation*. Springer. https://doi.org/10.1007/978-3-319-39246-2

Jayanti, A., & Avin, S. (2020). *It takes a village: The shared responsibility of "raising" an autonomous weapon*. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/resources/it-takes-village

Jeffery, C. R. (1957). The development of crime in early English society. *Journal of Criminal Law, Criminology, and Police Science*, *47*(6), 647–666. https://www.jstor.org/stable/1140057

Jinyuan, S. U. (2017). Space arms control: Lex lata and currently active proposals. *Asian Journal of International Law*, *7*(1), 61–93. https://doi.org/10.1017/S2044251315000223

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

John, T. M. (2020a). *Longtermist institutional design and policy: A literature review (draft)* [Manuscript in preparation]. The Forethought Foundation for Global Priorities Research.

John, T. M. (2020b). Empowering future people by empowering the young? [Manuscript in preparation]. https://docs.google.com/document/d/1NeP3NPd-m4NxhS9O5aqdsaphh4v__iFcEAjMCcPHM_0/edit

John, T. M., & MacAskill, W. (2021). *Longtermist institutional reform* [Forthcoming]. In N. Cargill & T. M. John (Eds.), *The long view*. https://philpapers.org/rec/JOHLIR

Johnson, J. (2020, June 24). *Boston bans facial recognition due to concern about racial bias*. VentureBeat. https://venturebeat.com/2020/06/24/boston-bans-facial-recognition-due-to-concern-about-racial-bias/

Jolls, C., Sunstein, C. R., & Thaler, R. (1998). A behavioral approach to law and economics. *Stanford Law Review*, *50*(5), 1471–1550. https://doi.org/10.2307/1229304

Jonason, P. K., Duineveld, J. J., & Middleton, J. P. (2015). Pathology, pseudopathology, and the dark triad of personality. *Personality and Individual Differences*, *78*, 43–47. https://doi.org/10.1016/j.paid.2015.01.028

Jones, N., O'Brien, M., & Ryan, T. (2018). Representation of future generations in United Kingdom policy-making. *Futures*, *102*, 153–163. https://doi.org/10.1016/j.futures.2018.01.007

Joshi, S. (2020, August 25). *We're (surprisingly) more positive about tackling bio risks: outcomes of a survey* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/oxdmyQWsnNwCGSLLC/we-re-surprisingly-more-positive-about-tackling-bio-risks

Jumper, J. et al. (2020). *AlphaFold: a solution to a 50-year-old grand challenge in biology*. DeepMind. https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Kaebnick, G. E., Gusmano, M. K., & Murray, T. H. (2014). The ethics of synthetic biology: Next steps and prior questions. *Synthetic Future*, *44*(S5), S4–S26. https://doi.org/10.1002/hast.392

Kagan, E. (2001). Presidential administration. *Harvard Law Review*, *114*(8), 2245–2385. https://www.jstor.org/stable/1342513

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, *5*(1), 193–206. https://doi.org/10.1257/jep.5.1.193

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi:10.2307/1914185

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*(2), 123–141. https://doi.org/10.1016/0010-0277(82)90022-1

Kanetake, M. (2018). Balancing innovation, development, and security. In N. Craik, C. S. G. Jefferies, S. L. Seck, & T. Stephens (Eds.), *Global environmental change and innovation in international law* (pp. 180–200). Cambridge University Press. https://doi.org/10.1017/9781108526081.011

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, *6*, 39589. https://doi.org/10.1038/srep39589

Kaplow, L. (1992). Rules versus standards: An economic analysis. *Duke Law Journal*, *42*, 557–692. https://scholarship.law.duke.edu/dlj/vol42/iss3/2/

Kaplow, L., & Shavell, S. (2002). *Fairness versus welfare*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674023642 (Version previously published as Kaplow, L., & Shavell, S. (2001). Fairness versus welfare. *Harvard Law Review*, *114*, 961–1388)

Karnofsky, H. (2011, August). *Why we can't take expected value estimates literally (even when they're unbiased)* [Blog post]. GiveWell. https://blog.givewell.org/2011/08/18/why-we-cant-take-expected-value-estimates-literally-even-when-theyre-unbiased/

Karnofsky, H. (2013, May). *Flow-through effects* [Blog post]. GiveWell. https://blog.givewell.org/2013/05/15/flow-through-effects/

Karnofsky, H. (2016a, May 6). *Potential risks from advanced artificial intelligence: The philanthropic opportunity*. Open Philanthropy. https://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity

Karnofsky, H. (2016b). *Some background on our views regarding advanced artificial intelligence*. Open Philanthropy. https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence

Kelle, A. (2013). Beyond patchwork precaution in the dual-use governance of synthetic biology. *Science and Engineering Ethics*, *19*(3), 1121–1139. https://doi.org/10.1007/s11948-012-9365-8

Kemp, L., Cihon, P., Maas, M. M., Belfield, H. ÓhÉigeartaigh, S., Leung, J., & Cremer, C. Z. (2019). *UN high-level panel on digital cooperation: A proposal for international AI governance*. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/resources/proposal-international-ai-governance

Kessler syndrome. (n.d.) Wikipedia. https://en.wikipedia.org/wiki/Kessler_syndrome

Khan, L. M. (2016). Amazon's antitrust paradox. *Yale Law Journal*, *126*(3), 710–805. https://digitalcommons.law.yale.edu/ylj/vol126/iss3/3

Khan, Y., O'Sullivan, T., Brown, A., Tracey, S., Gibson, J., Généreux, M., Henry, B., & Schwartz, B. (2018). Public health emergency preparedness: a framework to promote resilience. *BMC Public Health*, *18*(1), 1344. https://dx.doi.org/10.1186/s12889-018-6250-7

Klinkrad, H. (2010). Space debris. *Encyclopedia of Aerospace Engineering*. https://doi.org/10.1002/9780470686652.eae325

Klotz, L. (2019, February 25). *Human error in high-biocontainment labs: A likely pandemic threat*. Bulletin of the Atomic Scientists. https://thebulletin.org/2019/02/human-error-in-high-biocontainment-labs-a-likely-pandemic-threat/

Knopf, J. W. (2010). The fourth wave in deterrence research. *Contemporary Security Policy*. *31*(1), 1–33. https://doi.org/10.1080/13523261003640819

Kobokovich, A., West, R., Montague, M., Inglesby, T., & Gronvall, G. K. (2019). Strengthening security for gene synthesis: Recommendations for governance. *Health Security*, *17*(6), 419–429. http://doi.org/10.1089/hs.2019.0110

Koehler, A. (2020, August). *How to use your career to help reduce existential risk*. 80,000 Hours. https://80000hours.org/articles/how-to-reduce-existential-risk

Kofler, N., Collins, J. P., Kuzma, J., Marris, E., Esvelt, K., Nelson, M. P., Newhouse, A., Rothschild, L. J., Vigliotti, V. S., Semenov, M., Jacobsen, R., Dahlman, J. E., Prince, S., Caccone, A., Brown, T., Schmitz, O. J. (2018, November 2). Editing nature: Local roots of global governance. *Science*, *362*(6414), 527–529. https://doi.org/10.1126/science.aat4612

Kohli, P., Dvijotham, K., Uesato, J., & Gowal, S. (2019, March 28). *Identifying and eliminating bugs in learned predictive models*. DeepMind. https://deepmind.com/blog/article/robust-and-verified-ai

König, H., Dorado-Morales, P., & Porcar, M. (2015). Responsibility and intellectual property in synthetic biology: A proposal for using responsible research and innovation as a basic framework for intellectual property decisions in synthetic biology. *EMBO reports*, *16*(9), 1055–1059. https://doi.org/10.15252/embr.201541048

König, H., Frank, D., Heil, R., & Coenen, C. (2013). Synthetic genomics and synthetic biology applications between hopes and concerns. *Current Genomics*, *14*(1), 11–24. https://doi.org/10.2174/1389202911314010003

Korsgaard, C. M. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Kosal, M. E. (2014). A new role for public health in bioterrorism deterrence. *Frontiers in Public Health*, *2*(278), 1–4. https://doi.org/10.3389/fpubh.2014.00278

Krakovna, V., Kumar, R., Orseau, L., & Turner, A. (2019b, March 11). *Designing agent incentives to avoid side effects*. Medium. https://medium.com/@deepmindsafetyresearch/designing-agent-incentives-to-avoid-side-effects-e1ac80ea6107

Krakovna, V., Orseau, L., Kumar, R., Martic, M., & Legg, S. (2019a). *Penalizing side effects using stepwise relative reachability*. arXiv. https://arxiv.org/abs/1806.01186v2

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(3), 633–705.

Kuecken, M., Thuilliez, J., Valfort, M. A. (2014). *Does malaria control impact education? A study of the Global Fund in Africa.* https://halshs.archives-ouvertes.fr/halshs-00924112

Kun, K. E., Rose, D. A., Morris, T., Salter, M., Lamia, T., Bhalakia, A., & McLees, A. W. (2014). Conceptualizing and measuring community preparedness within public health preparedness and response: complexities and lessons learned. *Journal of Public Health Management and Practice*, *20*(4), E1–E5. https://doi.org/10.1097/phh.0b013e3182a5bbcc

Kuran, T. (2012). *The long divergence: How Islamic law held back the Middle East.* Princeton University Press. https://www.jstor.org/stable/j.ctt7t73p

Kurki, V. A. J. (2019). *A theory of legal personhood.* Oxford University Press.

Kurki, V. A. J. (2017). Why things can hold rights: Reconceptualizing the legal person. In V. A. J. Kurki, & T. Pietrzykowski (Eds.), *Legal personhood: Animals, artificial intelligence and the unborn* (pp. 69–89). Springer, Cham. https://doi.org/10.1007/978-3-319-53462-6

Kurki, V. A. J., & Pietrzykowski, T. (2017). *Legal personhood: Animals, artificial intelligence and the unborn.* Springer.

Kurzweil, R., Richter, R., Kurzweil, R., & Schneider, M. L. (1990). *The age of intelligent machines* (Vol. 579). Cambridge: MIT press.

Kuzma, J., & Rawls, L. (2016). Engineering in the wild: Gene drives and intergenerational equity, *Jurimetrics Journal*, *56*, 279–296. https://research.ncsu.edu/ges/files/2014/02/engineering_the_wild.authcheckdam.pdf

Kvinta, B. (2011). Quarantine powers, biodefense, and Andrew Speaker. *Journal of Biosecurity, Biosafety and Biodefense Law*, *1*(1), 1–17. https://doi.org/10.2202/2154-3186.1002

Kwisda, K., White, L., & Hübner, D. (2020). Ethical arguments concerning human-animal chimera research: A systematic review. *BMC Medical Ethics*, 21. https://dx.doi.org/10.1186/s12910-020-00465-7

Kysar, R. M. (2010). Lasting legislation. *University of Pennsylvania Law Review*, 159, 1007.

LaFreniere, D. (2019). Forgiveness or permission: How may the United States Government conduct experiments on the public or in public? *Journal of Biosecurity, Biosafety and Biodefense Law*, *10*(1), 1–8.

Lai, H-E., Canavan, C., Cameron, L., Moore, S., Danchenko, M., Kuiken, T., Sekeyová, Z., & Freemont, P. S. (2019). Synthetic biology and the United Nations. *Trends in Biotechnology*, *37*(11), 1146–1151. https://doi.org/10.1016/j.tibtech.2019.05.011

Laird S. A., & Wynberg R. P. (2018, January 10). *A fact finding and scoping study on digital sequence information on genetic resources in the context of the convention on biological diversity and Nagoya Protocol.* Convention on Biological Diversity, Ad Hoc Technical Expert Group on Synthetic Biology. https://www.cbd.int/doc/c/e95a/4ddd/4baea2ec772be28edcd10358/dsi-ahteg-2018-01-03-en.pdf

Lawsky, S. (2020). *Spring reported entry level hiring report 2020.* PrawfsBlawg. https://prawfsblawg.blogs.com/prawfsblawg/2020/05/spring-reported-entry-level-hiring-report-2020-1.html

Le Feuvre, R. A., & Scrutton, N. S. (2018). A living foundry for synthetic biological materials: a synthetic biology roadmap to new advanced materials. *Synthetic and Systems Biotechnology*, *3*(2), 105–112. https://doi.org/10.1016/j.synbio.2018.04.002

Legg, S., & Hutter, M. (2007a). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications*, *157*, 17–24.

