LEGAL PRIORITIES
RESEARCH:
A RESEARCH AGENDA

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

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

1. The Case for Legal Priorities Research  
   1.1 The Importance and Neglectedness of Legal Priorities Research  
   1.2 Objections to Legal Priorities Research  
2. Longtermism  
   2.1 Foundations of Longtermism  
   2.2 Objections to Longtermism  
3. Methodology  
   3.1 Methodology for Cause Prioritization  
   3.2 Methodology for Identifying Research Projects

## PART 2: EXPLORATION BY CAUSE AREAS

4. Artificial Intelligence  
   4.1 Reducing Existential Risks from AI  
   4.2 Reducing Suffering Risks from AI  
   4.3 Sharing the Benefits of AI  
   4.4 Meta-Research in AI  
5. Synthetic Biology and Biorisk  
   5.1 Preventing Intentional or Accidental Release of a Biological Agent  
   5.2 Coordination and Response  
   5.3 Sharing the Benefits of Synthetic Biology  
6. Institutional Design  
   6.1 Design of Legal Institutions  
   6.2 Judicial Decision-Making  
   6.3 Impact, Evaluation and Uncertainty in Law  
7. Meta-Research
PART 3: CAUSE AREAS FOR FURTHER ENGAGEMENT

8  Space Governance 105
     8.1 Coordination and Peace in Outer Space 106
     8.2 Sustainability of the Long-Term Presence of Humanity in Outer Space 109

9  Animal Law 113
     9.1 Incremental Changes 114
     9.2 Fundamental Changes 117

REFERENCES

APPENDIX: CLOSELY RELATED AREAS OF EXISTING ACADEMIC RESEARCH

A. Philosophy 168
B. Economics 170
C. Psychology 173
D. Macrohistory 176
E. Political Science 178
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. 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).
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). 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). 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). 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, 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.\footnote{All searches were performed in October 2020. The original and most-cited Chevron case (\textit{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 (\textit{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).}

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 \textit{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.\footnote{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.}

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).
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), 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).

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; Oster tag, 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 omniscient 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 \textit{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.
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.” 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). 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).
LONGTERMISM

neighborhood. 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 (Gal- lup & 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. Or, from a virtue ethics perspective, 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). 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). Many non-consequentialist

---

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

---

\(^{18}\) 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). Legal scholars are also in favor of prioritizing legal research questions based on longtermist considerations (Martinez & Winter, 2021). 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 \textit{ex ante} effects, as opposed to \textit{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. 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. 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. 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.”\textsuperscript{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).\textsuperscript{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).\textsuperscript{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).\textsuperscript{35} Likewise, laws focused on the short-term may face similar challenges. For example, environmental regulations\textsuperscript{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

\textsuperscript{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).

\textsuperscript{33} For recent informal discussion, see Greaves (2020).

\textsuperscript{34} Karnofsky (2013) refers to these as “flow-through effects”.

\textsuperscript{35} For more discussion on the relationship between economic growth and risks from emerging technology see Bostrom (2019) and Aschenbrenner (2020).

\textsuperscript{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). Additionally, Lempert (2019) has proposed decision support tools to address the “deep uncertainty” in policy decisions through Robust Decision Making (RDM). 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. 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. 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 satisficing, 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).\textsuperscript{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.\textsuperscript{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.\textsuperscript{47}

\textsuperscript{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).

\textsuperscript{46} See, for example, the Center on Long Term Risk, the Center for Reducing Suffering, and the Center for Emerging Risk Research.

\textsuperscript{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.” 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?”

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. 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). If one takes humanity’s current trajectory as a reference basis (*status quo*...
trajectory),\textsuperscript{52} it is conceivable that the conditions for welfare might drastically improve or deteriorate over time depending on humanity’s actions today.\textsuperscript{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.

\textsuperscript{52} For more information on the status quo trajectory, see Baum et al. (2019, pp. 6–9).

