Existential Advocacy: Lawyering for AI Safety and the Future of Humanity

John Bliss*

Abstract

Lawyers have played a central role in a wide range of social movements aiming to provide legal voice to marginalized communities. How might this tradition of social-change lawyering apply to the protection of future generations—a population that cannot advocate for themselves? This is a pressing question in the movement to mitigate "existential risk," which refers to events that would foreclose a meaningful existence for future generations either through human extinction or irreversible trajectories of human suffering. Over the past two decades, an Oxford-based academic community has been researching existential threats that could arise from emerging technology, such as advanced artificial intelligence and engineered pandemics. These concerns have sparked the founding of dozens of non-profit organizations working to preserve the future of humanity. In just the past couple of years, this movement has entered the realm of law and politics, where it has already helped establish new legal mechanisms addressing AI safety, existential risk, and the rights of future generations. At the same time, the movement has faced criticism for shifting attention away from current social injustices.

This article presents the first empirical study of the legal wing of this movement. It asks how "existential advocates" approach the key questions faced by all social-change lawyers regarding: (1) efficacy, which is framed here as the question of how to impact such a large-scale, uncertain, and abstract issue; and (2) accountability, which is framed here as the question of how to faithfully represent the future generations who are silent stakeholders in the decisions we make today. Drawing on a qualitative study embedded in this community of

* Assistant Professor of Law, University of Denver Sturm College of Law. I wish to thank my research assistants Juliana Todeschi and J.P. Scanlon. Invaluable feedback was provided by Catherine Albiston, Alan Chen, Scott Cummings, Meghan Dawe, Benjamin Eidelson, Noah Feldman, Bryon Fong, Calvin Morrill, Ann Southworth, and David Wilkins. I received helpful comments when presenting this work at the UC Berkeley Center for the Study of Law and Society, the Harvard Law School Center on the Legal Profession, the UC Irvine Center for Empirical Research on the Legal Profession, the University of Denver Summer Scholarship Series, the Multidisciplinary Forum on Longtermism and the Law at Universität Hamburg, and the Law and Society Association. I am grateful to Christoph Winter and other members of the Legal Priorities Project who granted me access to conduct an in-depth ethnography of their organization. I also wish to thank the 53 participants who voluntarily participated in research interviews. Funding for this study was provided by the Hughes-Ruud Research Professorship at the University of Denver. The study was approved by the University of Denver Institutional Review Board. All errors are mine.

legal advocates, the article describes the development of a distinct model of social-change lawyering—the "priorities methodology." This model aims to maximize impact using formal processes for selecting goals and strategies while minimizing cognitive biases. The priorities methodology is an innovation within the tradition of social-change lawyering, although it faces some difficult points of tension when seeking to mobilize a broader movement of lawyers and other actors. The article concludes with recommendations for adapting this model as the existential risk community scales up and pursues more direct and high-profile legal interventions.

TABLE OF CONTENTS

INTRODUCTION

> "In some ways [the movement to mitigate existential risk] is the most inclusive movement, because it's trying to preserve everything. It sort of sits above all social movements. . .as a continuation of social justice. I think it's really vitally important to every person in the world today and every single future person. . ."
>
> "[When presenting existential risk to lawmakers, they often respond] 'How can I think about this when we've got so many crocodiles closer to the boat?'"
>
> "If the core aggrieved group is future generations, they don't get to chime in because: no time travel."
>
> – Excerpts from research interviews

Could our technological progress be paving the way toward our own extinction? When thinking about extinction-level events, one might imagine climate change, asteroids, supervolcanoes, widespread nuclear conflict, or an engineered pandemic 500 times more deadly than COVID-19.[1] Or, like the advocates examined in this article, you might see advancing AI as the top source of existential risk. Even many of the leaders of the AI industry believe their technology could, in the worst case, lead to "lights for all of us,"[2] to quote Sam Altman of OpenAI.[3] This is a longstanding concern, dating back to Alan Turing and I.J. Good.[4] Advanced AI could conceivably threaten human survival by, e.g., widening the

---

1. Note that COVID-19, according to a recent World Health Organization report, has been responsible for killing one out of 500 people globally. Hence, this reference to a pandemic "500 times more deadly than COVID-19" is meant to suggest an upper limit of mortality amounting to human extinction. *See* Thomas Mulier & Clara Hernanz Lizarraga, *Covid Killed One out of Every 500 People, WHO Report Shows*, BLOOMBERG (May 5, 2022), http://www.bloomberg.news/articles/2022-05-05/covid-killed-about-1-out-of-every-500-people-who-report-shows [https://perma.cc/5Y45-JXWW]. For an analysis suggesting that such extreme pandemics are growing more likely, see Kevin Esvelt, *How a Deliberate Pandemic Could Crush Societies and What to Do About It*, BULLETIN OF THE ATOMIC SCIENTISTS (November 15, 2022), https://thebulletin.org/2022/11/how-a-deliberate-pandemic-could-crush-societies-and-what-to-do-about-it/ [https:perma.cc/6RNB-QAS8] (noting that advances in synthetic biology and "CRISPR-based gene drive systems" are increasingly able to produce new "pandemic-class agents" far more deadly than those found in nature, and describing scenarios where multiple synthetic pathogens could be released simultaneously). *See also* Abraar Karan & Stephen Luby, *A Natural Pandemic Has Been Terrible. A Synthetic One Would Be Even Worse*, STAT (Aug. 19, 2021), https://www.statnews.com/2021/08/19/natural-pandemic-terrible-synthetic-one-even-worse/ [https://perma.cc/6YTU-9SRZ] (describing pathogens that have been engineered to have the transmissibility of our most contagious diseases, such as measles, and the virulence of our most deadly diseases, such as Ebola); TOBY ORD, THE PRECIPICE: EXISTENTIAL RISK AND THE FUTURE OF HUMANITY 127–38 (2020) (detailing the history of gain-of-function research, including where researchers make deadly viruses more contagious, and the historical record of laboratory leaks, information hazards, and biological warfare, which could enable engineered pathogens to reach large populations). *See* discussion *infra* Part I.A.

2. Sarah Jackson, *The CEO of the company behind AI chatbot ChatGPT says the worst-case scenario for artificial intelligence is 'lights out for all of us'*, BUS. INSIDER (Jul. 4, 2023), https://www.businessinsider.com/chatgpt-openai-ceo-worst-case-ai-lights-out-for-all-2023-1 [perma.cc/AZ9D-XAPP].

3. *Id.*

4. *See* I. J. Good, *Speculations Concerning the First Ultraintelligent Machine*, 6 ADVANCES IN COMPUTS. 31 (1966) (discussing the possibility of artificial intelligence that renders human intelligence obsolete); Alan Turing, *Computing Machinery and Intelligence*, 59(236) MIND 433 (1950).

availability of information about how to create biological and chemical weapons, enabling near limitless production of autonomous drones and robots, accelerating the development of new weapons of mass destruction, influencing military decision-makers, undermining economic and political stability through surveillance, propaganda, and manipulation, unleashing cyber-attacks that devastate critical infrastructure, or even surpassing human intelligence and evading human control.[5] If the probability of such events occurring in the near future is more than negligible, how should we respond?

A rising community of advocates is working to mitigate these and other "existential risks," which they define as catastrophic events that would cause human extinction or other permanent destruction of a meaningful existence for future generations.[6] This article offers the first empirical study of the lawyers and other legal advocates in this community. Drawing on a multi-method research design, I examine how the tradition of social-change lawyering is being reimagined in the unique context of existential risk. What does it mean to be a lawyer for an issue, which, as described in the interviews quoted in the epigraph above, operates on an unimaginably large scale (affecting "every person in the world and every single future person") but is obscured by a host of cognitive biases and political incentives directing attention to issues that are seemingly more immediate (the "crocodiles closer to the boat")? Moreover, how can these lawyers faithfully represent future generations, a vast population of silent stakeholders in the decisions we make today (who cannot "chime in")? As detailed throughout this article, these advocates are engaged in a legal movement unlike any we have seen before and with a distinct understanding of what it means to be a lawyer for social change.

Over the past two decades, existential risk has been the subject of a growing interdisciplinary academic field of inquiry. In the canonical work of this field, *The Precipice: Existential Risk and the Future of Humanity* (2020), philosopher Toby Ord estimated that the next century brings a one-in-six chance of existential catastrophe.[7] Ord's analysis takes into account scientific and theoretical investigations

---

5. *See* Dan Hendrycks, Mantas Mazeika & Thomas Woodside, *An Overview of Catastrophic AI Risks*, ARXIV (Oct. 9, 2023), https://arxiv.org/pdf/2306.12001.pdf [https://perma.cc/D9A8-F5W6].