Legg, S., & Hutter, M. (2007b). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*(4), 391–444. https://doi.org/10.1007/s11023-007-9079-x

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). *Scalable agent alignment via reward modeling: A research direction*. arXiv. https://arxiv.org/abs/1811.07871v1

Lenman, J. (2000). Consequentialism and cluelessness. *Philosophy and Public Affairs*, *29*(4), 342–370. https://doi.org/10.1111/j.1088-4963.2000.00342.x

Lentzos, F. (2019). *Compliance and enforcement in the biological weapons regime*. United Nations Institute for Disarmament Research. https://www.unidir.org/sites/default/files/2020-02/compliance-bio-weapons.pdf

Leung, J. (2018, September 28). *Analyzing AI actors*. Effective Altruism. https://www.effectivealtruism.org/articles/ea-global-2018-analyzing-ai-actors

Leung, J. (2019). *Who will govern artificial intelligence? Learning from the history of strategic politics in emerging technologies* [Doctoral dissertation]. University of Oxford. https://ora.ox.ac.uk/objects/uuid:ea3c7cb8-2464-45f1-a47c-c7b568f27665

Lev, O. (2019). Regulating dual-use research: Lessons from Israel and the United States. *Journal of Biosafety and Biosecurity*, *1*(2), 80–85. https://doi.org/10.1016/j.jobb.2019.06.001

Levenda, K. (2013). Legislation to protect the welfare of fish. *Animal Law*, *20*, 119–144.

Lewis, G. (2018a, April). *The person-affecting value of existential risk reduction* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/dfiKak8ZPa46N7Np6/the-person-affecting-value-of-existential-risk-reduction

Lewis, G. (2018b, February 19). *Horsepox synthesis: A case of the unilateralist's curse?* Bulletin of the Atom Scientists. https://thebulletin.org/2018/02/horsepox-synthesis-a-case-of-the-unilateralists-curse/

Lewis, G. (2020, March). *Reducing global catastrophic biological risks*. 80,000 Hours. https://80000hours.org/problem-profiles/biosecurity/

Lewis, G., Jordan, J. L., Relman, D. A., Koblentz, G. D., Leung, J., Dafoe, A., Nelson, C., Epstein, G. L., Katz, R., Montague, M., Alley, E. C., Filone, C. M., Luby, S., Churche, G. M., Millett, P., Esvelt, K. M., Cameron, E. E., Inglesby, T. V. (2020). The biosecurity benefits of genetic engineering attribution. *Nature Communications*, *11*(6294). https://doi.org/10.1038/s41467-020-19149-2

Lewis, G., Millett, P., Sandberg, A., Snyder-Beattie, A., & Gronvall, G. (2019). Information hazards in biotechnology. *Risk Analysis*, *39*(5), 975–981. https://doi.org/10.1111/risa.13235

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

Lindblom, C. E. (1959). The science of "muddling through". *Public Administration Review*, *19*(2), 79–88. http://www.jstor.org/stable/973677?origin=JSTOR-pdf

Lindquist, S. A., & Cross, F. C. (2008). 1 *Stability, predictability and the rule of law: Stare decisis as reciprocity norm* [Manuscript]. University of Texas School of Law. https://law.utexas.edu/conferences/measuring/The%20Papers/Rule%20of%20Law%20Conference.crosslindquist.pdf

Liu, H. Y., Lauta, K. C., & Maas, M. M. (2018). Governing boring apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures*, *102*, 6–19. https://doi.org/10.1016/j.futures.2018.04.009

Livermore, M. A., & Revesz, R. L. (Eds.). (2013). *The globalization of cost-benefit analysis in environmental policy*. Oxford University Press.

Lyall, F., & Larsen, P. B. (2018). *Space law: A treatise*. Routledge.

Maas, M. M. (2019a). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy*, *40*(3), 285–311. https://doi.org/10.1080/13523260.2019.1576464

Maas, M. M. (2019b). Innovation-proof global governance for military artificial intelligence? *Journal of International Humanitarian Legal Studies*, *10*(1), 129–157. https://doi.org/10.1163/18781527-01001006

MacAskill, W. (2018). Understanding effective altruism and its challenges. In D. Boonin (Ed.), *The Palgrave Handbook of Philosophy and Public Policy* (pp. 441-453). Palgrave Macmillan, Cham. https://www.academia.edu/43357150/Understanding_Effective_Altruism_and_its_Challenges

MacAskill, W. (2019a, July). *'Longtermism'* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/qZyshHCNkjs3TvSem/longtermism

MacAskill, W. (2019b, September). When should an effective altruist donate? [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/2020/William_MacAskill_when_donate.pdf

MacAskill, W. (2020a). *Human extinction, asymmetry, and option value* [Manuscript in preparation]. https://docs.google.com/document/d/1hQI3otOAT39sonCHIM6B4na9BKeKjEl7wUKacgQ9qF8/edit

MacAskill, W. (2020b). *Are we living at the hinge of history?* [Working paper]. https://globalprioritiesinstitute.org/william-macaskill-are-we-living-at-the-hinge-of-history/

MacAskill, M., Bykvist, K., & Ord, T. (2020). *Moral Uncertainty*. Oxford University Press.

Machery, E., & O'Neill, E. (Eds.). (2014). *Current controversies in experimental philosophy*. Routledge.

Mahfoud, T., Aicardi, C., Datta, S., & Rose, N. (2018). The limits of dual use. *Issues in Science and Technology*, *34*(4), 73–78.

Makridakis, S. (2017). The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms. *Futures, 90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

Mandel, G. N., & Marchant, G. E. (2014). The living regulatory challenges of synthetic biology. *Iowa Law Review*, *100*, 155–200. https://ilr.law.uiowa.edu/print/volume-100-issue-1/the-living-regulatory-challenges-of-synthetic-biology/

Manheim, D. (2018). Questioning estimates of natural pandemic risk. *Health Security*, *16*(6), 381–390. https://dx.doi.org/10.1089%2Fhs.2018.0039

Manheim, D. (2019). Multiparty dynamics and failure modes for machine learning and artificial intelligence. *Big Data and Cognitive Computing, 3*(2), Article 21. https://doi.org/10.3390/bdcc3020021

Marcello, I., & Effy, V. (2018). Dual use in the 21st century: Emerging risks and global governance. *Swiss Medical Weekly*, *148*(14688). https://doi.org/10.4414/smw.2018.14688

Martinez, E., & Winter, C. K. (2021). *Legal priorities survey*. [Manuscript in preparation].

Martinez, R. (2019). Artificial intelligence: distinguishing between types & definitions. *Nevada Law Journal*, *19*(3), 1015–1041.

Masci, D. (2016, July 26). *Human enhancement: The scientific and ethical dimensions of striving for perfection*. Pew Research Center. https://www.pewresearch.org/science/2016/07/26/human-enhancement-the-scientific-and-ethical-dimensions-of-striving-for-perfection/

Massachusetts Body of Liberties. (1641). *Massachusetts body of liberties*. Online Library of Liberty. https://oll.libertyfund.org/page/1641-massachusetts-body-of-liberties

Matheny, J. G. (2007). Reducing the risk of human extinction. *Risk Analysis: An International Journal*, *27*(5), 1335–1344. https://doi.org/10.1111/j.1539-6924.2007.00960.x

McAdams, R. H. (1997). The origin, development, and regulation of norms. *Michigan Law Review*, *96*(2), 338–433. https://doi.org/10.2307/1290070

McAdams, R. H. (2000). A focal point theory of expressive law. *Virginia Law Review*, 1649–1729. https://doi.org/10.2307/1073827

McCabe, C., Claxton, K., & Culyer, A. J. (2008). The NICE cost-effectiveness threshold. *Pharmacoeconomics*, *26*(9), 733–744. https://doi.org/10.2165/00019053-200826090-00004

McCarthy, J. (2007). What is artificial intelligence? [Manuscript]. Stanford University. http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12–14. https://doi.org/10.1609/aimag.v27i4.1904

McCarthy-Jones, S. (2019). The autonomous mind: The right to freedom of thought in the 21st century. *Frontiers in Artificial Intelligence*, *2*, 19. https://doi.org/10.3389/frai.2019.00019

McDonald, M. M., Donnellan, M. B., & Navarrete, C. D. (2012). A life history approach to understanding the dark triad. *Personality and Individual Differences*, *52*(5), 601–605. https://doi.org/10.1016/j.paid.2011.12.003

McGee, V. (1991). We turing machines aren't expected-utility maximizers (even ideally). *Philosophical Studies*, *64*(1), 115–123. https://doi.org/10.1007/BF00356093

McIntyre, F., & Simkovic, M. (2018). Value of a law degree by college major. *Journal of Legal Education*, *68*, 585.

McKenna, S. (2020, May 20). *Human viruses can jump into animals, too—Sowing the seeds of future epidemics*. Scientific American. https://www.scientificamerican.com/article/human-viruses-can-jump-into-animals-too-sowing-the-seeds-of-future-epidemics/

McKinnon, C. (2017). Endangering humanity: An international crime? *Canadian Journal of Philosophy*, *47*(2–3), 395–415. https://doi.org/10.1080/00455091.2017.1280381

Means, L. (2019). IS and bioweapons: How can the BWC be used to intercede when a non-signator IS suspected of bioweapon creation? *Journal of Biosecurity, Biosafety, and Biodefense Law*, *10*(1), 1–9. https://doi.org/10.1515/jbbbl-2019-0002

Meenan, C. (2020 May 19). *The future of pandemic financing: Trigger design and 2020 hindsight.* Centre for Disaster Protection. https://www.disasterprotection.org/latest-news/the-future-of-pandemic-financing-trigger-design-and-2020-hindsight

Meichenbaum, D. (2009). Ways to improve political decision-making: Negotiating errors to be avoided. In *Psychological and Political Strategies for Peace Negotiation* (pp. 87–97). Springer. https://melissainstitute.org/wp-content/uploads/2015/10/DecisionMaking.pdf

Merriam-Webster. *'Pandemic' vs. 'epidemic': How they overlap and where they differ.* https://www.merriam-webster.com/words-at-play/epidemic-vs-pandemic-difference

Merryman, J. H. (1975). Legal education there and here: A comparison. *Stanford Law Review*, *27*(3), 859–878.

Merryman, J. H. (1977). Comparative law and social change: on the origins, style, decline & revival of the law and development movement. *The American Journal of Comparative Law*, *25*(3), 457–491.

Merryman, J. H. (1981). On the convergence (and divergence) of the civil law and the common law. *Stanford Journal of International Law*, *17*(2), 357–388.

Merryman, J. H., & Pérez-Perdomo, R. (2018). *The civil law tradition: An introduction to the legal systems of Europe and Latin America.* Stanford University Press.

Michaels, A. C. (2020). Artificial intelligence, legal change, and separation of powers. *University of Cincinnati Law Review*, *88*(4), 1083–1103. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3459069

Millett. P. D. (2017, January 17). Gaps in the international governance of dual-use research of concern. In National Academies of Sciences, Engineering, and Medicine. *Dual use research of concern in the life sciences.* The National Academies Press. https://doi.org/10.17226/24761

Millett, P. D., & Snyder-Beattie, A. (2017). Existential risk and cost-effective biosecurity. *Health Security*, *15*(4), 373–383. https://doi.org/10.1089/hs.2017.0028

Minsky, M. (Ed.) (1969). *Semantic information processing.* MIT Press.

Mody, S. (2002). Brown footnote eleven in historical context: Social science and the Supreme Court's quest for legitimacy. *Stanford Law Review*, *54*(4), 793–829. https://doi.org/10.2307/1229579

Mogensen, A. (2019, September). *Maximal cluelessness* [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/2020/Andreas_Mogensen_maximal_cluelessness.pdf

Mogensen, A. (2020, June). *Moral demands and the far future* [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/Working-Paper-1-2020-Andreas-Mogensen.pdf

Monett, D., & Lewis, C. W. (2018). Getting clarity by defining artificial intelligence—A survey. In *3rd conference on philosophy and theory of artificial intelligence* (pp. 212–214). Berlin: Springer.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*, *111*(20), 7176–7184. https://doi.org/10.1073/pnas.1319946111

Moses, L. B. (2011). *Recurring dilemmas: The law's race to keep up with technological change.* SSRN. http://dx.doi.org/10.2139/ssrn.979861

Muehlhauser. (2013). *What is AGI?* Machine Intelligence Research Institute. https://intelligence.org/2013/08/11/what-is-agi/

Muehlhauser, L., & Salamon, A. (2012). Intelligence explosion: Evidence and import. In *Singularity Hypotheses* (pp. 15–42). Springer, Berlin, Heidelberg.