\textsuperscript{53} But see also Section 2.2 regarding empirical uncertainties of influencing the far future.
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. Hence, legal research should aim at avoiding entering persistent or even permanent trajectories. 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.

**Reducing Existential Risks**

“Existential risks (x-risks)” 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). 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).

Reducing Risks of Astronomical Suffering

“Suffering risks (s-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). 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.

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.

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.

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.

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), and conflict involving powerful forms of artificial intelligence (Clifton, 2020).

Reducing Risks of Losing Astronomical Pleasure

Let us define “pleasure risks (p-risks)” 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. 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. 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 “tranquillism” (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. 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.

\[\text{Figure 4. Possible trajectories and distributions for p-opportunities, and x- and s- risks.}\]

**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). 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:\(^71\)

---

\(^70\) 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. 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. 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. 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. 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.”

---

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. 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. 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. 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).

CO₂ 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 °C.
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 bio-risk (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)\textsuperscript{80} could significantly shape the long-term future. On the one hand, it could enable scientific breakthroughs\textsuperscript{81} and the accumulation of unprecedented wealth. On the other hand, it could pose existential risks and cause

\textsuperscript{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).

\textsuperscript{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.\(^{86}\)

---


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-Bar ratt 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). Specific scenarios in which AI accidents cause an existential catastrophe have been described by Bostrom (2014) and Yudkowsky (2008a), as well as Christiano (2019). 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. 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), Quinonero-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 Christiano 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.\textsuperscript{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).\textsuperscript{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?

\textsuperscript{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).

\textsuperscript{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). 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? 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). A possible scenario involves some kind of war exacerbated by developments in AI (Aguirre, 2020; Allen & Chan, 2017; Avin & Amadue, 2019; Boulain et al., 2020; Dafoe, 2018; Geist & Lohn, 2018; Horowitz, 2019; Horowitz et al., 2019; Jayanti & Avin, 2020; Lieber & Press, 2017; Maas, 2019a, 2019b). 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**


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


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


**EXISTING INFORMAL DISCUSSION**


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). 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). 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). It seems at least somewhat plausible that artificial agents might exhibit behavior that resembles malevolent traits like psychopathy or sadism. 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


EXISTING INFORMAL DISCUSSION


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


EXISTING INFORMAL DISCUSSION


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. 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**


**Existing Informal Discussion**


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-n0G3QPq/ben-garfinkel-how-sure-are-we-about-this-ai-stuff


5 SYNTHETIC BIOLOGY AND BIORISK

Synthetic biology 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. Some extreme cases of this are

102 There is no generally accepted definition of “synthetic biology.” The term emerged at the turn of the millennia 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. 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). 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). 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,
which may implicate existential risks through loss of potential, as well as pleasure risks and suffering risks (Section 5.3).

5.1 Preventing Intentional or Accidental Release of a Biological Agent

The most desirable outcome is to avoid a catastrophic event entirely (Cotton-Barrett et al., 2020, p. 273). If we can prevent the intentional or accidental release 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). 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. 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)? 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, 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. 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

---

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. 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. 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. 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? 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.

---

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\textsuperscript{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.\textsuperscript{122} Technical forensics may aid in attribution,\textsuperscript{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

\textsuperscript{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.

\textsuperscript{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).

\textsuperscript{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).
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. 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.

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 & Commings, 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).
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, 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


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).


EXPLORATION BY CAUSE AREAS


**EXISTING INFORMAL DISCUSSION**


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-Barrett 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, 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, 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.\textsuperscript{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**


\textsuperscript{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).


**Existing Informal Discussion**


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 conversion 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 measures138 (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.


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. 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.
EXPLORATION BY CAUSE AREAS

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; Forsdam 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

---

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. 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


---

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-


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


**EXISTING INFORMAL DISCUSSION**


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. 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, 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?