6. *See generally* Nick Bostrom, *Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards*, 9 J. EVOLUTION & TECH. 1 (2002) (introducing the notion of existential risk); *see also* ORD, *supra* note 1, at 4 (defining existential threats as "risks that threaten the destruction of humanity's long-term potential," which is "most obvious" with human extinction but also includes other irreversible future trajectories of immense suffering); Nick Bostrom, *Existential Risk Prevention as Global Priority*, 4 GLOB. POL'Y 15 (2013); NICK BOSTROM & MILAN M. ĆIRKOVIĆ, GLOBAL CATASTROPHIC RISKS (2008); ANNETTE BAIER, THE RIGHTS OF PAST AND FUTURE PERSONS (1973); JOHN LESLIE, THE END OF THE WORLD: THE SCIENCE AND ETHICS OF HUMAN EXTINCTION (1996).

7. Note that this estimate finds support from some other researchers in the field, although always with the caveat that such estimates are highly speculative and could be off by multiple orders of magnitude in either direction. ORD, *supra* note 1, at 26 (estimating a one in six chance ("Russian roulette") of existential catastrophe over the next century, and a one in three chance over the long-term future); *see* Anders Sandberg & Nick Bostrom, *Global Catastrophic Risks Survey*, FUTURE OF HUMAN. INST. (2008), https://www.global-catastrophic-risks.com/docs/2008-1.pdf [https://perma.cc/F875-6C23] (reporting on a survey of participants at the Global Catastrophic Risks Conference at Oxford University in 2008, finding an average estimate of 19% chance of human extinction prior to

of a variety of threats, which he has explored in his position as a senior fellow at Oxford University's Future of Humanity Institute.[8]

For young people today, the notion that there is a significant probability of existential risk might not seem especially surprising, given their general sense that "humanity is doomed."[9] But this popular sense of doom is usually associated with climate change and other events that would devastate large populations over a number of generations, and thus are worthy of great concern, but are projected to be very unlikely to foreclose the long-term future of humanity.[10] The greatest threats on the existential scale are thought to arise not from the familiar issues of climate change, asteroids, or nuclear weapons, but rather from new and emerging technologies, and the sense that we have only just started what will become a growing list of means to bring about existential catastrophes.[11]

Ord, and other scholars of existential risk, view artificial intelligence as the current leading threat category.[12] Amid recent transformative advances in generative AI systems, which have passed the Uniform Bar Exam and a wide range of other tests of aptitude and even common sense, the notion that AI may soon pose an existential risk has become an increasingly mainstream concern.[13] The three

---

2100, which considers a number of risk categories from great power wars, pandemics and super-intelligent AI). *Contra* WILLIAM MACASKILL, WHAT WE OWE THE FUTURE (2022) (suggesting a lower estimate of existential risk, likely below one percent over the next century); *see also* David Thorstad, *Existential Risk Pessimism and Time of Perils,* GLOB. PRIORITIES INST. (2022), https://globalprioritiesinstitute.org/wp-content/uploads/David-Thorstad-Existential-risk-pessimism-.pdf [https://perma.cc/TQD4-MNJX]; Michael Arid, *Database of Existential Risk Estimates*, EFFECTIVE ALTRUISM F. (Apr. 15, 2020), https://forum.effectivealtruism.org/posts/JQQAQrunyGGhzE23a/database-of-existential-risk-estimates [https://perma.cc/D7LY-YYYB].

8. *See generally* Ord, *supra* note 1.

9. Caroline Hickman, Elizabeth Marks, Panu Pihkala, Susan Clayton, R. Eric Lewandowski, Elouise E. Mayall, Britt Wray, Catriona Mellor & Lisa van Susteren, *Young People's Voices on Climate Anxiety, Government Betrayal and Moral Injury: A Global Phenomenon*, 5.12 THE LANCET PLANETARY HEALTH 863, 863 (2021) (reporting a 10-country survey finding that a majority of people under 25 years old believe that "humanity is doomed."); *see* Angela Lashbrook, *'No Point in Anything Else': Gen Z Members Flock to Climate Careers*, THE GUARDIAN (Sept. 6, 2020), https://www.theguardian.com/environment/2021/sep/06/gen-z-climate-change-careers-jobs [https://perma.cc/NJ2E-ZKGH] (describing the overwhelming concern for climate change among Gen Z as reflected in their career aspirations and choices); *see generally* Madhukar Pai, *Young Climate Justice Activists Are Fighting for Our Collective Survival*, FORBES (July 28, 2022), https://www.forbes.com/sites/madhukarpai/2022/07/28/young-climate-justice-activists-are-fighting-for-our-collective-survival [https://perma.cc/3P3C-2BZH].

10. *See* Pai, *supra* note 9.

11. *See, e.g.*, ORD, *supra* note 1, at 148 (estimating the existential risk associated with both nuclear war and climate change at 1 in 1,000 over the next century, but offering far higher estimates for unaligned artificial intelligence (1 in 10), engineered pandemics (1 in 30), and unforeseen anthropogenic risks (1 in 30)); *see also* Holden Karnofsky, *Forecasting Transformative AI, Part 1: What Kind of AI?*, COLD TAKES (Aug. 10, 2021), https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai [https://perma.cc/XY8L-H8DN] (suggesting that transformative artificial intelligence may radically accelerate scientific and technological development, which could lead to "technology capable of wiping humans out of existence").

12. *See* ORD, *supra* note 1; Nick Bostrom, *Ethical Issues in Advanced Artificial Intelligence*, *in* COGNITIVE, EMOTIVE & ETHICAL ASPECTS OF DECISION MAKING IN HUMANS & IN ARTIFICIAL INTELLIGENCE 12 (George Eric Lasker, Wendell Wallach, Iva Smit eds., 2d ed. 2003).

13. *See GPT-4*, OPENAI, https://openai.com/research/gpt-4 [https://perma.cc/5XZU-94ED] (last visited Sept. 23, 2023) (describing GPT-4's performance on a wide range of exams, including scoring in the 99th percentile on

most-cited AI researchers have signed a public statement, along with many other experts in the field, stating that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[14] A recent survey suggested that around half of AI experts believe that there is a 5% or greater chance that AI will cause human extinction within this century.[15] A poll of American adults found that nearly half (46%) were either "somewhat concerned" or "very concerned" about "the possibility that AI will cause the end of the human race on Earth."[16]

The problem of existential risk is made more difficult by the global coordination that may be required to contain threats of this magnitude: an existential catastrophe arising in one jurisdiction would destroy humanity in all jurisdictions.[17] Moreover, by definition, we have no experience with events that terminate humanity. We cannot afford a single failure to prevent such a catastrophe, nor can we rely on a reactive trial-and-error approach to designing our responses.[18] These observations have led some leading minds in this field to suggest that we are entering a new era of history, walking along a "precipice" in humanity's "most important century" as we make our initial encounter with existential

---

the GRE verbal section and 88th percentile on the LSAT); Zach Stein-Perlman & Katja Grace, *2022 Expert Survey on Progress in AI*, AI IMPACTS (Aug. 3, 2022), https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/ [https://perma.cc/7Q7S-4GAT] (finding that nearly half of AI researchers who participated in the survey believe there is a 5–10% chance AI will cause human extinction); *How Concerned, If At All, Are You About the Possibility That AI Will Cause the End of the Human Race on Earth?,* YOUGOV (Apr. 3, 2023), https://today.yougov.com/topics/technology/survey-results/daily/2023/04/03/ad825/3 [https://perma.cc/Q3GK-HX8R] (finding that 46% of survey respondents are somewhat or very concerned that AI "will cause the end of the human race on Earth"); *The End of Humanity: How Real is the Risk?*, 201 TIME S. PAC. 21 (2023); Cade Metz, *How Could A.I. Destroy Humanity?*, N.Y. TIMES (June 10, 2023), https://www.nytimes.com/2023/06/10/technology/ai-humanity.html [https://perma.cc/HJU6-8BZE] (observing that many AI technical experts only recently came to view AI as an existential threat, believing that the existential risk scenarios were not "all that plausible until the last year or so").