Mueller, G. O. (1961). The German Penal Code of 1871 (p. 107). Rothman. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=130449

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 555–572). Springer. https://doi.org/10.1007/978-3-319-26485-1_33

Nakad-Weststrate, H. W. R, van den Herik, H. J., Jongbloed, A. T., & Salem, A. B. M. (2015). The rise of the robotic judge in modern court proceedings. In *The 7th International Conference on Information Technology*. http://icit.zuj.edu.jo/icit15/DOI/Artificial_Intelligence/0009.pdf

National Aeronautics and Space Administration (2018). *The global exploration strategy framework: Executive summary*. https://www.nasa.gov/pdf/296751main_GES_framework.pdf

National Aeronautics and Space Administration (2020). *Planetary protection*. https://sma.nasa.gov/sma-disciplines/planetary-protection

National Academies of Sciences, Engineering, and Medicine. (2016a, July 28). *Gene drives on the horizon: Advancing science, navigating uncertainty, and aligning research with public values*. The National Academies Press. https://doi.org/10.17226/23405

National Academies of Sciences, Engineering, and Medicine. (2016b). *Global health risk framework: Pandemic financing: Workshop summary*. The National Academies Press. https://doi.org/10.17226/21855

National Academies of Sciences, Engineering, and Medicine. (2017a). *A Proposed Framework for Identifying Potential Biodefense Vulnerabilities Posed by Synthetic Biology: Interim Report*. The National Academies Press. https://doi.org/10.17226/24832

National Academies of Sciences, Engineering, and Medicine. (2017b). *Dual use research of concern in the life sciences*. National Academies Press. https://doi.org/10.17226/24761

National Academies of Sciences, Engineering, and Medicine. (2018a). *Biodefense in the age of synthetic biology*. National Academies Press. https://doi.org/10.17226/24890

National Academies of Sciences, Engineering, and Medicine. (2018b). *Governance of dual use research in the life sciences*. National Academies Press. https://doi.org/10.17226/25154

National Academies of Sciences, Engineering, and Medicine. (2020a). *Evidence-based practice for public health emergency preparedness and response*. The National Academies Press. https://doi.org/10.17226/25650

National Academies of Sciences, Engineering, and Medicine. (2020b). *Framework for equitable allocation of COVID-19 vaccine*. The National Academies Press. https://doi.org/10.17226/25917

National Academies of Sciences, Engineering, and Medicine. (2020c). *A Framework for equitable allocation of vaccine for the Novel Coronavirus*. https://www.nationalacademies.org/our-work/a-framework-for-equitable-allocation-of-vaccine-for-the-novel-coronavirus#

National Academies of Sciences, Engineering, and Medicine (2020d). *Ethical, legal, and regulatory issues associated with neural chimeras and organoids*. https://www.nationalacademies.org/our-work/ethical-legal-and-regulatory-issues-associated-with-neural-chimeras-and-organoids

National Institute of Corrections (2013). *Evidence-based practices in the criminal justice system*. U.S Department of Justice. https://info.nicic.gov/nicrp/system/files/026917.pdf

National Research Council. (2004). *Biotechnology research in an age of terrorism*. National Academies Press.

National Research Council. (2014). *Science Needs for Microbial Forensics: Developing Initial International Research Priorities*. https://www.ncbi.nlm.nih.gov/books/NBK234876/

Nature Biotechnology. (2009). What's in a name? *Nature Biotechnology, 27*, 1071–1073. https://doi.org/10.1038/nbt1209-1071

Naudé, W., & Dimitri, N. (2020). The race for an artificial general intelligence: Implications for public policy. *AI and Society*, *35*, 367–379. https://doi.org/10.1007/s00146-019-00887-x

Negowetti, N. E. (2018). Establishing and enforcing animal welfare labeling claims: Improving transparency and ensuring accountability. *Journal of Animal and Natural Resource Law*, *14*, 131–158.

Nell, V. (2006). Cruelty's rewards: The gratifications of perpetrators and spectators. *Behavioral and Brain Sciences*, *29*(3), 211–224. https://doi.org/10.1017/S0140525X06009058

Nelson, C., Lurie, N., Wasserman, J., & Zakowski, S. (2007). Conceptualizing and defining public health emergency preparedness. *American Journal of Public Health, 97*, S9–S11. https://doi.org/10.2105/AJPH.2007.114496

Ngo, R. (2019, February 21). *Disentangling arguments for the importance of AI safety* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/posts/w6d7XBCegc96kz4n3/the-argument-from-philosophical-difficulty

Ngo, R (2020, September 28). *AGI safety from first principles* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ

Niiler, E. (2019, March 25). *Can AI be a fair judge in court? Estonia thinks so*. Wired. https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/

Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge University Press.

Nindler, R. (2019). The United Nation's capability to manage existential risks with a focus on artificial intelligence. *International Community Law Review*, *21*(1), 5–34.

Noble, C., Min, J., Olejarz, J., Buchthal, J., Chavez, A., Smidler, A. L., DeBenedictis, E. A., Church, G. M., Nowak, M. A., & Esvelt, K. M. (2019). Daisy-chain gene drives for the alteration of local populations. *Proceedings of the National Academy of Sciences, 116*(17), 8275-8282. https://doi.org/10.1073/pnas.1716358116

Northpointe, Inc. (2015, March 19). *Practitioner's guide to COMPAS core*. http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

NSCS. (2018). I*ntroduction to the evidence-based judicial decision making curriculum*. Courts and Jails. https://www.ncsc.org/__data/assets/pdf_file/0018/14490/intro-to-ebj-decision-making.pdf

Nussbaum, M. C. (2009). *Frontiers of justice: Disability, nationality, species membership*. Harvard University Press.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American Economic Review, 89*(1), 103–124. https://doi.org/10.1257/aer.89.1.103

Organisation for Economic Co-operation and Development. (2019). *Recommendation of the council on artificial intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Organisation for Economic Co-operation and Development (2009). *Regulatory impact analysis: A tool for policy coherence*. https://doi.org/10.1787/19900481

Office for AI. (2020). *A guide to using artificial intelligence in the public sector*. Gov.UK. https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector

ÓhÉigeartaigh, S., Whittlestone, J., Liu, Y., Zeng, Y., & Liu, Z. (2020). Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy and Technology, 33*, 571–593. https://doi.org/10.1007/s13347-020-00402-x

World Organisation for Animal Health. (2020, November 27). *Questions and answers on COVID-19*. https://www.oie.int/scientific-expertise/specific-information-and-recommendations/questions-and-answers-on-2019novel-coronavirus

Oldham, P., Hall, S., & Burton, G. (2012). Synthetic biology: Mapping the scientific landscape. *PLOS One*, *7*(4), e34368. https://doi.org/10.1371/journal.pone.0034368

Open Philanthropy. (2017, September). *Genspace—DIYbio labs project*. Open Philanthropy: Focus areas. https://www.openphilanthropy.org/focus/global-catastrophic-risks/biosecurity/genspace-diy-bio-labs-project

Open Philanthropy. (2020a). *Cause selection*. Open Philanthropy: Research & Ideas. https://www.openphilanthropy.org/research/cause-selection

Open Philanthropy. (2020b). *Global catastrophic risks*. Open Philanthropy: Focus areas. https://www.openphilanthropy.org/focus/global-catastrophic-risks

Ord, T. (2013). The moral imperative toward cost-effectiveness in global health. In H. Greaves & T. Pummer (Eds.), *Effective Altruism: Philosophical Issues* (pp. 29–36). Oxford University Press. https://doi.org/10.1093/oso/9780198841364.003.0002

Ord, T. (2014, July 3). *The timing of labour aimed at reducing existential risk*. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/the-timing-of-labour-aimed-at-reducing-existential-risk/

Ord, T. (2020). *The precipice: existential risk and the future of humanity*. Hachette Books.

Orseau, L., & Armstrong, S. (2016). *Safely interruptible agents*. Future of Humanity Institute. https://www.fhi.ox.ac.uk/wp-content/uploads/Interruptibility.pdf

Ortega, P. A., Maini, V., & the DeepMind safety team. (2018, September 27). *Building safe artificial intelligence: specification, robustness, and assurance*. Medium. https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1

Ortiz-Ospina, E., & Roser, M. (2013). *Global extreme poverty* [Online resource]. Our World in Data. https://ourworldindata.org/extreme-poverty

Ortiz-Ospina, E., & Roser, M. (2016). *Global health* [Online resource]. Our World in Data. https://ourworldindata.org/health-meta

Osman, N. D. (2018). The legal regulation of biosafety risk: A comparative legal study between Malaysia and Singapore [Unpublished thesis]. Nottingham Trent University. https://doi.org/10.13140/RG.2.2.26177.10084

Ostertag, J. R. (1993). Legal education in Germany and the United States—A structural comparison. *Vanderbilt Journal of Transnational Law*, *26*, 301–340.

Ovadya, A., & Whittlestone, J. (2019). *Reducing malicious use of synthetic media research: Considerations and potential release practices for machine learning*. arXiv. https://arxiv.org/abs/1907.11274

O'Keefe, C. (2018). *Stable agreements in turbulent times: A toolkit for constrained temporal decision transmission* [Technical report]. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/Stable-Agreements.pdf

O'Keefe, C. (2020a). *Antitrust-compliant AI industry self-regulation* [Manuscript in preparation]. https://cullenokeefe.com/blog/antitrust-compliant-ai-industry-self-regulation

O'Keefe, C. (2020b). *How will national security considerations affect antitrust decisions in AI? An examination of historical precedents* [Technical report]. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/How-Will-National-Security-Considerations-Affect-Antitrust-Decisions-in-AI-Cullen-OKeefe.pdf

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., & Dafoe, A. (2020, February). *The windfall clause: Distributing the benefits of AI for the common good* [Technical report]. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/windfallclause

O'Keefe, C., Lansky, D., Clark, J., & Payne, C. (2019). *Before the United States Patent and Trademark Office Department of Commerce: Comment regarding request for comments on intellectual property protection for artificial intelligence, Innovation Docket No. PTO–C–2019–0038, Comment of OpenAI, LP addressing question 3*. https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf archived at https://perma.cc/ZS7G-2QWF

O'Neill, G. (1974). The colonization of space. In *Space Manufacturing Facilities* (p. 2041). https://doi.org/10.2514/6.1975-2041

Page, M., Aiken, C., & Murdick, D. (2020, October). *Future indices: How crowd forecasting can inform the big picture*. Center for Security and Emerging Technology, Georgetown University. https://cset.georgetown.edu/research/future-indices

Palmer, M. J. (2020). Learning to deal with dual use. *Science*, *367*(6482), 1057. https://doi.org/10.1126/science.abb1466

Palmer, M. J., Tiu, B. C., Weissenbach, A. S., & Relman, D. A. (2017). On Defining Global Catastrophic Biological Risks. *Health Security*, *15*(4), 347–348. https://doi.org/10.1089/hs.2017.0057

Pamlin, D., & Armstrong, S. (2015). *12 risks that threaten human civilisation*. Global Challenges Foundation. https://www.pamlin.net/s/12-Risks-that-threaten-human-civilisation-GCF-Oxford-2015.pdf

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Pasquale, F. (2019). A rule of persons, not machines: The limits of legal automation. *George Washington Law Review*, *87*, 1–55. https://www.gwlr.org/a-rule-of-persons-not-machines-the-limits-of-legal-automation/

Peters, A. (2020). Toward international animal rights. In A. Peters (Ed.), *Studies in global animal law* (pp. 109–120). Springer. https://doi.org/10.1007/978-3-662-60756-5_10

Pejovic, C. (2001). Civil law and common law: Two different paths leading to the same goal. *Victoria University Wellington Law Review*, *32*, 817–841.

Pigou, A. C. (2013). *The economics of welfare*. Palgrave Macmillan.

Pillai, M. (2019, August 16). *China now has AI-powered judges*. Radii. https://radiichina.com/china-now-has-ai-powered-robot-judges/

Pinker, S. (2012). *The better angels of our nature: Why violence has declined*. Penguin Group USA.

Pinker, S. (2018). *Enlightenment now: The case for reason, science, humanism, and progress*. Penguin.

Pogge, T. W. (2002). Moral universalism and global economic justice. *Politics, Philosophy and Economics*, *1*(1), 29–58. https://doi.org/10.1177/1470594X02001001002

Polyakova, A., & Meserole, C. (2019, August). *Exporting digital authoritarianism. The Russian and Chinese models*. Brookings. https://www.brookings.edu/research/exporting-digital-authoritarianism/

Pop, V. (2008). *Who owns the moon? Extraterrestrial aspects of land and mineral resources ownership* (Vol. 4). Springer Science & Business Media.

Porsdam Mann, S., Sun, R., & Hermerén, G. (2019) A framework for the ethical assessment of chimeric animal research involving human neural tissue. *BMC Medical Ethics*, *20*, Article 10. https://doi.org/10.1186/s12910-019-0345-2

Posner, R. A. (1973). Economic analysis of law. Boston: Little, Brown and Company.

Posner, R. A. (1987). The law and economics movement. *The American Economic Review*, *77*(2), 1–13. https://www.jstor.org/stable/1805421

Posner, R. A. (1993). Legal scholarship today. *Stanford Law Review*, *45*, 1647–1658. https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=3074&context=journal_articles

Posner, R. A. (1998). Rational choice, behavioral economics, and the law. *Stanford Law Review*, *50*(5), 1551–1575. https://chicagounbound.uchicago.edu/journal_articles/1880/

Pray, L., Relman, D. A., & Choffnes, E. R. (Eds.). (2011). *The science and applications of synthetic and systems biology: Workshop summary*. National Academies Press.

Prunkl, C., & Whittlestone, J. (2020, February). 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* (pp. 138–143). https://doi.org/10.1145/3375627.3375803

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). *Dataset shift in machine learning*. The MIT Press.