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.
conventional 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? Given that rigid constitutional provisions can create strong lock-in effects, 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, 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? 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-Rico y and Axel Gossy (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. Gossy & 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 Crofoot (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? 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)? 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%.

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.

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.\textsuperscript{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?

\textit{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?

\textit{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

\textsuperscript{153} See also John (2020b).
new national and/or international criminal laws against the endangerment of humanity?\textsuperscript{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.\textsuperscript{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,\textsuperscript{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

\textsuperscript{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).

\textsuperscript{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 institutions protections for the interests of all sentient beings.

\textsuperscript{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 $10^{19}$), 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?\footnote{See, e.g., Art. 13 TFEU; Chessman, 2018; Reddy, 2020; World Animal Protection, 2014; see also Section 9.}

### 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


EXISTING INFORMAL DISCUSSION


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 Union160 and “balancing tests” in the United States Supreme Court.161 What sorts of cognitive


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


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. 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). 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 decisions?

---


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 (Cavola 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. 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 (Crotof, 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**


---

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


EXISTING INFORMAL DISCUSSION


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,
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


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. 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?

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).

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:


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


**EXISTING INFORMAL DISCUSSION**


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. 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.

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, 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


**EXISTING INFORMAL DISCUSSION**


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 are 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**


Caviola, L. (2019). *How we value animals: The psychology of speciesism* [Doctoral dissertation]. [https://ora.ox.ac.uk/objects/uuid:9000a8be-b6de-4e63-88e8-974b9daaa83a](https://ora.ox.ac.uk/objects/uuid:9000a8be-b6de-4e63-88e8-974b9daaa83a)


References

80,000 Hours (2020, May). Our current view of the world’s most pressing problems. https://80000hours.org/problem-profiles/


Aguirre, A. (2020, November 11). Why those who care about catastrophic and existential risk should care about autonomous weapons [Online forum post]. Effective Altruism
REFERENCES


REFERENCES


REFERENCES


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


REFERENCES


REFERENCES


Duda, R., & Koehler, A. (2016, April). *Climate change (extreme risks)*. 80,000 Hours. https://80000hours.org/problem-profiles/climate-change/


REFERENCES


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
REFERENCES

Biosecurity, Biosafety and Biodefense Law, 7(1), 25–45. https://doi.org/10.1515/jbbbl-2016-0007


REFERENCES


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.


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


REFERENCES


John, T. M. (2020b). Empowering future people by empowering the young? [Manuscript in preparation]. https://docs.google.com/document/d/1NeP3NPd-m4NhxhS9O5aqdsaphh4v__iFcEajMCCePHM_0/edit


REFERENCES


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


REFERENCES


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


Merriam-Webster. 'Pandemic' vs. 'epidemic': How they overlap and where they differ. https://www.merriam-webster.com/words-at-play/epidemic-vs-pandemic-difference


REFERENCES


REFERENCES


REFERENCES


REFERENCES


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/


REFERENCES


REFERENCES


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


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


**RELEVANT LITERATURE ON LONGTERMISM**


**RELEVANT LITERATURE ON NORMATIVE UNCERTAINTY AND DECISION THEORY**


**RELEVANT LITERATURE ON EXPERIMENTAL JURISPRUDENCE**

See also Appendix C: Psychology


**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**


**RELEVANT LITERATURE ON BEHAVIORAL LAW AND ECONOMICS**


### Relevant Literature on Comparative Law and Economics


### Relevant Literature on Law, Happiness, and Well-Being

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


**RELEVANT LITERATURE ON MORAL PSYCHOLOGY, NEUROSCIENCE, AND THE COGNITIVE SCIENCES MORE GENERALLY**


**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**


**RELEVANT LITERATURE ON MACROHISTORY AND THE FUTURE**


RELEVANT LITERATURE ON MACROHISTORY GENERALLY


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).

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


**RELEVANT LITERATURE ON POLITICAL SCIENCE AND INSTITUTIONAL DECISION-MAKING**