14. Statement on AI Risk, CENTER FOR AI SAFETY, https://www.safe.ai/work/statement-on-ai-risk#open-letter [perma.cc/A35V-445Z] (last visited Apr. 18, 2024); Google Scholar, https://scholar.google.com/citations?hl=en&view_op=search_authors&mauthors=label%3Aartificial_intelligence+OR+label%3Aai&btnG= [perma.cc/3GWB-MY99] (last visited May 21, 2024) (indicating the three most-cited researchers for "artificial intelligence" and "ai"); *see also* Billy Perrigo, *AI Is as Risky as Pandemics and Nuclear War, Top CEOs Say, Urging Global Cooperation*, TIME (May 30, 2023), https://time.com/6283386/ai-risk-openai-deepmind-letter/ [perma.cc/GE5S-BXP4].

15. Will Henshall, *When Might AI Outsmart Us? It Depends on Who You Ask*, TIME (Jan. 19, 2024), https://time.com/6556168/when-ai-outsmart-humans/[perma.cc/L4ZC-7827] ("The median respondent thought it was 5% likely that AGI leads to "extremely bad," outcomes, such as human extinction.")

16. Importantly, the population surveyed was only U.S. adult citizens and not all U.S. adults. Taylor Orth & Carl Bialik, *AI doomsday worries many Americans. So does apocalypse from climate change, nukes, war, and more*, YOUGOV (Apr. 14, 2023) https://today.yougov.com/technology/articles/45565-ai-nuclear-weapons-world-war-humanity-poll?redirect_from=%2Ftopics%2Ftechnology%2Farticles-reports%2F2023%2F04%2F14%2Fai-nuclear-weapons-world-war-humanity-poll [https://perma.cc/PM7G-B4NX].

17. *See* ORD, *supra* note 1, at 176 (observing that the absence of strong multilateral mechanisms of global governance presents a formidable challenge for regulating these risks in the nearly 200 countries of the world).

18. *See* Bostrom, *Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards*, *supra* note 6, at 3 ("Our approach to existential risks cannot be one of trial-and-error. There is no opportunity to learn from errors. The reactive approach—see what happens, limit damages, and learn from experience— is unworkable. Rather, we must take a proactive approach. This requires *foresight* to anticipate new types of threats and a willingness to take decisive *preventive action* and to bear the costs (moral and economic) of such actions.").

threats.[19] As the issue is often framed in this field, humanity could either be approaching its end, or, if we are able to survive the advent of technologies capable of producing existential catastrophes, and establish global regulatory systems to assure that any new existential threats that arise would be contained as well, we could still be in the early infancy of humanity with an extraordinarily long arc of (hopefully good) experience ahead of us.[20] This article contributes a new layer to this aspirational vision by considering what forms of legal advocacy might be most effective when seeking to persuade legal and political decision-makers to intervene in existential threats.

The Oxford-based academic study of existential risk has sparked a larger movement that not only conducts research, but is increasingly engaging in advocacy for "existential security." Non-profit organizations have proliferated in this field,[21] as well as a wave of student-led "existential risks initiatives" at leading universities, including Stanford, Cambridge, Harvard, and MIT.[22] The field has seen extraordinary fundraising in recent years, currently totaling at least several billion U.S. dollars, surpassing many of the leading philanthropic foundations of the world.[23]

This movement has only recently started exploring the role of law and lawyers.[24] There are some initial signs that the legal profession might be receptive to addressing existential risk. The two most cited scholars in U.S. legal academia have written books on large-scale catastrophes, although neither connect their

---

19. *See* ORD, *supra* note 1, at 4–5; Holden Karnofsky, *The 'Most Important Century' Blog Post Series*, COLD TAKES (Sept. 19, 2021), https://www.cold-takes.com/most-important-century/ [https://perma.cc/87MM-GAY4]. *C.f.* William MacAskill, *Are We Living at the Hinge of History?* 2–4 (Glob. Priorities Inst., Working Paper No. 12, 2020), https://globalprioritiesinstitute.org/wp-content/uploads/William-MacAskill_Are-we-living-at-the-hinge-of-history.pdf [https://perma.cc/589Z-NPUS] (weighing arguments that we are likely to be living in the most important century).

20. *See* ORD, *supra* note 1, at 1–2.

21. Organizations established at Oxford University include the Future of Humanity Institute and the Global Priorities Institute. Nearby Cambridge University is home to the Centre for the Study of Existential Risk. In the United States, organizations focused on existential risk include the Future of Life Institute and the Global Catastrophic Risk Institute. Grantmaking organizations include Open Philanthropy, Longview Philanthropy, and the Berkeley Existential Risk Initiative.

22. These student-led initiatives involve a range of activities including summer fellowships, conferences, and reading groups. *See generally* STAN. EXISTENTIAL RISKS INITIATIVE, https://seri.stanford.edu [https://perma.cc/7VRH-UZMN] (last visited Sept. 23, 2023); HARV.-MIT X-RISK, https://harvardmitxrisk.org [https://perma.cc/4CH3-GKNQ] (last visited Sept. 23, 2023).

23. *See* Naina Bajekal, *Want to Do More Good? This Movement Might Have the Answer*, TIME (Aug. 10, 2022), https://time.com/6204627/effective-altruism-longtermism-william-macaskill-interview [https://perma.cc/WC3U-8TTN] (noting that, as of August 2022, the funding committed to Effective Altruism, where existential risk figures prominently as a leading cause area, far exceeds that raised by the Ford Foundation (roughly $16 billion) and the Rockefeller Foundation (roughly $6 billion)). Funding in this field has declined precipitously over the past year with the downfall of FTX and the decline in the net worth of other funders. Total funding committed still likely exceeds $10 billion. *See* Benjamin Todd, *Is Effective Altruism Growing? An Update on the Stock of Funding vs People*, 80,000 HOURS (July 28, 2021), https://80000hours.org/2021/07/effective-altruism-growing [https://perma.cc/PV6U-H9VQ].

24. *See* discussion *infra* Part III.A.

work to the literature on existential risk.[25] An international survey found that legal scholars generally believe that legal action taken today can impact existential risk and the long-term future.[26] Moreover, the participants in this study have identified a new landscape of promising legal interventions, spanning from local efforts to prevent specific threats,[27] to far-reaching efforts to establish the legal interests of future generations.[28] As researchers in this field have reported, future generations are now referenced in eighty-one national constitutions[29] and in a host of domestic laws and international agreements.[30] Moreover, some courts have recently shown an unprecedented willingness to enforce these provisions in the context of climate change litigation.[31] Perhaps the most striking example is the 2021 German Constitutional Court decision striking down a national climate law, citing "intertemporal guarantees of freedom" and a "special duty of care . . . for the benefit of future generations."[32]

There are also promising signs in the domain of legislation and policy. New parliamentary groups have been formed to protect future generations from the impacts of low-probability/high-impact catastrophes.[33] In the United States, the

---

25. *See* RICHARD POSNER, CATASTROPHE: RISK AND RESPONSE (2004); CASS SUNSTEIN, WORST CASE SCENARIOS (2007); CASS SUNSTEIN, AVERTING CATASTROPHE (2021); *see also* CHRISTOPH WINTER, JONAS SCHUETT, ERIC MARTINEZ, VAN ARSDALE, RENAN ARAUJO, NICK HOLLMAN, JEFF SEBO, ANDREW STEWASZ, CULLEN O'KEEFE & GIULIANA ROTOLA, LEGAL PRIORITIES RESEARCH: A RESEARCH AGENDA 5, 33–4 (2021), https://www.legalpriorities.org/research_agenda.pdf [https://perma.cc/973D-EUFP] (observing that legal scholars have paid little attention to existential risk).

26. *See* Eric Martínez & Christoph Winter, *Protecting Future Generations: A Global Survey of Legal Academics* (Legal Priorities Project, Working Paper No. 1-2021, 2021), https://ssrn.com/abstract=3931304 [https://perma.cc/55PN-54H6] (last visited Dec. 4, 2023).

27. For example, suing an AI lab to prevent the release of a risky product.

28. For example, judicial recognition of personhood, standing, or a right to life.

29. *See* Renan Araújo & Leonie Koessler, *The Rise of the Constitutional Protection of Future Generations* (Legal Priorities Project, Working Paper No. 7-2021, 2021), https://ssrn.com/abstract=3933683 [https://perma.cc/6A9H-JWLU] (last visited Sept. 14, 2023).

30. *Id.* at 4 (referencing, *inter alia*, the concern expressed in the United Nations Charter for "generations to come," the influential 1974 Swedish constitutional amendment protecting the environment in light of concerns for future generations, and the 1997 UNESCO Declaration on the Responsibilities of the Present Generations Towards Future Generations).