Quint, M., Delker, C., Franklin, K. A., Wigge, P. A., Halliday, K. J., & van Zanten, M. (2016). Molecular and genetic control of plant thermomorphogenesis. *Nature Plants*, *2*(1), 1–9. https://doi.org/10.1038/nplants.2015.190

Ram, N. (2017). Science as speech. *Iowa Law Review*, *103*(3), 1187–1238. https://ilr.law.uiowa.edu/print/volume-102-issue-3/science-as-speech/

Ramsey, F. P. (1928). A mathematical theory of saving. *The Economic Journal*, *38*(152), 543–559. https://doi.org/10.2307/2224098

Rappert, B. (2014). Why has not there been more research of concern? *Frontiers in Public Health*, *2*, 74. https://dx.doi.org/10.3389%2Ffpubh.2014.00074

Re, R. M., & Solow-Niederman, A. (2019). Developing artificially intelligent justice. *Stanford Technology Law Review*, *22*(2), 242–289.

Reddy, R. (2020, January 22). *Enshrining animal sentience into law: Global developments and implications*. ABA. https://www.americanbar.org/groups/tort_trial_insurance_practice/publications/committee-newsletters/enshrining_animal_sentience_into_law/

Regan, T. (2001). The radical egalitarian case for animal rights. *Environmental Ethics*, *5*, 82–90. https://philpapers.org/rec/REGTRE

Resnik D. B. (2013). H5N1 avian flu research and the ethics of knowledge. *The Hastings Center report*, *43*(2), 22–33. https://doi.org/10.1002/hast.143

Revill, J. (2017). Past as prologue? The risk of adoption of chemical and biological weapons by non-state actors in the EU. *European Journal of Risk Regulation*, *8*(4), 626–642. https://doi.org/10.1017/err.2017.35

Ribeiro, B., & Shapira, P. (2019). Anticipating governance challenges in synthetic biology: Insights from biosynthetic menthol. *Technological Forecasting and Social Change*, *139*, 311–320. https://doi.org/10.1016/j.techfore.2018.11.020

Roberts, K. H., & Bea, R. (2001). Must accidents happen? Lessons from high-reliability organizations. *Academy of Management Perspectives*, *15*(3), 70–78. https://doi.org/10.5465/ame.2001.5229613

Roeder, O. (2017, October 17). *The supreme court is allergic to math*. FiveThirtyEight. https://fivethirtyeight.com/features/the-supreme-court-is-allergic-to-math/

Roell, M. S., & Zurbriggen, M. D. (2020). The impact of synthetic biology for future agriculture and nutrition. *Current Opinion in Biotechnology*, *61*, 102–109. https://doi.org/10.1016/j.copbio.2019.10.004

Rogelj, J., Den Elzen, M., Höhne, N., Fransen, T., Fekete, H., Winkler, H., Schaeffer, R., Sha, F., Riahi, K., & Meinshausen, M. (2016). Paris Agreement climate proposals need a boost to keep warming well below 2 C. *Nature*, *534*(7609), 631–639. https://doi.org/10.1038/nature18307

Rooke, J. (2013). Synthetic biology as a source of global health innovation. *Systems and Synthetic Biology*, *7*(3), 67–72. https://dx.doi.org/10.1007%2Fs11693-013-9117-3

Rosling, H. (2019). *Factfulness*. Flammarion.

Rose, M. (2018). *Zukünftige generationen in der heutigen demokratie: theorie und praxis der proxy-repräsentation*. Springer. https://www.springer.com/de/book/9783658188450

Ross, J. (2006). Rejecting ethical deflationism. *Ethics*, *116*(4), 742–768.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin Random House.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: a modern approach*. Pearson.

Ryan, J. E. (2002). The limited influence of social science evidence in modern desegregation cases. *North Carolina Law Review*, *81*, 1659–1702. https://scholarship.law.unc.edu/nclr/vol81/iss4/8/

Rylott, E. L., & Bruce, N. C. (2020). How synthetic biology can help bioremediation. *Current Opinion in Chemical Biology*, *58*, 86–95. https://doi.org/10.1016/j.cbpa.2020.07.004

Sachs, J., & Malaney, P. (2002). The economic and social burden of malaria. *Nature*, *415*(6872), 680–685.

Sagan, C., & Newman, W. I. (1983). The solipsist approach to extraterrestrial intelligence. *Quarterly Journal of the Royal Astronomical Society*, *24*, 113–121.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB neuroscience*, *11*(2), 88–95. https://doi.org/10.1080/21507740.2020.1740350

Sandberg, A., & Bostrom, N. (2008). *Global catastrophic risks survey* [Technical report]. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/reports/2008-1.pdf

Sandberg, A., & Nelson, C. (2020, June 10). Who should we fear more: Biohackers, disgruntled postdocs, or bad governments? A simple risk chain model of biorisk. *Health Security*, *18*(3), 155–163. https://doi.org/10.1089/hs.2019.0115

Santosuosso, A., Sellaroli, V., & Fabio, E. (2007). What constitutional protection for freedom of scientific research? *Journal of Medical Ethics*, *33*(6), 342–344. http://dx.doi.org/10.1136/jme.2007.020594

Schell, J. (2000). *The fate of the Earth and the abolition.* Stanford University Press.

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harvard Journal of Law and Technology, 29*(2), 353–400. http://jolt.law.harvard.edu/articles/pdf/v29/29HarvJLTech353.pdf

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., . . . Lombard, M. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, *358*(6363), 652–655. https://doi.org/10.1126/science.aao6266

Schmidt, M., & Pei, L. (2010). Synthetic toxicology: Where engineering meets biology and toxicology. *Toxicological Sciences*, *120*, S204–S224. https://doi.org/10.1093/toxsci/kfq339

Schoch-Spana, M., Cicero, A., Adalja, A., Gronvall, G., Kirk Sell, T., Meyer, D., Nuzzo, J. B., Ravi, S., Shearer, M. P., Toner, E., Watson, C., Watson, M., & Inglesby, T. (2017). Global catastrophic biological risks: Toward a working definition. *Health Security*, *15*(4), 323–328. https://doi.org/10.1089/hs.2017.0038

Schrogl, K. U., Hays, P. L., Robinson, J., Moura, D., & Giannopapa, C. (Eds.). (2020). *Handbook of space security*. Springer.

Schubert, S., Caviola, L., & Faber, N. S. (2019). The psychology of existential risk: Moral judgments about human extinction. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-50145-9

Schuett, J. (2019). *A legal definition of AI.* arXiv. https://arxiv.org/abs/1909.01095

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy, 39*, 98–119. https://doi.org/10.1111/misp.12032

Scoles, S. (2020, August 5). *How do we know if a virus is bioengineered?* Medium Future Human. https://futurehuman.medium.com/how-do-we-know-if-a-virus-is-bioengineered-541ff6f8a48f

Scrivner, S. (2018). Regulations and resolutions: Does the BWC prevent terrorists from accessing bioweapons? *Journal of Biosecurity, Biosafety, and Biodefense Law*, *9*(1), 1–5.

Sebo, J. (2018). The moral problem of other minds. *The Harvard Review of Philosophy*, *25*, 51–70. https://doi.org/10.5840/harvardreview20185913

Sebo, J. (2020). *All we owe to animals.* Aeon. https://aeon.co/essays/we-cant-stand-by-as-animals-suffer-and-die-in-their-billions

Sebo, J. (2021a). *Animals and climate change*. In M. Budolfson, T. McPherson, & D. Plunkett (Eds.), *Philosophy and climate change* [Forthcoming]. Oxford University Press.

Sebo, J. (2021b). *Animal ethics in a human world* [Forthcoming]. Oxford University Press.

Secretariat of the Convention on Biological Diversity. (2015). *Synthetic biology*. CBD Technical Series No. 82. https://www.cbd.int/doc/publications/cbd-ts-82-en.pdf

Sen, A. (2000). The discipline of cost-benefit analysis. *Journal of Legal Studies*, *29*(S2), 931–952.

Shah, R. (2018, December 3). *Coherence arguments do not imply goal-directed behavior* [Online forum post]. AI Alignment Forum. https://www.alignmentforum.org/s/4dHMdK5TLN6xcqtyc/p/NxF5G6CJiof6cemTw

Shapiro, F. R. (1995). The most-cited law review articles revisited. *Chicago-Kent Law Review*, *71*, 751–779. https://scholarship.kentlaw.iit.edu/cgi/viewcontent.cgi?article=3037&context=cklawreview

Shapiro, F. R. (2000a). The most-cited law reviews. *Journal of Legal Studies*, *29*(1), Pt. 2. https://ssrn.com/abstract=231574

Shapiro, F. R. (2000b). The most-cited legal scholars. *Journal of Legal Studies*, *29*(S1), 409–426. https://doi:10.1086/468080

Shapiro, F. R., & Pearse, M. (2012). The most-cited law review articles of all time. *Michigan Law Review*, 1483-1520. https://www.jstor.org/stable/23217010

Shavell, S. (2009). *Economic analysis of accident law*. Harvard University Press. https://doi.org/10.4159/9780674043510

Sheldon, K. M., & Krieger, L. S. (2004). Does legal education have undermining effects on law students? Evaluating changes in motivation, values, and well-being. *Behavioral Sciences and the Law*, *22*(2), 261–286. https://doi.org/10.1002/bsl.582

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., von der Heydt, A. S., Knutti, R., . . . Zelinka, M. D. (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678. https://doi.org/10.1029/2019RG000678

Shevlane, T., & Dafoe, A. (2020). The offense-defense balance of scientific knowledge: Does publishing AI research reduce misuse? *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 173–179. https://doi.org/10.1145/3375627.3375815

Shore, D. (2020). Divergence and convergence of royalty determinations between compulsory licensing under the TRIPS Agreement and ongoing royalties as an equitable remedy. *American Journal of Law and Medicine*, *46*, 55–88. https://doi.org/10.1177/0098858820919553

Shulman, C. (2020, October 15). *What do historical statistics teach us about the accidental release of pandemic bioweapons?* [Blog post]. Reflective Disequilibrium. https://reflectivedisequilibrium.blogspot.com/2020/10/what-do-historical-statistics-teach-us.html

Sidgwick, H. (2019). *The methods of ethics*. Good Press.

Simester, A. P., & von Hirsch, A. (2009). Remote harms and non-constitutive crimes. *Criminal Justice Ethics*, *28*(1), 89–107. https://doi.org/10.1080/07311290902831441

Simkovic, M., & McIntyre, F. (2014). The economic value of a law degree. *Journal of Legal Studies*, *43*(2), 249–289. https://doi.org/10.1086/677921

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161–176. http://innovbfa.viabloga.com/files/Herbert_Simon___theories_of_bounded_rationality___1972.pdf

Singer, P. (1975). *Animal liberation*. Random House.

Singer, P. (2011). *The expanding circle: Ethics, evolution, and moral progress*. Princeton University Press.

Sinnott-Armstrong, W. (2019). *Consequentialism*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/consequentialism/

Sirakaya, A. (2019). Balanced options for access and benefit-sharing: Stakeholder insights on provider country legislation. *Frontiers in Plant Science*, *10*, 1175. https://doi.org/10.3389/fpls.2019.01175

Sirleaf, M. (2018a). Ebola does not fall from the sky: Structural violence & international responsibility. *Vanderbilt Journal of Transnational Law*, *51*(2), 477–554.

Sirleaf, M. (2018b). Responsibility for epidemics. *Texas Law Review*, *97*(2), 285–351. https://texaslawreview.org/responsibility-for-epidemics/

Slovic, P. (2010). If I look at the mass I will never act: Psychic numbing and genocide. In S. Roeser (Ed.), *Emotions and risky technologies* (pp. 37–59). Springer.

Slovic, P., Zionts, D., Woods, A. K., Goodman, R., & Jinks, D. (2013). Psychic numbing and mass atrocity. In E. Shafir (Ed.), *The behavioral foundations of public policy* (pp. 126–142). New Jersey: Princeton University Press.

Smith, H. M. (2010). Subjective rightness. *Social Philosophy and Policy*, *27*(2), 64–110. https://doi.org/10.1017/S0265052509990161

Smith, K. K. (2012). *Governing animals: Animal welfare and the liberal state*. Oxford University Press.

Smuha, N. A. (2019). *From a 'Race to AI' to a 'Race to AI Regulation': Regulatory competition for artificial intelligence*. SSRN. http://dx.doi.org/10.2139/ssrn.3501410

Snyder-Beattie, A. E., Ord, T., & Bonsall, M. B. (2019). An upper bound for the background rate of human extinction. *Scientific Reports*, *9*(1), 1–9. https://dx.doi.org/10.1038%2Fs41598-019-47540-7

Soares, N. (2016a). *The value learning problem* [Technical report]. Machine Intelligence Research Institute. https://intelligence.org/files/ValueLearningProblem.pdf

Soares, N. (2016b, July 23). *Submission to the OSTP on AI outcomes*. Machine Intelligence Research Institute. https://intelligence.org/2016/07/23/ostp

Soares, N., & Fallenstein, B. (2014). Agent foundations for aligning machine intelligence with human interests: A technical research agenda. In V. Callaghan, J. Miller, R. Yampolskiy, & S. Armstrong (Eds.), *The technological singularity: Managing the Journey* (pp. 103–125). Springer. https://doi.org/10.1007/978-3-662-54033-6_5

Solow, R. M. (1974). The economics of resources or the resources of economics. In C. Gopalakrishnan (Ed.), *Classic papers in natural resource economics* (pp. 257–276). London: Palgrave Macmillan. https://www.palgrave.com/gp/book/9780333777633

Sommers, R. (2020). Commonsense consent. *Yale Law Journal*, *129*(8), 2232–2324. https://www.yalelawjournal.org/article/commonsense-consent

Stafforini, P. (2016, December 3). *Paul Christiano on cause prioritization* [Blog post]. Pablo's miscellany. http://www.stafforini.com/blog/paul-christiano-on-cause-prioritization

Stauffer, M. (2019). Tactical models to improve institutional decision-making [Blog post]. Effective Altruism Geneva. https://eageneva.org/blog/2019/2/27/tactical-models-to-improve-institutional-decision-making

Stawasz, A. (2020, July 4). *Why and how to value nonhuman animals in cost-benefit analyses*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3643473

Steele, K., & Stefánsson, H. (2020). *Decision theory*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/decision-theory

Stephanopoulos, N. O., & McGhee, E. M. (2014). Partisan gerrymandering and the efficiency gap. *University of Chicago Law Review*, *82*(2), 831–900.

Stern, N. (2008). The economics of climate change. *American Economic Review*, *98*(2), 1–37. https://doi.org/10.1017/CBO9780511817434

Stirling, A., Hayes, K. R., & Delborne, J. (2018). Towards inclusive social appraisal: Risk, participation and democracy in governance of synthetic biology. *BMC Proceedings*, *12*(8), 15. https://doi.org/10.1186/s12919-018-0111-3

Stucki, S., & Winter, C. K. (2019, June). *Of chicks and men: Anmerkungen zum BVerwG-Urteil über die Tötung männlicher Küken*. Verfassungsblog. https://verfassungsblog.de/of-chicks-and-men/

Sullivan, K. M. (1992). Foreword: The justices of rules and standards. *Harvard Law Review*, *106*, 22–123.

Sunstein, C. R. (1996). Social norms and social roles. *Columbia Law Review*, *96*(4), 903–968. https://doi.org/10.2307/1123430

Sunstein, C. R. (2018). *The cost-benefit revolution*. MIT Press.

Sunstein, C. R. (2020). Maximin. *Yale Journal on Regulation*, *37*(3), 940–980.

Sutton, V. (2005). A multidisciplinary approach to an ethic of biodefense and bioterrorism. *Journal of Law, Medicine and Ethics*, *33*(2), 310–322. https://doi.org/10.1111/j.1748-720X.2005.tb00496.x

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). *Intriguing properties of neural networks*. arXiv. https://arxiv.org/abs/1312.6199 v4

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022

Tarsney, C. *The epistemic challenge to longtermism* [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/Working-paper-10-Christian-Tarsney.pdf

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). *Alignment for advanced machine learning systems*. Machine Intelligence Research Institute. https://intelligence.org/files/AlignmentMachineLearning.pdf

Tetley, W. (1999). Mixed jurisdictions: Common law v. civil law (codified and uncodified). *Louisiana Law Review*, *60*(3), 677–738. https://digitalcommons.law.lsu.edu/lalrev/vol60/iss3/2/

The Association of American Law Schools & Gallup. (2018). *Before the JD: Undergraduate views on law school*. https://www.aals.org/research/bjd/

Thomas, J. (2011). From swine flu to smallpox: Government compensation programs for vaccines and terrorism. *Journal of Biosecurity, Biosafety and Biodefense Law*, *1*(1). https://doi.org/10.2202/2154-3186.1005

Thomas, J. (2016). In defense of philosophy: a review of Nick Bostrom, Superintelligence: Paths, Dangers, Strategies. *Journal of Experimental and Theoretical Artificial Intelligence*, *28*(6), 1089–1094. https://doi.org/10.1080/0952813X.2015.1055829

Thomas, T. (2018). Some possibilities in population axiology. *Mind*, *127*(507), 807–832. https://doi.org/10.1093/mind/fzx047

Thompson, P. B. (2020). Philosophical ethics and the improvement of farmed animal lives. *Animal Frontiers*, *10*(1), 21–28. https://doi.org/10.1093/af/vfz054

Thorn, P. D. (2015). Nick Bostrom: Superintelligence: Paths, Dangers, Strategies. *Minds and Machines, 25*(3), 285–289. https://doi.org/10.1007/s11023-015-9377-7

Thorstad, D., & Mogensen, A. (2020). Heuristics for clueless agents: How to get away with ignoring what matters most in ordinary decision-making [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/David-Thorstad-Andreas-Mogensen-Heuristics-for-clueless-agents.pdf

Tiedemann, P. (2020). *Philosophical foundation of human rights*. Springer. https://doi.org/10.1007/978-3-030-42262-2_20

Todd, B. (2017a, October). *Introducing longtermism*. 80,000 Hours. https://80000hours.org/articles/future-generations

Todd, B. (2017b, October). *The case for reducing extinction risk*. 80,000 Hours. https://80000hours.org/articles/extinction-risk/

Todd, B. (2019, May 7). *80,000 hours annual review—December 2018*. 80,000 Hours. https://80000hours.org/2019/05/annual-review-dec-2018

Todd, B. (2020a, August) *The emerging school of patient longtermism*. 80,000 Hours. https://80000hours.org/2020/08/the-emerging-school-of-patient-longtermism/

Todd, B. (2020b, August). *Why I've come to think global priorities research is more important than I thought*. 80,000 Hours. https://80000hours.org/2020/08/global-priorities-research-update/

Tomasik, B. (2014). *Do artificial reinforcement-learning agents matter morally?* arXiv. https://arxiv.org/abs/1410.8233

Tomasik, B. (2015a, December). *Charity cost-effectiveness in an uncertain world*. Center on Long-Term Risk. https://longtermrisk.org/charity-cost-effectiveness-in-an-uncertain-w orld

Tomasik, B. (2015b). *The importance of wild-animal suffering*. Center on Long-Term Risk. https://doi.org/10.7358/rela-2015-002-toma

Tomasik, B. (2017). *Machine sentience and robot rights* [Blog post]. Essays on Reducing Suffering. https://reducing-suffering.org/machine-sentience-and-robot-rights

Tomasik, B. (2018, December 13). *Astronomical suffering from slightly misaligned artificial intelligence*. Essays on Reducing Suffering. https://reducing-suffering.org/near-miss

Tomasik, B. (2019a, August 7). *How many wild animals are there?* [Blog post]. Essays on Reducing Suffering. https://reducing-suffering.org/how-many-wild-animals-are-there

Tomasik, B. (2019b, July 2). *Risks of astronomical future suffering*. Center on Long-Term Risk. https://longtermrisk.org/risks-of-astronomical-future-suffering

Tonn, B., & Stiefel, D. (2013). Evaluating methods for estimating existential risks. Risk Analysis, *33*(10), 1772–1787. https://doi.org/10.1111/risa.12039

Torres, P. (2020a). *International criminal law and the future of humanity: A theory of the crime of omnicide* [Unpublished manuscript]. https://www.xriskology.com/publications

Trajtenberg, M. (2018). *AI as the next GPT: A political-economy perspective* [Working paper]. National Bureau of Economic Research. https://doi.org/10.3386/w24245

Trammell, P. (2020). *Patience and philanthropy*. Global Priorities Institute, University of Oxford. https://philiptrammell.com/static/PatienceAndPhilanthropy.pdf

Trammel, P., & Korinek, A. (2020, October). *Economic growth under transformative AI: A guide to the vast range of possibilities for output growth, wages, and the labor share* [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/philip-trammell-and-anton-korinek-economic-growth-under-transformative-ai

Tribe, L. H., & Gudridge, P. O. (2003). The anti-emergency constitution. *Yale Law Journal, 113*, 1801–1870. https://repository.law.miami.edu/cgi/viewcontent.cgi?httpsredir=1&article=1333&context=fac_articles

Tronchetti, F. (2009). *The exploitation of natural resources of the Moon and other celestial bodies: a proposal for a legal regime (Vol. 4).* Martinus Nijhoff Publishers.

Tronchetti, F. (2013). *Fundamentals of space law and policy*. Springer.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2020). *The ethics of algorithms: Key problems and solutions*. SSRN. http://dx.doi.org/10.2139/ssrn.3662302

Turner, J. (2018). *Robot rules: Regulating artificial intelligence*. Springer.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tyagi, A., Kumar, A., Aparna, S. V., Mallappa, R. H., Grover, S., & Batish, V. K. (2016). Synthetic biology: applications in the food sector. *Critical Reviews in Food Science and Nutrition*, *56*(11), 1777–1789. https://doi.org/10.1080/10408398.2013.782534

U.S. Department of Health and Human Services. (2020). *Screening framework guidance for providers of synthetic double-stranded DNA*. Public Health and Emergency. https://www.phe.gov/Preparedness/legal/guidance/syndna/Documents/syndna-guidance.pdf

U.S. House Judiciary Subcommittee on Antitrust, Commercial and Administrative Law. (2020). *Investigation of competition in digital markets*. U.S. House of Representatives. https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf

Unger, P. K. (1996). *Living high and letting die: Our illusion of innocence*. Oxford University Press, USA.

United Nations Conference on Trade and Development (2019). *Synthetic biology and its potential implications for biotrade and access and benefit-sharing* (UNCTAD/DITC/TED/INF/2019/12). https://unctad.org/system/files/official-document/ditctedinf2019d12_en.pdf

United Nations Office for Outer Space Affairs. (2007). *Space debris mitigation guidelines of the committee on the peaceful uses of outer space*. https://www.unoosa.org/pdf/publications/st_space_49E.pdf

United Nations Office for Outer Space Affairs. (2020a). *Treaty on principles governing the activities of states in the exploration and use of outer space, including the moon and other celestial bodies*. http://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/introouterspacetreaty.html

United Nations Office for Outer Space Affairs. (2020b). *FAQs*. https://www.unoosa.org/oosa/en/informationfor/faqs.html

United Nations Office for Outer Space Affairs. (2020c). *Long-term sustainability of outer space activities*. https://www.unoosa.org/oosa/en/informationfor/faqs.html

Van Houten, J., & Fleming, D. O. (1993). Comparative analysis of current US and EC biosafety regulations and their impact on the industry. *Journal of Industrial Microbiology 11*, 209–215. https://doi.org/10.1007/BF01569593

Varian, H. R. (2001). High-technology industries and market structure. *Proceedings – Economic Policy Symposium – Jackson Hole, Federal Reserve Bank of Kansas City*, 65–101. Archived at https://perma.cc/DZ2B-E7GT

Villasenor, J. (2020, July 31). *Soft law as a complement to AI regulation*. Brookings Institution. https://www.brookings.edu/research/soft-law-as-a-complement-to-ai-regulation

von der Dunk, F. G. (2009). Studies in space law. In F. G. von der Dunk (Ed.), *National Space Legislation in Europe* (pp. 381–381). Leiden: Brill Nijhoff.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI and Society*, *22*(4), 565–582. https://doi.org/10.1007/s00146-007-0099-0

Wallach, W., Saner, M., & Marchant, G. (2018). *Beyond Cost-Benefit Analysis in the Governance of Synthetic Biology*. Hastings Center Report, *48*, S70–S77. https://onlinelibrary.wiley.com/doi/full/10.1002/hast.822

Warmbrod, K. L., Kobokovich, A., West, R., Ray, G., Trotochaud, M., & Montague, M. (2020, May 18). *Gene drives: Pursuing opportunities, minimizing risk*. Johns Hopkins Bloomberg School of Public Health, Center for Health Security. https://www.centerforhealthsecurity.org/our-work/publications/gene-drives-pursuing-opportunities-minimizing-risk

Watson, A. (1991). *Roman law & comparative law*. University of Georgia Press.