31. *See* BVerfG, 1 BvR 2656/18, Mar. 24, 2021, https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/2021/03/rs20210324_1bvr265618en.html [https://perma.cc/P73Z-3KH6]; *see also* HR 20 december 2019, NJ 1959, 19/00135 m.nt Engels (State of the Netherlands/Urgenda) (Neth.) (requiring emissions reductions on the grounds of right to life and right to respect for private and family life from Articles 2 and 8 of the European Convention on Human Rights and well as the UNFCC provision to "protect the climate system for the benefit of present and future generations of humankind"); Rechtbank Den Haag 26 mei 2021, HA ZA 2021, 19-379 (Vereniging Milieudefensie et. al/Royal Dutch Shell PLC) (Neth.) (applying a "standard of care" as a human rights obligation that applies to future generations). *Cf.* Juliana v. United States, 947 F.3d 1159, 1175 (9th Cir. 2020) (finding that a group of twenty-one young plaintiffs lacked Article III standing for lack of redressability and dismissing all claims).

32. *See* 1 BvR 2656/18, *supra* note 31.

33. *See, e.g.*, ALL-PARTY PARLIAMENTARY GROUP FOR FUTURE GENERATIONS, https://www.appgfuturegenerations.com/about [https://perma.cc/6AJX-PNVV] (last visited Sept. 15, 2023) (aiming to establish national wellbeing goals relating to future generations); *Wellbeing of Future Generations Bill [HL]*, U.K. PARLIAMENT (Feb. 8, 2022), https://bills.parliament.uk/bills/2869 [https://perma.cc/A5JP-3Q8U]; Colum Lynch,

Global Catastrophic Risk Management Act (passed in December 2022) requires federal agencies to assess and mitigate "existential risk," which is defined in the Act as risk with the "potential for an outcome that would result in human extinction."[34] The British Prime Minister, Rishi Sunak, has acknowledged existential risk relating to artificial intelligence.[35] Nelson Mandela's world peace organization, The Elders, has framed existential risk as a central concern for humanity.[36] The Secretary General of the United Nations has issued a major report on existential risk and related threats to future generations.[37] The U.N. has scheduled a "Summit of the Future" for 2024, which is expected to bring together heads of state from around the world to establish a U.N. Declaration on Future Generations, a Special Envoy for Future Generations, and an obligation to issue reports on global catastrophic risks.[38]

Regarding existential threats from AI, governments have recently started taking an array of actions. President Biden issued an executive order on "Safe, Secure, and Trustworthy" AI, which directed several federal agencies to address AI risks and established the United States AI Safety Institute.[39] The U.K. has also established an AI Safety Institute, and the U.S. and the U.K. have created the first bilateral agreement on AI risk, enabling their respective safety institutes to share information and evaluation tools.[40] The E.U. has passed an act banning high-risk AI systems and creating specific regulations for "general purpose AI," including limits on training runs above 10^25 floating point operations, a standard that the

---

DevExplains: In Short, Longtermism Has Arrived, DEVEX (Dec. 22, 2022), https://www.devex.com/news/devexplains-in-short-longtermism-has-arrived-104626 [https://perma.cc/7JGA-BNES] (referencing institutions to protect future generations within the governments of Singapore, Finland, the United Arab Emirates, and Sweden).

34. The Global Catastrophic Risk Management Act of 2022, H.R. 7776, 117 Cong. (2022).

35. Alex Hern & Kiran Stacey, *No 10 Acknowledges 'Existential' Risk of AI for First Time*, THE GUARDIAN (May 25, 2023), https://www.theguardian.com/technology/2023/may/25/no-10-acknowledges-existential-risk-ai-first-time-rishi-sunak [https://perma.cc/3NVN-KEVY].

36. *The Elders' New Strategy Sets Out to Address Humanity's Existential Threats*, THE ELDERS (Jan. 24, 2023), https://theelders.org/news/elders-new-strategy-sets-out-address-humanity-s-existential-threats [https://perma.cc/99JP-77VH].

37. *See* ANTÓNIO GUTERRES, UNITED NATIONS, OUR COMMON AGENDA: REPORT OF THE SECRETARY-GENERAL 27, 64 (2021), https://www.un.org/en/content/common-agenda-report/assets/pdf/Common_Agenda_Report_English.pdf [https://perma.cc/G2VK-73R6] (referencing "solidarity" between current and future generations); *see also* UNITED NATIONS DEVELOPMENT PROGRAMME, HUMAN DEVELOPMENT REPORT (2020) (including an essay on existential risk written by Toby Ord).

38. *The Summit of the Future in 2024*, U.N. (Mar. 30, 2023), https://www.un.org/en/ common-agenda/summit-of-the-future [https://perma.cc/BED6-BJLX].

39. Exec. Order No. 14110, 88 Fed. Reg. 75,191 (Oct. 30, 2023) (requiring federal agencies to enforce new standards for safety testing of new AI systems, limiting military use of AI, and drawing attention to AI-related threats to "critical infrastructure" as well as chemical, biological, radiological, nuclear, and cybersecurity risks).

40. *See* U.K. DEP'T. FOR SCI., INNOVATION AND TECH., INTRODUCING THE AI SAFETY INSTITUTE (2023); *U.S. and UK Announce Partnership on Science of AI Safety*, U.S. DEP'T. OF COMMERCE (Apr. 1, 2024), https://www.commerce.gov/news/press-releases/2024/04/us-and-uk-announce-partnership-science-ai-safety [perma.cc/DN35-MYWL].

Act recommends updating based on risk assessments.[41] The Council of Europe has recently agreed on the "first rules for AI in the world", emphasizing safety and broad human rights protections.[42] The U.N. has also passed a resolution on AI.[43] Further legislation has been proposed in many U.S. jurisdictions requiring safety testing for frontier AI systems, as well as addressing intersections between AI and biological and nuclear weapons.[44] Other proposals would establish licensing regimes, mandatory impact assessments, third-party auditing, and liability and insurance frameworks for the risks imposed by AI companies.[45]

The field of advocates examined in this article has taken a leading role in many of these and related efforts. They have advised national governments and international policy organizations and consulted with lawyers pursuing litigation relevant to existential risk (e.g., by finding liability for imposing existential risk as a matter of tort, human rights, or criminal law). To preserve confidentiality, I will only share details about specific advocacy projects where I have permission from research participants.

At the same time that these advocates have made remarkable progress, they also have been the subject of widespread backlash.[46] Critical reactions to the movement intensified in late 2022 following the downfall of a major donor, Sam Bankman-Fried, who had pledged much of his wealth—once valued at $29 billion—to the mitigation of existential risk but has now been convicted on charges of fraud and conspiracy.[47] The common narrative of mass and social media

---

41. European Commission Interinstitutional File 2021/0106 (COD), Presidency, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - Analysis of the final compromise text with a view to agreement (Jan. 26, 2024).

42. European Council Press Release 986/23, The Council, Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world (Sept. 12, 2023).

43. *General Assembly adopts landmark resolution on artificial intelligence*, United Nations (Mar. 21, 2024), https://news.un.org/en/story/2024/03/1147831 [https://perma.cc/V4RL-NWJY].

44. U.S. Artificial Intelligence Policy: Legislative and Regulatory Developments, COVINGTON (Oct. 20, 2023), https://www.cov.com/en/news-and-insights/insights/2023/10/us-artificial-intelligence-policy-legislative-and-regulatory-developments [perma.cc/WG2A-AZ3V].

45. EDOUARD HARRIS, JEREMIE HARRIS & MARK BEALL, DEFENSE IN DEPTH: AN ACTION PLAN TO INCREASE THE SAFETY AND SECURITY OF ADVANCED AI; Gabriel Weil, *Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4694006 [perma.cc/Q3TM-3MWV].

46. *See* Kieran Setiya, *The New Moral Mathematics*, BOS. REV. (Aug. 15, 2022), https://www.bostonreview.net/articles/the-new-moral-mathematics/ [https://perma.cc/Z4WV-Z3NW]; Emile P. Torres, *The Dangerous Ideas of "Longtermism" and "Existential Risk,"* CURRENT AFFAIRS (July 28, 2021), https://www.currentaffairs.org/2021/07/the-dangerous-ideas-of-longtermism-and-existential-risk [https://perma.cc/LV4Q-XXFP].