Watson, C. (2018, August 9). Assessing global catastrophic biological risks [Talk]. https://www.effectivealtruism.org/articles/ea-global-2018-assessing-gcbr/

Way, J. C., Collins, J. J., Keasling, J. D., & Silver, P. A. (2014). Integrating biological redesign: where synthetic biology came from and where it needs to go. *Cell*, *157*(1), 151–161. https://doi.org/10.1016/j.cell.2014.02.039

Weinrib, E. J. (1980). The case for a duty to rescue. *The Yale Law Journal*, *90*(2), 247–293. https://doi.org/10.2307/795987

Weiss Evans, S. (2018, October 23). *The use and abuse of science and technology: rethinking dual-use* [Blog post]. Sam Weiss Evans' Research. https://evansresearch.org/2018/10/the-use-and-abuse-of-science-and-technology-rethinking-dual-use/

Wentworth, J. (2020, August 8). *The fusion power generator scenario* [Online forum post]. AI Alignment forum. https://www.alignmentforum.org/posts/2NaAhMPGub8F2Pbr7/the-fusion-power-generator-scenario

Whittlestone, J. (2017a, September). *Improving institutional decision-making*. 80,000 Hours. https://80000hours.org/problem-profiles/improving-institutional-decision-making/#improving-decision-making-could-help-us-to-solve-almost-all-other-problems

Whittlestone, J. (2017b, November). *The long-term future* [Online forum post]. Effective Altruism. https://www.effectivealtruism.org/articles/cause-profile-long-run-future/

Whittlestone, J., & Ovadya, A. (2020). *The tension between openness and prudence in responsible AI research*. arXiv. https://arxiv.org/abs/1910.01170

Wiblin, R. (2016, January). *The Important/Neglected/Tractable framework needs to be applied with care* [Online forum post]. Effective Altruism Forum. https://forum.effectivealtruism.org/posts/74oJS32C6CZRC4Zp5/the-important-neglected-tractable-framework-needs-to-be

Wiblin, R. (2017a, September). *Why the long-term future of humanity matters more than anything else, and what we should do about it* [Podcast]. 80,000 Hours. https://80000hours.org/podcast/episodes/why-the-long-run-future-matters-more-than-anything-else-and-what-we-should-do-about-it/

Wiblin, R. (2017b, March). *Positively shaping the development of artificial intelligence*. 80,000 Hours. https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence

Wiblin, R. (2019, October). *One approach to comparing global problems in terms of expected impact*. 80,000 Hours. https://80000hours.org/articles/problem-framework/

Wiblin, R., & Harris, K. (2018a). *Tackling the ethics of infinity, being clueless about the effects of our actions, and having moral empathy for intellectual adversaries, with philosopher Dr Amanda Askell* [Podcast]. 80,000 Hours. https://80000hours.org/podcast/episodes/amanda-askell-moral-empathy

Wiblin, R., & Harris, K. (2018b). Our descendants will probably see us as moral monsters. What should we do about that? [Podcast]. 80,000 Hours. https://80000hours.org/podcast/episodes/will-macaskill-moral-philosophy

Wilkinson, H. (2020, September). *In defence of fanaticism* [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/Hayden-Wilkinson_In-defence-of-fanaticism.pdf

Wilson, G. (2013). Minimizing global catastrophic and existential risks from emerging technologies through international law. *Virginia Environmental Law Journal*, *31*(2), 307–364. https://www.jstor.org/stable/44679544

Winter, C. K. (2020a). The value of behavioral economics for EU judicial decision-making. *German Law Journal*, *21*(2), 240–264. https://doi.org/10.1017/glj.2020.3

Winter, C. K. (2020b). *Towards a dual-process theory of criminalization* [Manuscript in preparation]. https://www.christophwinter.net/s/Towards-a-Dual-Process-Theory-of-Criminalization.pdf

Winter, C. K. (2021a). Exploring the challenges of artificial judicial decision-making for liberal democracy [Forthcoming]. In P. Bystranowski, P. Janik, & M. Próchnicki (Eds.), *Judicial decision-making: Integrating empirical and theoretical perspectives*. https://www.christophwinter.net/s/AI-Judiciary.pdf

Winter, C. K. (2021b). *Metamoralisches Strafrecht* [Unpublished manuscript].

Wise, S. M. (2000). *Rattling the cage: Toward legal rights for animals*. Da Capo Press. https://dacapopress.com/titles/steven-wise/rattling-the-cage/9780306824005/

Wittes, B., & Chong, J. (2014, September). *Our cyborg future: Law and policy implications*. Brookings Institution. https://www.brookings.edu/research/our-cyborg-future-law-and-policy-implications

Woodhouse, J. (2019, May). *Universal declaration of sentient rights*. Sentientism. https://www.researchgate.net/publication/338345937_Universal_Declaration_of_Sentient_Rights

World Animal Protection. (2020). *Animal protection index (API) 2020*. World Animal Protection. https://api.worldanimalprotection.org/country/usa

World Bank. (2017). *From panic and neglect to investing in health security: Financing pandemic preparedness at a national level (English)* [Report]. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/979591495652724770/from-panic-and-neglect-to-investing-in-health-security-financing-pandemic-preparedness-at-a-national-level

World Health Organization, World Intellectual Property Organization, & World Trade Organization (2013). *Promoting access to medical technologies and innovation: Intersections between public health, intellectual property and trade*. World Health Organization. https://www.who.int/phi/promoting_access_medical_innovation/en/

World Health Organization. (2016). *Human challenge trials for vaccine development: Regulatory considerations*. World Health Organization. https://www.who.int/biologicals/expert_committee/Human_challenge_Trials_IK_final.pdf

World Health Organization. (2017). *Annex 10: Human challenge trials for vaccine development: Regulatory considerations*. World Health Organization. https://www.who.int/biologicals/expert_committee/WHO_TRS_1004_web_Annex_10.pdf?ua=1

World Health Organization. (2020). *Key criteria for the ethical acceptability of COVID-19 human challenge studies*. World Health Organization. https://www.who.int/ethics/publications/key-criteria-ethical-acceptability-of-covid-19-human-challenge/en

Wurtzel, E. T., Vickers, C. E., Hanson, A. D., Millar, A. H., Cooper, M., Voss-Fels, K. P., Nickel, P. I., & Erb, T. J. (2019). Revolutionizing agriculture with synthetic biology. *Nature Plants*, *5*, 1207–1210. https://doi.org/10.1038/s41477-019-0539-0

Yassif, J. (2017). Reducing global catastrophic biological risks. *Health Security*, *15*(4), 329–330. https://dx.doi.org/10.1089%2Fhs.2017.0049

Young, M., Katell, M., & Krafft, P. M. (2019). Municipal surveillance regulation and algorithmic accountability. *Big Data and Society*, *6*(2), 1–14 https://doi.org/10.1177/2053951719868492

Yudkowsky, E. (2004). *Coherent extrapolated volition*. Machine Intelligence Research Institute. https://intelligence.org/files/CEV.pdf

Yudkowsky, E. (2008a). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308–345). Oxford University Press. https://intelligence.org/files/AIPosNegFactor.pdf

Yudkowsky, E. (2008b). Cognitive biases potentially affecting judgment of global risks. In N. Bostrom, & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 91–119). Oxford University Press.

Zhang, J., Marris, C., & Rose, N. (2011, May). *The transnational governance of synthetic biology: Scientific uncertainty, cross-borderness and the 'art' of governance*. London: BIOS (Centre for the Study of Bioscience, Biomedicine, Biotechnology and Society). http://openaccess.city.ac.uk/16098/

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

Appendix

Closely Related Areas of Existing Academic Research

As alluded to at various points so far in this agenda, legal priorities research by its very nature is an interdisciplinary affair. Below is an overview of some of the most closely related areas of existing literature that serve as particularly relevant background for the topics in this research agenda. This appendix is organized around the general academic disciplines of philosophy (A), economics (B), psychology (C), macrohistory (D) and political science (E). Within each discipline we identify both general examples of interdisciplinary research between law and that respective discipline, as well as more specific research programs/areas within those disciplines that are likely to be particularly useful for legal priorities research. Consequently, interested researchers who have a strong background in one or more of these areas are likely to be particularly good fits for legal priorities research.

A. Philosophy

Recent insights from the philosophical literature, and in particular those from moral and political philosophy, have played a central role in motivating the development of this research agenda. Overall, the discipline of philosophy is an indispensable component of prioritization research more generally, from determining the appropriate evaluative criteria on which to prioritize to evaluating potential solutions to the prioritized research questions. Moreover, within the field of philosophy there are several sub-areas that are likely to be particularly useful in addressing many of the meta- and object-level research questions presented in this agenda, including philosophical longtermism, normative decision theory, normative uncertainty, and experimental jurisprudence.

Relevant Literature on Law and Philosophy Generally

Coleman, J. L., Shapiro, S., & Himma, K. E. (Eds.). (2002). *The Oxford handbook of jurisprudence and philosophy of law*. Oxford University Press.

Hart, H. L. A. (1958). Positivism and the separation of law and morals. *Harvard Law Review*, *71*(4), 593–629. https://doi.org/10.2307/1338225

Marmor, A., & Sarch, A. (2019). *The nature of law*. Stanford Encyclopedia of Philosophy Archive. https://plato.stanford.edu/archives/spr2020/entries/lawphil-nature

Shapiro, S. J. (2011). *Legality*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674725782

Wenar, L. (2020). *Rights*. Stanford Encyclopedia of Philosophy Archive. https://plato.stanford.edu/archives/spr2020/entries/rights

RELEVANT LITERATURE ON LONGTERMISM

Beckstead, N. (2019). A brief argument for the overwhelming importance of shaping the far future. In H. Greaves & T. Pummer (Eds.), *Effective altruism: Philosophical issues*. Oxford University Press. https://doi.org/10.1093/oso/9780198841364.003.0006

Feinberg, J. (1974). The rights of animals and future generations. In W. Blackstone (Ed.), Philosophy and environmental crisis (pp. 43–68). Athens, Georgia: University of Georgia Press. http://www.animal-rights-library.com/texts-m/feinberg01.pdf

Greaves, H. (2017a). Discounting for public policy: A survey. Economics & Philosophy, 33(3), 391–439. https://doi.org/10.1017/S0266267117000062

Greaves, H. (2017c). Discounting future health. PhilPapers. https://philpapers.org/rec/GREDFH

Greaves, H., & MacAskill, W. (2019). The case for strong longtermism [Working paper]. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/2020/Greaves_MacAskill_strong_longtermism.pdf

Greaves, H., Mogensen, A. L., & MacAskill, W. (2021). Longtermism for risk-averse altruists [Manuscript in preparation].

John, T. M., & MacAskill, W. (2021). Longtermist institutional reform [Forthcoming]. In N. Cargill & T. M. John (Eds.), The Long View. https://philpapers.org/rec/JOHLIR

Thomas, T. (2019). The asymmetry, uncertainty, and the long term. Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/teruji-thomas-the-asymmetry-uncertainty-and-the-long-term/

RELEVANT LITERATURE ON NORMATIVE UNCERTAINTY AND DECISION THEORY

Askell, A. (2019). Evidence neutrality and the moral value of information. In H. Greaves & T. Pummer (Eds.), *Effective altruism: Philosophical issues*. Oxford University Press.

Barry, C., & Tomlin, P. (2019). Moral Uncertainty and the Criminal Law. In L. Alexander & K. K. Ferzan (Eds.), *The Palgrave handbook of applied ethics and the criminal law* (pp. 445–467). New York: Palgrave. https://www.palgrave.com/gp/book/9783030228101

Bostrom, N. (2009). Pascal's mugging. *Analysis*, *69*(3), 443–445. https://www.jstor.org/stable/40607655

Buchak, L. (2013). *Risk and rationality*. Oxford University Press.

Gustafsson, J. E., & Torpman, O. (2014). In defence of my favourite theory. *Pacific Philosophical Quarterly*, *95*(2), 159–174. https://doi.org/10.26556/jesp.v12i2.223

Lockhart, T. (2000). *Moral uncertainty and its consequences*. Oxford University Press.

MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral uncertainty*. Oxford University Press.

Mogensen, A. L. (2021). Maximal cluelessness. *The Philosophical Quarterly*, *71*(1), 141–162. https://doi.org/10.1093/pq/pqaa021

Sepielli, A. (2013). Moral uncertainty and the principle of equity among moral theories. *Philosophy and Phenomenological Research*, *86*(3), 580–589. https://doi.org/10.1111/j.1933-1592.2011.00554.x

Tarsney, C. (2018). Moral uncertainty for deontologists. *Ethical Theory and Moral Practice*, *21*(3), 505–520. https://doi.org/10.1007/s10677-018-9924-4

Winter, C. K. (2021b). *Metamoralisches Strafrecht* [Unpublished manuscript].

RELEVANT LITERATURE ON EXPERIMENTAL JURISPRUDENCE

See also Appendix C: Psychology

Donelson, R., & Hannikainen, I. (2020). Fuller and the folk: The inner morality of law revisited. In T. Lombrozo, J. Knobe, & S. Nichols (Eds.), *Oxford studies in experimental philosophy* (Vol. 3). Oxford University Press.

Klapper, S., Schmidt, S. & Tarantola, T. (2020). *Ordinary meaning from ordinary people*. SSRN. https://ssrn.com/abstract=3593917

Knobe, J., & Nichols, S. (Eds.). (2013). *Experimental Philosophy* (Vol. 2). Oxford University Press.

Kneer, M., & Bourgeois-Gironde, S. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition*, *169*, 139–146. https://doi.org/10.1093/pq/pqaa021

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003

Sommers, R. (2020). Commonsense consent. *Yale Law Journal*, *129*(8), 2232–2605.

Struchiner, N., de Almaida G., & Hainniken, I. R. (2020). Legal decision-making and the abstract/concrete paradox. *Cognition*, *205*, 104421. https://doi.org/10.1016/j.cognition.2020.104421

Tobia, K. P. (2019). *Essays in experimental jurisprudence* [Doctoral dissertation]. Yale University.

Tobia, K. P. (2020a). *Experimental jurisprudence*. SSRN. http://dx.doi.org/10.2139/ssrn.3680107

Tobia, K. P. (2020b). Testing ordinary meaning. *Harvard Law Review*, *134*, 726–806. https://harvardlawreview.org/2020/12/testing-ordinary-meaning/

Winter, C. K. (2021b). *Metamoralisches Strafrecht* [Unpublished manuscript]

B. ECONOMICS

Since the mid 20th century, economic analysis of law has played an increasingly prominent role in judicial decisions (Posner, 1973) as well as in legal academic research (Jolls, Sunstein & Thaler, 1998; Kaplow & Shavell, 2002; Shavell, 2009). Although the traditional goals of law-and-economics research have not been squarely aligned with longtermism or legal priorities research *per se*, the tools of law-and-economics researchers, as well as of the field of economics in general, are nonetheless useful in evaluating cause areas, research questions, and individual solutions within legal priorities research.

RELEVANT LITERATURE ON LAW AND ECONOMICS GENERALLY

Garoupa, N. (2014) Economic theory of criminal behavior. In G. Bruinsma & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice*. Springer. https://doi.org/10.1007/978-1-4614-5690-2_409

Kaplow, L., & Shavell, S. (2001). Fairness versus welfare, *Harvard Law Review*, *114*, 961–1388.

Kornhauser, L. (2017). *The economic analysis of law*. Stanford Encyclopedia of Philosophy Archive. https://plato.stanford.edu/archives/fall2017/entries/legal-econanalysis/

Kornhauser, L. A. (1984). The great image of authority. *Stanford Law Review*, *36*, 349–389 https://doi.org/10.2307/1228686

Listokin, Y. (2017). Law and macroeconomics: The law and economics of recessions. *Yale Journal on Regulation*, *34*, 791–856. https://digitalcommons.law.yale.edu/yjreg/vol34/iss3/3/

Mackaay, E. (1999). *History of law and economics*. Encyclopedia of Law & Economics. https://reference.findlaw.com/lawandeconomics/0200-history-of-law-and-economics.pdf

Mackaay, E. (2014). *Law and economics for civil law systems*. Edward Elgar Publishing.

Polinsky, A. M. (2018). *An introduction to law and economics*. Wolters Kluwer Law & Business.

Posner, R. A. (1979). Utilitarianism, economics and legal theory. *Journal of Legal Studies*, *8*, 103–140. https://doi.org/10.1086/467603

Posner, R. A. (2014). *Economic analysis of law*. Wolters Kluwer Law & Business.

Samuels, W. J. (1974). The Coase theorem and the study of law and economics. *Natural Resources Journal*, *14*, 1–33.

Shavell, S. (2009). *Foundations of economic analysis of law*. Harvard University Press.

RELEVANT LITERATURE ON BEHAVIORAL LAW AND ECONOMICS

Alemanno, A., & Sibony, A. L. (Eds.). (2015). *Nudge and the law: A European perspective*. Bloomsbury Publishing.

Engel, C. (2007). *Institutions for intuitive man*. MPI Collective Goods Preprint. https://dx.doi.org/10.2139/ssrn.1015765

Engel, C. (2013). *Behavioral law and economics: Empirical methods*. MPI Collective Goods Preprint. https://dx.doi.org/10.2139/ssrn.2207921

Hoffman, E., & Spitzer, M. (1985). Experimental law and economics: An introduction. *Columbia Law Review*, *85*(5), 991–1036. https://doi.org/10.2307/1122460

Jolls, C. (2007). *Behavioral law and economics*. National Bureau of Economic Research. https://www.nber.org/papers/w12879

Jolls, C., & Sunstein, C. R. (2006). Debiasing through law. *The Journal of Legal Studies*, *35*(1), 199–242. https://doi.org/10.1086/500096

Jolls, C., Sunstein, C. R., & Thaler, R. (1998). A behavioral approach to law and economics. *Stanford Law Review*, *50*, 1471–1550. https://doi.org/10.2307/1229304

Korobkin, R. B., & Ulen, T. S. (2000). Law and behavioral science: Removing the rationality assumption from law and economics. *California Law Review*, *88*, 1051–1144.

Mathis, K. (Ed.). (2015). *European perspectives on behavioural law and economics* (Vol. 2). Springer.

Rachlinski, J. J. (2011). The psychological foundations of behavioral law and economics. *University of Illinois Law Review*, *2011*(5), 1675–1696.

Rachlinski, J. J., Guthrie, C., & Wistrich, A. J. (2011). Probable cause, probability, and hindsight. *Journal of Empirical Legal Studies*, *8*, 72–98. https://doi.org/10.1111/j.1740-1461.2011.01230.x

Schmid, A. A. (1994). Institutional law and economics. *European Journal of Law and Economics*, *1*, 33–51. https://doi.org/10.1007/BF01540990

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Ulen, T. S. (2013). Behavioral law and economics: Law, policy, and science. *Supreme Court Economic Review*, *21*(1), 5–42. https://doi.org/10.1086/675264

Winter, C. K. (2020a). The value of behavioral economics for EU judicial decision-making. *German Law Journal*, *21*(2), 240–264. https://doi.org/10.1017/glj.2020.3

Zamir, E., & Teichman, D. (Eds.) (2014). *The Oxford handbook of behavioral economics and the law*. Oxford University Press.

RELEVANT LITERATURE ON COMPARATIVE LAW AND ECONOMICS

Caterina, R. (2012). Comparative law and economics. In J. M. Smits (Ed.), *Elgar encyclopedia of comparative law* (2nd Ed.). Edward Elgar Publishing.

Cooter, R. D., & Ginsburg, T. (1996). Comparative judicial discretion: An empirical test of economic models. *International Review of Law and Economics*, *16*(3), 295–313. https://doi.org/10.1016/0144-8188(96)00018-X

Faust, F. (2008). Comparative law and economic analysis of law. In M. Reimann & R. Zimmermann (Eds.), *The Oxford handbook of comparative law*. Oxford University Press.

Mattei, U. (1997). *Comparative law and economics*. University of Michigan Press. https://doi.org/10.3998/mpub.11209

Mattei, U., & Pardolesi, R. (1991). Law and economics in civil law countries: A comparative approach. *International Review of Law and Economics*, *11*(3), 265–275. https://doi.org/10.1016/0144-8188(91)90004-W

RELEVANT LITERATURE ON LAW, HAPPINESS, AND WELL-BEING

Bagaric, M., & McConvill, J. (2005). Goodbye justice, hello happiness: Welcoming positive psychology to the law. *Deakin Law Review*, *10*, 1–26. https://doi.org/10.21153/dlr2005vol10no1art265

Bandes, S. A., & Blumenthal, J. A. (2012). Emotion and the law. *Annual Review of Law and Social Science*, *8*, 161–181. https://doi.org/10.1146/annurev-lawsocsci-102811-173825

Bronsteen, J., Buccafusco, C., & Masur, J. (2009a). Happiness and punishment. *University of Chicago Law Review*, *76*(3), 1037–1082. https://www.jstor.org/stable/27793400

Bronsteen, J., Buccafusco, C., & Masur, J. S. (2009b). Welfare as happiness. *Georgetown Law Journal*, *98*, 1583–1641.

Bronsteen, J., Buccafusco, C., & Masur, J. S. (2014). Happiness and the law. University of Chicago Press.

Huang, P. H. (2008). Authentic happiness, self-knowledge and legal policy. *Minnesota Journal of Law Science & Technology*, *9*, 755–784.

Huang, P. H. (2010). Happiness studies and legal policy. *Annual Review of Law and Social Science*, *6*, 405–432. https://doi.org/10.1146/annurev-lawsocsci-102209-152828

Huang, P. H. (2018). Subjective well-being and the law. In E. Diener, S. Oishi, & L. Tay (Eds.), *Handbook of well-being*. DEF Publishers.

Huang, P. H., & Blumenthal, J. A. (2009). Positive institutions, law, and policy. In S. J. Lopez & C. R. Snyder (Eds.), *Oxford library of psychology*. Oxford handbook of positive psychology (pp. 589–597). Oxford University Press. https://psycnet.apa.org/record/2009-05143-056

Sunstein, C. R. (2018). *The cost-benefit revolution*. MIT Press.

Sunstein, C., & Posner, E. (2010). *Law and happiness*. University of Chicago Press.

C. PSYCHOLOGY

Although not particularly emphasized in the existing prioritization research literature, the discipline of psychology and the related fields of the cognitive sciences play a significant role in properly addressing many of the questions outlined in this research agenda. Given that one of the primary purposes of law (and in particular, the legal interventions highlighted in this agenda) is to influence and control human behavior, doing so requires a sophisticated understanding of the human mind. Psychology also interacts with many of the other highlighted disciplines of this appendix, such as economics (for example, behavioral law and economics), philosophy (for example, experimental jurisprudence), and political science (for example, political psychology) an understanding of which is likewise useful for this agenda, particularly with regard to institutional design and decision-making.

RELEVANT LITERATURE ON LAW AND PSYCHOLOGY GENERALLY

Bartol, C. R., & Bartol, A. M. (2018). *Psychology and law*. SAGE Publications.

Bersoff, D. N. (1986). Psychologists and the judicial system: Broader perspectives. *Law and Human Behavior*, *10*(1–2), 151–165. https://doi.org/10.1007/BF01044566

Busey, T. A., & Loftus, G. R. (2007). Cognitive science and the law. *Trends in Cognitive Sciences*, *11*(3), 111–117. https://doi.org/10.1016/j.tics.2006.12.004

Goodenough, O. R., & Tucker, M. (2010). Law and cognitive neuroscience. *Annual Review of Law and Social Science*, *6*, 61–92. https://doi.org/10.1146/annurev.lawsocsci.093008.131523

Greene J., Shariff A. F., Clark C. J., Baumeister, R. F., Luguri J., Vohs K. D., & Karremans J. C. (2014). Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological Science*, *25*(8), 1563–1570. https://doi.org/10.1177%2F0956797614534693

Kapardis, A. (2009). *Psychology and law: A critical introduction*. Cambridge University Press.

Kahan, D. M. (2015). Laws of cognition and the cognition of law. *Cognition*, *135*, 56–60. https://doi.org/10.1016/j.cognition.2014.11.025

Sznycer, D., & Patrick, C. (2020). The origins of criminal law. *Nature Human Behaviour*, *4*(5), 506–516. https://doi.org/10.1038/s41562-020-0827-8

Tapp, J. L. (1976). Psychology and the law: An overture. *Annual Review of Psychology*, *27*(1), 359–404. https://doi.org/10.1146/annurev.ps.27.020176.002043

Tushnet, R. (2007). Gone in sixty milliseconds: Trademark law and cognitive science. *Texas Law Review*, *86*, 507–568.

RELEVANT LITERATURE ON COGNITIVE BIASES AND DESCRIPTIVE DECISION THEORY

Arceneaux, K. (2012). Cognitive biases and the strength of political arguments. *American Journal of Political Science*, *56*(2), 271–285. https://doi.org/10.1111/j.1540-5907.2011.00573.x

Caviola, L., Everett, J. A., & Faber, N. S. (2019). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology*, *116*(6), 1011. https://psycnet.apa.org/doi/10.1037/pspp0000182

Cushman, F. (2013) Action, outcome, and value: A dual system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–329. https://doi.org/10.1177/1088868313495594

Fox, C. R., Erner, C., & Walters, D. J. (2015). *Decision under risk: From the field to the laboratory and back*. The Wiley Blackwell handbook of judgment and decision making, 43–88. https://doi.org/10.1002/9781118468333.ch2

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In L. C. MacLean & W. T. Ziemba (Eds.), *Handbook of the fundamentals of financial decision making: Part I* (pp. 99–127). https://doi.org/10.1142/9789814417358_0006

Levy, J. S. (1997). Prospect theory, rational choice, and international relations. *International Studies Quarterly*, *41*(1), 87–112. https://doi.org/10.1111/0020-8833.00034

Nordgren, L. F., & McDonnell, M. H. M. (2011). The scope-severity paradox: Why doing more harm is judged to be less harmful. *Social Psychological and Personality Science*, *2*(1), 97–102. https://doi.org/10.1177/1948550610382308

Oliver, A. (Ed.) (2013). *Behavioural public policy*. Cambridge University Press.

Schubert, S., Caviola, L., & Faber, N. S. (2019). The psychology of existential risk: Moral judgments about human extinction. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-50145-9

Slovic, P. (1987). Perception of risk. *Science*, *246*(4799), 280–285. https://doi.org/10.1126/science.3563507

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge University Press.