47. *See* Luc Cohen & Jody Godoy, *Bankman-Fried sentenced to 25 years for multi-billion dollar FTX fraud*, REUTERS (Mar. 28, 2024), https://www.reuters.com/technology/sam-bankman-fried-be-sentenced-multi-billion-dollar-ftx-fraud-2024-03-28/#:~:text=NEW%20YORK%2C%20March%2028%20 (Reuters,former%20billionaire%20wunderkind's%20dramatic%20downfall [https://perma.cc/3K2W-JYAV]; *see also* Gerrit De Vynck, *U.S. Charges FTX Founder Sam Bankman-Fried with Criminal Fraud*, WASH. POST (Dec. 13, 2022), https://www.washingtonpost.com/technology/2022/12/13/sbf-sec-fraud-charges [https://perma.cc/Q722-RYGM].

suggests that advocating for the mitigation of existential risk tends to, perhaps by design, distract from the injustice and suffering that already exists in the world, and instead directs attention toward the techno-utopian fantasies of billionaire donors.[48] But these public reactions, whatever the ultimate merit of the critiques being raised, are based on a very shallow and deeply inaccurate view into the existential risk community, and generally no view at all into the sub-community of legal and political advocates. This article provides an empirical window into this field of advocacy drawing from ethnography, semi-structured interviews (n=53), and a systematic review of online materials.

The participants in this study included members and affiliates of over two dozen organizations.[49] Since my aim is to understand the role of law and lawyers within this movement, the study centered on the Legal Priorities Project (hereinafter "LPP"). Founded in 2020 as a 501(c)(3) by students and a visiting professor at Harvard Law School, LPP is the only legal organization in the world focused on existential risk.[50] In its first year of operation, LPP wrote a research agenda and began to build a network of young lawyers and law students via summer fellowships and speaker programs.[51] In their second year, they shifted attention to policy advising, providing a legal perspective on new legislative and regulatory efforts relating to existential risk with a focus on AI.[52] During the ethnography, LPP formed an impact-litigation team, hiring a Costa Rican human rights attorney, a Dutch judge, and an American public interest lawyer.[53] This litigation team recently shifted toward a greater focus on providing legal guidance on legislative and regulatory proposals.[54] LPP is a highly global organization—the membership has a clear tilt toward Europe and the United States, but over the period of ethnography, the roughly twelve members who attended LPP's internal all-hands meetings included nationalities and legal training from Africa, Australia,

---

48. *See* Jennifer Szalai, *How Sam Bankman-Fried Put Effective Altruism on the Defensive*, N.Y. TIMES (Dec. 13, 2022), https://www.nytimes.com/2022/12/09/books/review/effective-altruism-sam-bankman-fried-crypto.html [https://perma.cc/GJJ5-BU5M] (suggesting that scholars of existential risk generally hold that "considerations of immediate need pale next to speculations about existential risk—not just earthly concerns about climate change and pandemics but also . . . more extravagant theorizing about space colonization and A.I."); Emily Frey & Noah Giansiracusa, *The Moral Failing of 'Effective Altruism'*, BOS. GLOBE (Nov. 22, 2022), https://www.bostonglobe.com/2022/11/22/opinion/moral-failing-effective-altruism [https://perma.cc/C4BP-RE6W] (suggesting that tech billionaires may tend to be especially drawn to the notion that emerging technology poses a major threat to the future of humanity because this issue is framed as matter relating to their technological expertise).

49. In total, interview participants were employed full-time by eighteen different organizations, while some participants had part-time positions and affiliations at other organizations.

50. Jonas Schuett & Legal Priorities Project, *Introducing the Legal Priorities Project*, EFFECTIVE ALTRUISM F. (Aug. 30, 2020), https://forum.effectivealtruism.org/posts/PvBLDPkqKvdHQkKPn/introducing-the-legal-priorities-project [https://perma.cc/6UND-95YY].

51. *See* WINTER ET AL., *supra* note 25; *see also id.*

52. *See* Schuett, *supra* note 50; *see also* Parra, *infra* note 53.

53. *See* Alfredo Parra & Cristoph Winter, *Annual Report 2022*, LEGAL PRIORITIES PROJECT (May 2, 2023), https://www.legalpriorities.org/blog/2023/lpp-annual-report-2022/ [https://perma.cc/L7Z4-5DGP].

54. *See* discussion *infra* Part III.

Central America, Continental Europe, North America, and South America.[55] Most LPP operations take place remotely with occasional in-person meetings. Thus, the study was largely conducted via video-conferencing calls in addition to in-person visits to key sites including Oxford, London, Geneva, Washington, and Cambridge (Massachusetts).

The main empirical finding of this article is that these advocates have developed a distinct model of social-change lawyering—the "priorities methodology" drawn from the philanthropic framework of Effective Altruism. As elaborated in Part III, this model is an effort to help the most people to the greatest extent, considering the magnitude and probability of the estimated net effects of one's actions. The model begins with first principles of morality, such as enhancing overall human well-being or reducing suffering. Advocates then select cause areas based on certain criteria (see the Part III discussion of importance, neglect, and tractability as key criteria) in an effort to maximize impact on their moral commitments. Existential risk is prioritized by the advocates examined in this Article because it is a neglected issue that could severely affect a great number of people. None of the advocates I met claimed that the world is highly likely to end in the near future, but they do believe that decision theory and public policy considerations favor intervening in existential threats given the non-negligible likelihood of grave harms.

Having selected a cause area, these advocates then analyze strategic decisions in an effort to determine which option has the greatest expected value toward their chosen goals. As described in anthropological detail in Part V, this model's commitment to optimizing impact is supported by formal decision-making processes as well as a daily culture of reinforcing scientific, truth-seeking norms, including (1) embracing epistemic humility and normalizing uncertainty, (2) fostering a warm, inclusive, and empathetic atmosphere of "supportive dissent," and (3) limiting cognitive biases that would interfere with their focus on maximizing impact. In practice, these advocates are remarkably adherent to this set of cultural norms, although they acknowledge that these norms conflict with some aspects of human nature and the professional identities of lawyers. Moreover, this model raises difficult points of tension when seeking to represent the interests of voiceless future generations while also working to incorporate a broader set of current-person voices in strategic decision-making processes. The full model is depicted in Figure 1.

55. It should be noted that this degree of geographic diversity is perhaps uncommon among organizations in this field. This issue is discussed *infra* Part IV.

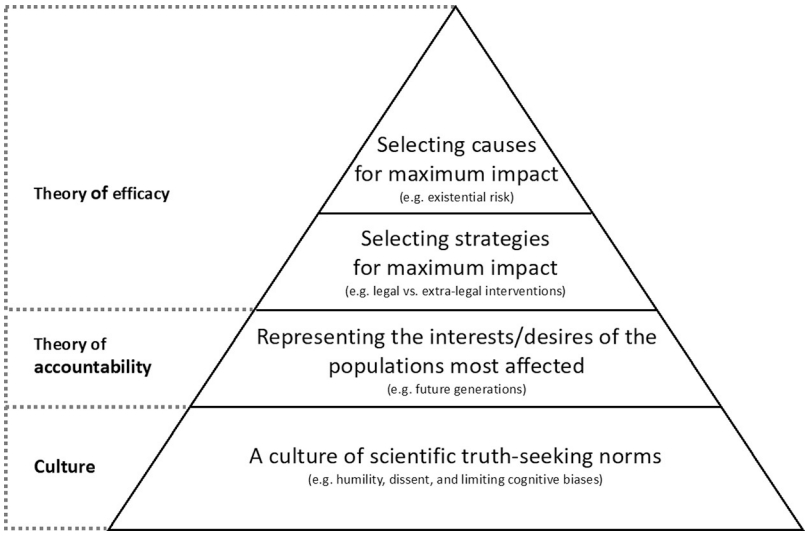

FIGURE 1. The Priorities Methodology for Social-Change Lawyering

In some respects, the priorities methodology is consistent with leading recommendations from the literature on law and social change. Empirical research has suggested that many public interest lawyers lack explicit theories of change, formal processes for setting priorities, and measures of their performance.[56] Scholars have called on social-change lawyers to adopt more rigorous approaches to strategic planning and assessment.[57] The existential advocates seem to take up this call very directly, even taking it to an extreme with their commitment to maximizing their overall counterfactual impact.

But these advocates depart from the literature's recommendations in one key respect: While they recognize that favorable public opinion and grassroots mobilization have supported the accountability and long-term efficacy of legal activism in other contexts, these advocates tend to worry that "going broad" would compromise their priorities methodology and politicize the issue of existential risk. The cultural values outlined in this article (relating to the value of dissent, uncertainty, and "system 2" deliberative reasoning rather than strong emotional motivations) are seemingly point-by-point the exact opposite of the "mobilizing frames" that scholars have found to be the essential ingredients for building a broad social movement.[58] This observation reveals the central point of tension

56. *See generally* Deborah L. Rhode, *Public Interest Law: The Movement at Midlife*, 60 STAN. L. REV. 2027, 2049–53 (2008) (finding that only 14% of public interest law organizations undertake "extensive" formal decision-making processes).

57. *See, e.g.*, *id.* at 2057.

58. *See* David A. Snow, Rens Vliegenthart & Pauline Ketelaars, *The Framing Perspective on Social Movements: Its Conceptual Roots and Architecture*, *in* THE WILEY BLACKWELL COMPANION TO SOCIAL MOVEMENTS 392 (David A. Snow, Sarah A. Soule, Hanspeter Kriesi & Holly J. McCammon eds., 2nd ed. 2018).

explored throughout this article. These advocates seek to enhance their inclusiveness and democratic legitimacy without undermining what their community does so well—overcoming cognitive biases to maintain a focus on the neglected issue of protecting future generations from low-probability/high-impact events.

Before detailing these empirical findings, Part I provides background regarding (1) the issue of existential risk and (2) the socio-legal literature examining how lawyers contribute to movements for social change. Part II describes the methods of this study. Parts III through V present this Article's empirical portrait of existential advocacy and the priorities methodology, beginning with theories of efficacy and accountability, and then discussing the movement's underlying cultural commitments. Part VI concludes the article with recommendations for adapting the priorities methodology as this movement scales up and pursues more direct and high-profile legal interventions.

I. BACKGROUND

A. THE PROBLEM OF EXISTENTIAL RISK

On the scale of human history, it is only very recently that we have developed technologies capable of foreclosing the human future.[59] Nick Bostrom, the Oxford philosopher who coined the term "existential risk," and defined it to include threats of human extinction or other irreversible destruction of human potential, notes that until the advent of nuclear weapons there were "probably no significant existential risks in human history . . . and certainly none that it was within our power to do something about."[60] Researchers in the emerging field of existential risk studies have now spent over two decades analyzing a wide range of conceivable threats, which has led many in this field to the conclusion that existential risk should be assigned a substantial probability.[61]

Ord suggests we are entering a new historical era, which he labels "the Precipice," marked by our initial meeting with existential threats.[62] In Ord's optimistic framing, this era may (hopefully) be remembered as the time when humanity "open[ed] its eyes" to existential risks and guaranteed a "long and flourishing future" through transformations in our legal, political, normative, and cognitive

59. *See* THOMAS MOYNIHAN, X-RISK: HOW HUMANITY DISCOVERED ITS OWN EXTINCTION (2020) (discussing religious traditions of apocalyptic prophecy, which portend a "sense of an ending," but explaining that existential risk, and human extinction in particular, is a "comparatively novel idea" involving an "ending of sense," a more complete cessation of human experience).

60. Nick Bostrom, *Dinosaurs, Dodos, Humans?*, 2006 GLOB. AGENDA 230, 230.

61. *See id.* (defining existential risks as threats to either "annihilate Earth-originating intelligent life or permanently and drastically curtail its potential"); ORD, *supra* note 1 (defining existential risk to include extinction, locked-in totalitarianism ("world in chains"), and irreversible collapse of civilization "where humanity across the globe loses civilization, a world without writing, cities, law, or any trappings of civilization"); *see generally Policy Idea Database*, GLOB. CATASTROPHIC RISK POL'Y, https://www.gcrpolicy.com/ideas [https://perma.cc/4UVS-SRK2] (last visited Oct. 2, 2023) (showing a total of 111 publications in the field of existential and catastrophic risk studies with an increase after 2017 to 75 to 82 publications per year).

62. ORD, *supra* note 1.

frameworks."[63] Ord offers a set of best-guess estimates of how likely it is that we would experience different existential events.[64] While conceding that assigning these probabilities is highly speculative, he notes that it can be a helpful exercise in comparing different threats and avoiding the human tendency to dismiss risks that are vaguely described as "low-probability" or "unlikely."[65]

Although climate change and nuclear weapons pose severe risks to the planet and humanity, Ord notes that models of these threats suggest a relatively small chance of reaching the existential scale. He assigns these categories only a 1/1000 likelihood of irreversibly foreclosing the human future over the next century.[66] Existential threats from natural sources are similarly considered quite unlikely to materialize in the near future.[67] Although asteroids loom large in fictional accounts, as well in prehistory as the source of the Cretaceous–Paleogene Extinction, this category is now well studied.[68] By looking to the skies and to the earth (via the geological record), scientists have shown that such events occur infrequently and thus appear to pose a relatively slight danger in the near term.[69]

Ord suggests a dramatic contrast between the probability of natural and anthropogenic existential risks. For example, he estimates a 1/10,000 probability that natural pandemics will bring about an existential catastrophe over the next century, while engineered pandemics are assigned a probability of 1/30.[70] Moreover, engineered pandemics are part of a growing category of concerns relating to synthetic biology, including bioweapons, pathogens escaping laboratories, information hazards (where information required to create dangerous biological materials is not kept confidential), and under-regulation of the DNA synthesis industry.[71] Other frontiers of scientific endeavor could also pose existential threats, such as

---

63. *Id.* at 31.

64. *Id.*

65. *Id.*

66. ORD, *supra* note 1 (noting that nuclear weapons and climate change "awoke us to the possibilities of destroying ourselves," and noting that the threat of climate change could grow to existential proportions with a runaway greenhouse effect that boils the oceans or sets off a cascade of ecosystem failures, although experts tend to suggest that such scenarios are highly unlikely). Existential risk is often viewed as a subset of the broader category of global catastrophic risk, which refers to events that would cause widespread harm but may lack the "terminal" intensity of existential risks. *See generally* Nick Bostrom & Milan M. Ćirković, *Introduction*, *in* GLOBAL CATASTROPHIC RISKS 1, 2 (2020) (defining "global catastrophe" as "a catastrophe that [would cause] 10 million fatalities or 10 trillion dollars worth of economic loss").

67. ORD, *supra* note 1.

68. *See* Adam Mann, *Odds of Death by Asteroid? Lower Than Plan Crash, Higher Than Lightning*, WIRED (Feb. 15, 2013), https://www.wired.com/2013/02/asteroid-odds [https://perma.cc/DY25-MMG5] (reporting on findings from NASA's Near Earth Objects Observations Program, which examines the risk of asteroid impacts).

69. *See* Bostrom & Ćirković, *supra* note 66, at 5–6.

70. *See* ORD, *supra* note 1, at 162 (reviewing the long history of the use of disease as a weapon, as well as examples of lab escapes and information hazards including the publication of the smallpox genome).

71. *Id.*; *see also* Kelsey Piper, *It's Time to Close the Gene Synthesis Loophole That Could Lead to a Human-Made Pandemic*, VOX (July 27, 2023), https://www.vox.com/future-perfect/2023/7/27/23808920/gene-dna-synthesis-biotechnology-pandemic-viruses-twist-bioscience-pathogens-ginkgo-bioworks [https://perma.cc/GN2R-LRJR].

bringing back unpredictable materials from space exploration, or running "radical scientific experiments" with unknown risks that could theoretically reach the scale of destroying life on Earth and beyond.[72]

But for Ord, along with most scholars in this field, the greatest single category of existential risk relates to the development of transformative artificial intelligence ("AI"), estimated by Ord as a 1/10 existential threat over the next century.[73] A recent survey found that nearly half of AI researchers offer a similar estimate of AI's existential threat.[74] With new developments in deep learning, a growing number of technology forecasters believe that we may be approaching the horizon of artificial general intelligence ("AGI"), which is variously defined as AI that meets or exceeds human-level performance across a wide range of cognitive tasks.[75] Some define it as the moment when humans are no longer the most intelligent beings on the planet.[76] One prominent aggregator of expert predictions in emerging technology has seen a dramatic shortening of AGI timelines over the past year—with forecasters moving from an average prediction of the year 2057 to the year 2031.[77] This shortening of the AGI timeline is likely influenced by recent advances in generative AI—systems that write essays, poetry, and music, create original images and video, and assist in scientific inquiries—and systems that can outperform humans at standardized tests and games of increasing complexity (e.g., chess and Go) as well as games based on psychology and stratagem (e.g., poker and Diplomacy).[78]

---

72. *See* ORD, *supra* note 1, at 93, 156 (discussing "radical scientific experiments" and citing the example of the advent of nuclear weapons when some scientists theorized that the first detonation would ignite Earth's atmosphere and set off an existential catastrophe).