Relevant Literature on Moral Psychology, Neuroscience, and the Cognitive Sciences more Generally

Cushman, F., & Young, L. (2009). The psychology of dilemmas and the philosophy of morality. *Ethical Theory and Moral Practice*, *12*, 9–24. https://doi.org/10.1007/s10677-008-9145-3

Frankish, K., & Ramsey, W. (Eds.) (2012). The Cambridge handbook of cognitive science. Cambridge University Press.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, *2*(3), 528–554. https://doi.org/10.1111/j.1756-8765.2010.01094.x

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47) (pp. 55–130). Academic Press. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66–77. https://doi.org/10.1016/j.cognition.2017.03.004

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Huang, K., Bernhard, R., Barak-Corren, N., Bazerman, M., & Greene, J. D. (2020). *Veil-of-ignorance reasoning favors allocating resources to younger patients during the COVID-19 crisis*. PsyArXiv. https://doi.org/10.31234/osf.io/npm4v

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-ignorance reasoning favors the greater good. *Proceedings of the National Academy of Sciences*, *116*(48), 23989–23995. https://doi.org/10.1073/pnas.1910125116

Lapsley, D. K. (2018). *Moral psychology*. Routledge.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

Sapolsky, R. M. (2017). *Behave: The biology of humans at our best and worst*. Penguin.

Zeki, S., Goodenough, O. R., Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1451), 1775–1785. https://doi.org/10.1098/rstb.2004.1546

D. MACROHISTORY

Insofar as large-scale societal trends repeat themselves in explainable and comprehensible ways, macrohistorical analysis may be one of the best tools in predicting the long-term future. In the context of law, macrohistory may be particularly useful in evaluating potential lock-in effects from constitutional and other legal instruments, as well as in assessing law's ability more generally to influence the long-term future in comparison to other non-legal mechanisms.

RELEVANT LITERATURE ON MACROHISTORIC ANALYSES OF LAW

Eisner, M. (2003). Long-term historical trends in violent crime. *Crime and Justice*, *30*, 83–142.

Elkins, Z., Ginsburg, T., & Melton, J. (2009, October 15). *The lifespan of written constitutions*. University of Chicago Law School. https://www.law.uchicago.edu/news/lifespan-written-constitutions

Plucknett, T. F. T. (2001). *A concise history of the common law*. The Lawbook Exchange, Ltd.

Stein, P. (1999). *Roman law in European history*. Cambridge University Press.

Tamanaha, B. Z. (2004). *On the rule of law: History, politics, theory*. Cambridge University Press.

RELEVANT LITERATURE ON MACROHISTORY AND THE FUTURE

Barrow, J. D., & Tipler, F. J. 1986. *The anthropic cosmological principle*. Oxford University Press.

Bostrom, N. 2009. The future of humanity. In J.-K. Berg Olsen, E. Selinger & S. Riis (Eds.), *New waves in philosophy of technology*. Palgrave McMillan.

Chaisson, E. (2007). Energy, ethics and the far future. In *Energy challenges: The Next 1000 Years, Foundation for the Future Proceedings* (pp. 131–138).

Clarke, A. C. (1973). *Profiles of the future: An inquiry into the limits of possibility*. Harper & Row.

Diamond, J. (2004). *Collapse: How societies choose to fail or succeed*. Viking.

Drexler, K. E. (2013). *Radical abundance: How a revolution in nanotechnology will change civilization*. Public Affairs.

Grinin, L. (2012). *Macrohistory and globalization*. Uchitel.

Harris, J. (2019). *How tractable is changing the course of history?* Sentience Institute. https://www.sentienceinstitute.org/blog/how-tractable-is-changing-the-course-of-history

Inayatullah, S. (1998). Macrohistory and futures studies. *Futures*, *30*(5), 381–394. https://doi.org/10.1016/S0016-3287(98)00043-3

Inayatullah, S. (2017). Macrohistory and timing the future as practice. *World Futures Review*, *9*(1), 26–33. https://doi.org/10.1177%2F1946756716686788

Korotayev, A. V. & LePoire D. J. (2020). *The 21st century singularity and global futures: A big history perspective*. Springer. https://doi.org/10.1177%2F106939717300800103

Mazlish, B. (1993). *The fourth discontinuity: The co-evolution of humans and machines*. Yale University Press.

More, M., & Vita-More N. (Eds.) (2013). *The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future*. Wiley-Blackwell.

Rees, M. (2003). *Our final hour: A scientist's warning: How terror, error, and environmental disaster threaten humankind's future in this century—on Earth and beyond*. Basic Books.

Rifkin, J. (2011). *The third industrial revolution: How lateral power is transforming energy, the economy, and the world*. Palgrave Macmillan.

Sagan, C. (1994). *Pale blue dot: A vision of the human future in space*. Ballantine Books.

Smolin, L. (1999). *The origins of life: From the birth of life to the origins of language*. Oxford University Press.

Steffen, W. et al. The trajectory of the Anthropocene: The Great Acceleration. *The Anthropocene Review*, *2*(1), 81–98. https://doi.org/10.1177/2053019614564785

Steffen, W., Sanderson, R. A., Tyson, P. D., Jäger, J., Matson, P. A., Moore III, B., Oldfield, F., Richardson, K., Schellnhuber, H.-J., Turner, B. L., Wasson, R. J. (2004). *Global change and the earth system: A planet under pressure*. The IGPB Series. http://www.igbp.net/publications/igbpbookseries/igbpbookseries/globalchangeandtheearthsystem2004.5.1b8ae20512db692f2a680007462.html.

Stock, G. (1993). *Metaman: The merging of humans and machines into a global superorganism*. Simon & Schuster.

Wright, R. (2000). *Nonzero: The logic of human destiny*. Random House.

Relevant Literature on Macrohistory Generally

Bryson, B. (2004). *A short history of nearly everything*. Transworld Publishers.

Chaisson, E. (2005). *Epic of evolution: Seven ages of the cosmos*. Columbia University Press.

Christian, D. (2005). *Maps of time: An introduction to big history*. University of California Press.

Christian, D. (2018). *Origin story: A big history of everything*. Little, Brown and Company.

Corballis, M. C. (2011). *The recursive mind: The origins of human language, thought, and civilization*. Princeton University Press.

Diamond, J. (1998). *Guns, germs and steel: The fates of human societies*. Vintage.

Galtung, J., & Inayatullah, S. (1997). *Macrohistory and macrohistorians: Perspectives on individual, social, and civilizational change*. Greenwood Publishing Group.

Harari, Y. N. (2014). *Sapiens: A brief history of humankind*. Vintage.

Henrich, J. (2017). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Jantsch, E. (1980). *The self-organizing universe: Scientific and human implications of the emerging paradigm of evolution*. Pergamon Press.

Kant, I. (1784). Idea for a universal history from a cosmopolitan point of view. In H. Reiss (Ed.), *Kant's political writing* (pp. 41–53).

Kardashev, N. S. (1997). Cosmology and civilizations. *Astrophysics and Space Science*, *252*(1–2), 25–40.

Lineweaver, C. H., Davies, P. C. W., & Ruse, M. (Eds.). (2013). *Complexity and the arrow of time*. Cambridge University Press.

Lovelock, J. (2000). *Gaia: A new look at life on earth*. Oxford University Press.

Maynard Smith, J., & Szathmáry, E. (1999). *The origins of life: From the birth of life to the origins of language*. Oxford University Press.

Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. Penguin.

Samson, P. R., & Pitt, D. (Eds.) (1999). *The biosphere and noosphere reader: Global environment, society and change*. Routledge.

Smil, V. (1994). *Energy in world history*. Westview Press.

Turchin, P. (2003). *Historical dynamics: Why states rise and fall*. Princeton University Press.

E. POLITICAL SCIENCE

Since at least the 1920s, scholars and judges have recognized that law and politics are deeply intertwined. Better understanding the interaction between these two disciplines, including both the effect of law on political institutions and the influence of political institutions on shaping the law, seems particularly critical in determining how legal systems can and ought to be set up so as to most positively impact the long-term trajectory.

RELEVANT LITERATURE ON POLITICAL SCIENCE AND INSTITUTIONAL DESIGN

Blyth, M. (2002). *Great transformations: Economic ideas and institutional change in the twentieth century*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087230

Conant, L. J. (2002). *Justice contained: Law and politics in the European Union*. Cornell University Press.

Engel, C. (2007). *Institutions for intuitive man. MPI Collective Goods Preprint* (No. 2007/12).

Finnemore, M., & Toope, S. J. (2001). Alternatives to "legalization": Richer views of law and politics. *International Organization*, *55*(3), 743–758. https://doi.org/10.1162/00208180152507614

Glaeser, E. L., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2004). Do institutions cause growth? *Journal of Economic Growth*, *9*(3), 271–303. https://doi.org/10.1023/B:JOEG.0000038933.16398.ed

González-Ricoy, I. & Gosseries, A. (Eds.) (2016). *Institutions for future generations*. Oxford University Press. doi.org/10.1093/acprof:oso/9780198746959.001.0001

Guerrero, A. A. (2014). Against elections: The lottocratic alternative. *Philosophy & Public Affairs*, *42*(2), 135–178. https://doi.org/10.1111/papa.12029

Hafner-Burton, E. M., Victor, D. G., & Lupu, Y. (2012). Political science research on international law: The state of the field. *American Journal of International Law*, *106*(1), 47–97. https://doi.org/10.5305/amerjintelaw.106.1.0047

Jacobs, A. M. (2011). *Governing for the long term: Democracy and the politics of investment*. Cambridge University Press. https://doi.org/10.1017/CBO9780511921766

Jones, N., O'Brien, M., & Ryan, T. (2018). Representation of future generations in United Kingdom policy-making. *Futures*, *102*, 153–163. https://doi.org/10.1016/j.futures.2018.01.007

Levitsky, S., & Murillo, M. (2009). Variation in institutional strength. *Annual Review of Political Science*, *12*, 115–133.

Loughlin, M. (1992). *Public law and political theory*. Clarendon Press.

Posner, E. (2014). *The twilight of human rights law*. Oxford University Press.

Raz, J. (1994). *Ethics in the public domain: essays in the morality of law and politics* (pp. 261–309). Clarendon Press.

Tankard, M. E., & Paluck E. L. (2916). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022

Tännsjö, T. (2007). Future people, the all affected principle, and the limits of the aggregation model of democracy. In T. Rønnow-Rasmussen, B. Petersson, J. Josefsson, & D. Egonsson (Eds.), *Homage à Wlodek: 60 philosophical papers dedicated to Wlodek Rabinowicz*. Lund. https://www.researchgate.net/publication/251687495_Future_People_the_All_Affected_Principle_and_the_Limits_of_the_Aggregation_Model_of_Democracy_1

Tonn, B. (1966). A design for future-oriented government. *Futures*, *28*(5), 413–431. https://doi.org/10.1016/0016-3287(96)00017-1

Voigt, S. (2009). How (not) to measure institutions. *Journal of Institutional Economics*, *9*(1), 1–26. https://doi.org/10.1017/S1744137412000148

Whittington, K. E., Kelemen, R. D., & Caldeira, G. A. (Eds.) (2010). *The Oxford handbook of law and politics* (Vol. 3). Oxford University Press on Demand.

RELEVANT LITERATURE ON POLITICAL SCIENCE AND INSTITUTIONAL DECISION-MAKING

Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. New York: Crown Publishing Group.

Arnold, R. D. (1992). *The logic of congressional action*. Yale University Press.

Cross, F. B. (1997). Political science and the new legal realism: A case of unfortunate interdisciplinary ignorance. *Northwestern University Law Review*, *92*(1), 251–326.

Deutsch, J. G. (1967). Neutrality, legitimacy, and the Supreme Court: Some intersections between law and political science. *Stanford Law Review*, *20*, 169–261. https://digitalcommons.law.yale.edu/fss_papers/1886

Fallon Jr., R. H. (2010). The Supreme Court, habeas corpus, and the war on terror: An essay on law and political science. *Columbia Law Review*, *110*(2), 352–398. https://www.jstor.org/stable/27806621

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351–401. https://doi.org/10.1257/002205102320161311

Iaryczower, M., Spiller, P. T., & Tommasi, M. (2002). Judicial independence in unstable environments, Argentina 1935–1998. *American Journal of Political Science*, *46*(4), 699–716. https://doi.org/10.2307/3088428

Jacobs, A. M., & Matthews, S. J. (2012). Why do citizens discount the future? Public opinion and the timing of policy consequences. *British Journal of Political Science*, *42*(4), 903–35. https://doi.org/10.1017/S0007123412000117

McGuire, K. T., & Stimson, J. A. (2004). The least dangerous branch revisited: New evidence on Supreme Court responsiveness to public preferences. *Journal of Politics*, *66*(4), 1018–1035. https://doi.org/10.1111/j.1468-2508.2004.00288.x

Nordhaus, W. D. (1975). The political business cycle. *Review of Economic Studies*, *42*(2), 169–190. https://doi.org/10.2307/2296528.

Steiner, J., Bächtiger, A., Spörndli, M., & Steenbergen, M. R. *Deliberative politics in action: Analyzing parliamentary discourse.* Cambridge University Press. https://doi.org/10.1017/CBO9780511491153