73. *Id*. at 138–42.

74. *See* Stein-Perlman & Grace, *supra* note 13.

75. *See* Ben Goertzel, *Artificial General Intelligence: Concept, State of the Art, and Future Prospects*, 5(1) J. ARTIFICIAL GENERAL INTELLIGENCE 1, 1 (2014) (discussing the origins of the term "Artificial General Intelligence," which refers to human level "thinking machines" and has been defined in terms influenced by the fields of mathematics, engineering, and biology).

76. *See* Nick Bostrom, *What Happens When Our Computers Get Smarter than We Are?*, YOUTUBE (Apr. 27, 2015), https://www.youtube.com/watch?v=MnT1xgZgkpk [https://perma.cc/N97J-AAW4]; Kelsey Piper, *OpenAI Wants to Build Systems Smarter than Us*, VOX (Mar 1, 2023), https://www.vox.com/future-perfect/23619354/openai-chatgpt-sam-altman-artificial-intelligence-regulation-sydney-microsoft-ai-safety [https://perma.cc/Q6E7-RT26].

77. *See* Matthew Barnett, *When Will the First General AI System Be Devised, Tested, and Publicly Announced?*, METACULUS (Aug. 23, 2020), https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/ [https://perma.cc/HG5H-VBLW] (asking when the first "general AI system [will] be devised, tested, and publicly announced" where "general AI" is defined as passing a two-hour multimodal adversarial Turing Test); *see also* Keith Wynroe, David Atkinson & Jaime Sevilla, *Literature Review of Transformative Artificial Intelligence Timelines*, EPOCH (Jan. 17, 2023), https://epochai.org/blog/literature-review-of-transformative-artificial-intelligence-timelines [https://perma.cc/M7AJ-Y6EJ] (reporting median estimates of the arrival of artificial general intelligence from five different surveys, centering on the years 2039, 2040, 2043, 2059, 2060, also noting "model-based" estimates centering on the year 2089).

78. *See* Seb Krier, *AI from Superintelligence to ChatGPT*, WORKS IN PROGRESS (Dec. 8, 2022) https://www.worksinprogress.co/issue/ai-from-superintelligence-to-chatgpt/ [https://perma.cc/W2SK-AKVB].

Advanced AI could give rise to existential risks under a number of malicious and accidental scenarios, such as where the technology is used to create and deploy biological, chemical, and autonomous weapons, or where AI interferes with political processes and encourages great-power wars, or where AI systems are inherently unsafe. The risk that AI will be developed without proper safety precautions may be heightened by emerging race dynamics as companies and governments compete to achieve "AI dominance."[79] An even more fundamental issue is that it may be difficult or impossible, even with our best efforts, to align these systems with human values and interests. An existential risk might arise if we fail to develop AI so that it reliably respects, at a minimum, human life and humanity's interest in avoiding locked-in dystopic futures. One of the concerns arising around efforts to reduce algorithmic bias (e.g., AI outputs that demonstrate bias on the grounds of race or gender) is that this may be a sign of the more general challenge of aligning transformative AI systems with human values. This challenge is exacerbated by the further question of how to determine a set of good human values and who should make this determination.[80] It is also worth noting here that Ord and others argue that transformative AI may have both positive and negative effects on other existential threats.[81] For instance, AI systems may help us develop tools to deal with problems like climate change (although AI may also accelerate climate change given the industry's energy demands) while also facilitating scientific advances that give rise to new threats on the existential scale.[82]

Finally, the notion of "unforeseen" risks may seem vague and speculative, but some consider this category the most alarming. In his writing on the "vulnerable world hypothesis,"[83] Bostrom posits that humanity's never-ending practice of drawing out new inventions from the metaphorical "urn of creativity" may eventually yield a "black ball" technology, that is, a technology that "invariably or by default destroys the civilization that invents it."[84] An example could be a device with the power of nuclear weapons but an ease of assembly that requires only a commercially available 3D printer or the equipment in a typical garage.[85] It is perhaps only by good fortune that humanity has not yet discovered a black ball,

79. *See generally* Wim Naudé & Nicola Dimitri, *The Race For An Artificial General Intelligence: Implications for Public Policy*, 35 AI & SOC'Y 367, 367 (2020); Kathleen Walch, *Why the Race for AI Dominance is More Global Than You Think*, FORBES (Feb. 9, 2020), https://www.forbes.com/sites/cognitiveworld/2020/02/09/why-the-race-for-ai-dominance-is-more-global-than-you-think [https://perma.cc/66A6-KPNU] (describing the Chinese political leadership's 2030 goal to achieve "AI dominance," and recent U.S. efforts to limit Chinese access to high-end semiconductors).

80. *See generally* Iason Gabriel, *Artificial Intelligence, Values, and Alignment*, 30 MINDS & MACHINES 411, 411 (2020); NICK BOSTROM, SUPERINTELLIGENCE (2014).

81. ORD, *supra* note 1, at 138–49.

82. *See id.*

83. *See* Nick Bostrom, *The Vulnerable World Hypothesis*, 10(4) GLOB. POL'Y 455, 455 (2019).

84. *Id.*

85. *Id.* at 455–56 (exploring the possibility of apocalyptic technology created "with a piece of glass, a metal object, and a battery arranged in a particular configuration").

and it may be only a matter of time before we do. If transformative AI greatly accelerates scientific and technological development (via "PASTA" a "Process for Automating Scientific and Technological Advancement"), the invention of black-ball technologies may grow increasingly likely.[86]

Estimates of existential risk are based not only on assessing the destructive capacities of various technologies, but also crucially on assessing our ability to prevent catastrophic use of those technologies. Ord notes that his 1/6 estimate of existential risk this century assumes that we would cut existential risk by half because we would, in the coming decades, "get our act together and start taking these risks very seriously."[87] Any such awakening in our political and legal systems is inhibited by a wide range of cognitive biases that make it difficult to recognize the scope of existential risk. To cite just a few key examples, as humans we generally find it difficult to imagine events on a scale we have never seen before (the availability heuristic), using a moral lens that is evolutionarily tuned to small-scale and nearby harms (scope neglect) that befall known individuals (the identifiable victim effect) who are alive today (present bias).[88] We tend to dismiss threats of low probability events unless our emotions are primed.[89] In popular understandings of existential risk, these cognitive biases may be exacerbated by the association of existential risk with science fiction, irrational doomsayers, "preppers" (people who gather resources and construct shelters in anticipation of catastrophes), and, in a common caricature of the existential risk community, hyper-rational "Silicon Valley tech bros."[90]

In the political realm, these cognitive biases may be exacerbated by short-term incentives that prioritize currently living (and voting and lobbying) people over

---

86. *See* Holden Karnofsky, *Forecasting Transformative AI, Part 1: What Kind of AI?*, COLD TAKES (Aug. 10, 2021), https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai [https://perma.cc/A27S-DAR8].

87. ORD, *supra* note 1, at 170.

88. *See* Tyler M. John & William MacAskill, *Longtermist Institutional Reform* 5–6 (Glob. Priorities Inst., Working Paper No. 14-2020, 2020) (discussing cognitive biases that limit our concern for risks that affect future generations); CASS R. SUNSTEIN, BEHAVIORAL SCIENCE AND PUBLIC POLICY 3 (2020) (noting that our concern is generally diminished when catastrophic threats are framed as a risk primarily to future generations—those living in "Laterland"); WINTER ET AL., *supra* note 25, at 17 (arguing that human cognition tends to be limited in its ability to consider "the vastness of the future, in particular . . . human extinction scenarios").

89. *See* CASS SUNSTEIN, WORST CASE SCENARIOS, *supra* note 25 (describing the general lack of emotional response around threats that are rare or unprecedented, and our tendency to overreact when low-probability events are emotionally salient, such as when similar events have recently occurred); CASS SUNSTEIN, AVERTING CATASTROPHE, *supra* note 25 (observing that some catastrophes are the result of exponential rather than linear growth, which leads to under-reaction due to "exponential growth neglect"); Eliezer Yudkowsky, *Cognitive Biases Potentially Affecting Judgement of Global Risks*, *in* GLOBAL CATASTROPHIC RISKS, 20 (Nick Bostrom & Milan M. Ćirković eds., 2d ed. 2020) (summarizing relevant cognitive biases and noting that humanity tends to "overestimate the predictability of the past and underestimate the surprise of the future").

90. *See* JOSHUA SCHUSTER & DEREK WOODS, CALAMITY THEORY: THREE CRITIQUES OF EXISTENTIAL RISK 3 (2021) (observing that "existential risk theory has been conducive to a warm reception by a 'tech bro' Silicon Valley audience."); POSNER, *supra* note 25, at 100–10 (arguing that irresponsible doomsday predictions can lead others to dismiss the risk of large-scale catastrophe, but noting that thoughtful science fiction can also help to illuminate these threats).

future generations.[91] Moreover, any effort to mitigate existential risk must overcome collective action problems associated with goods that are public, global, and intergenerational.[92] These political conditions could be made far more perilous by "existential risk factors" in the coming years, such as great power wars, extreme environmental impacts, or other events that make it less likely that we as a global community will be willing and able to work together to address threats on the existential scale.

But maybe these "biases" against concerning ourselves with existential risk are actually pointing toward something true. Does existential risk really matter? This question has both empirical and normative dimensions. This section has so far focused on the empirical side—assessing the probability that an existential event will occur. If this probability is extremely low (or extremely high and unavoidable), we might have little reason to invest in the prevention of existential threats. The normative dimension asks whether an existential event would be undesirable. One response from participants in this study is that existential risks threaten to harm a great number of people who actually exist.[93] This includes people whose lives would be ended (perhaps involving immense suffering) by an existential catastrophe or whose lives would be made much worse in an unrecoverable dystopia scenario. Moreover, efforts to mitigate existential risk greatly overlap with efforts to mitigate sub-existential risk—and these smaller catastrophes may be more likely to occur in the near-term and would have devastating effects on people currently living.

The more complex normative response considers the impact of human extinction on future generations, where the impact would be "felt" by people who would not have the opportunity to exist. A full treatment of this issue and related theories of population ethics is beyond the scope of this article.[94] This issue is the subject of a growing field of philosophical inquiry known as "longtermism," which considers the moral weight of future generations.[95] But it is worth noting briefly that many participants in this study analyze this question in the following terms: potential future persons could outnumber us (current persons) to such a

---

91. *See* Tyler John, *Representing Future Generations*, YOUTUBE (Mar. 21, 2020), https://youtu.be/095kFEA-jpE [https://perma.cc/2QUJ-Z94T] (observing that elected officials and other political leaders tend to consider future effects only on the scale of 2-5 years or "the next decade," due to cognitive biases, time preference, and election incentives).

92. *See* ORD, *supra* note 1, at 192.

93. *See* discussion, *infra* Part IV.

94. *See generally* DEREK PARFIT, REASONS AND PERSONS (2d ed. 1986) (introducing the hypothetical comparison of an event that ends 100% of human life and an event that ends 99% of human life, where the latter event permits the continuation of future generations, thus raising the question of how much we should value people who could one day exist).

95. *See generally* Bostrom, *Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards, supra* note 6; Hilary Greaves & William MacAskill, *The Case for Strong Longtermism* (Glob. Priorities Inst., Working Paper No. 5-2021, 2021); Christian Tarsney, *The Epistemic Challenge to Longtermism*, 201 SYNTHESE 195 (2023); William MacAskill, *The Case for Longtermism*, N.Y. TIMES (Aug. 5, 2022), https://www.nytimes.com/2022/08/05/opinion/the-case-for-longtermism.html [https://perma.cc/7C9Q-55XL].

radical degree that if one assigns any non-negligible value to the existence of future persons, and if one believes that future persons will tend to have lives worth living, one might view human extinction as a great harm.[96] For the participants in this study, these concerns relating to human extinction scenarios (for both the actual people who would die in such a catastrophe and the potential future people who would not have a chance to exist) and permanent dystopic scenarios (for many actual people current and future) combine to make a strong case that existential risk matters.

B. LITERATURE ON LAW AND SOCIAL CHANGE

As context for this article's discussion of existential advocacy, this section provides a brief background on how scholars have generally understood the role of lawyers in movements for social change.[97] Much of this literature has focused on the dark side of movement lawyering, criticizing lawyers who exaggerate the value of court-led social change, especially where lawyers fall under a "myth of rights" and a "hollow hope" that de jure victories in the courts will, on their own, bring about de facto social change.[98] The lawyers of the Civil Rights Movement have been cited as examples of this overly legalistic orientation, prioritizing litigation while discouraging grassroots organizing and legislative advocacy.[99] These scholars point to examples where judicial victories have sparked backlash and countermobilization, undermining the impact that court-centered activism seemed to promise.[100] Moreover, some scholars note that lawyers tend to dominate movement agendas, marginalizing the grassroots voices of the most affected constituencies.[101] Lawyers may tend to de-radicalize movements, both because of

---

96. Holden Karnofsky, *Debating Myself on Whether "Extra Lives Lived" Are As Good As Deaths Prevented*, COLD TAKES (Mar. 29, 2022), https://www.cold-takes.com/debating-myself-on-whether-extra-lives-lived-are-as-good-as-deaths-prevented [https://perma.cc/W255-YDWA].

97. *See generally* Sameer M. Ashar, *Movement Lawyers in the Fight for Immigrant Rights*, 64 UCLA L. REV. 1464 (2017); TOMIKO BROWN-NAGIN, COURAGE TO DISSENT: ATLANTA AND THE LONG HISTORY OF THE CIVIL RIGHTS MOVEMENT (2011); Susan D. Carle, *A Social Movement History of Title VII Disparate Impact Analysis*, 63 FLA. L. REV. 251 (2011); SCOTT L. CUMMINGS, BLUE AND GREEN: THE DRIVE FOR JUSTICE AT AMERICA'S PORT (2018); SCOTT L. CUMMINGS, AN EQUAL PLACE: LAWYERS IN THE STRUGGLE FOR LOS ANGELES (2021); MICHAEL W. MCCANN, RIGHTS AT WORK: PAY EQUITY REFORM AND THE POLITICS OF LEGAL MOBILIZATION (1994); MICHAEL J. KLARMAN, FROM JIM CROW TO CIVIL RIGHTS: THE SUPREME COURT AND THE STRUGGLE FOR RACIAL EQUALITY (2004); KENNETH W. MACK, REPRESENTING THE RACE: THE CREATION OF THE CIVIL RIGHTS LAWYER (2012); Douglas NeJaime, *Marriage Equality and the New Parenthood*, 129 HARV. L. REV. 1185 (2016); AUSTIN SARAT AND STUART A. SCHEINGOLD, CAUSE LAWYERS AND SOCIAL MOVEMENTS (2006); ANN SOUTHWORTH, LAWYERS OF THE RIGHT: PROFESSIONALIZING THE CONSERVATIVE COALITION (2008).

98. *See* GERALD N. ROSENBERG, THE HOLLOW HOPE: CAN COURTS BRING ABOUT SOCIAL CHANGE (2d ed. 2008); STUART A. SCHEINGOLD, THE POLITICS OF RIGHTS: LAWYERS, PUBLIC POLICY, AND POLITICAL CHANGE (2d ed. 2004).

99. *See* KLARMAN, *supra* note 97.

100. *See* ROSENBERG, *supra* note 98.

101. *See* Derrick A. Bell Jr., *Serving Two Masters: Integration Ideals and Client Interests in School Desegregation Litigation*, 85 YALE L.J. 470 (1976).

the lawyers' own preferences for working within institutional channels and because of the nature of the law as a conservative, precedential system for maintaining the status quo of the legal order.[102] These observations have led to calls for lawyers to take a reduced role in movement leadership.[103]

In contrast, recent empirical studies challenge the portrayal of social-change lawyers as narrowly legalistic and strategically unsophisticated.[104] Scholars increasingly see a revival of movement lawyering under the rubric of "integrated advocacy," in which lawyers coordinate their distinctively legal work (litigation and other legal services) with other tactics such as building movements, shaping public opinion, and advocating for new legislation.[105] By embedding lawyers within movements, integrated advocacy serves to enhance lawyers' accountability to the populations most affected by an issue.

Integrated advocacy also appears to enhance efficacy in at least some contexts. Professor Scott Cummings has described in empirical detail several examples of this "new canon" of social-change lawyering.[106] He notes that the marriage equality movement in particular has reshaped the test case litigation model with a greater emphasis on fostering favorable public opinion ("hearts and minds") through local legislative campaigns.[107] This creates a sense of collective demand for reform, which may help persuade judges to make what would have previously seemed very bold decisions (e.g., *Obergefell v. Hodges* in 2015), while also persuading the public to support enforcement of those decisions.[108] Cummings stresses that social change is a long-term project, marked by continual dynamics

---

102. *See* Catherine Albiston, *The Dark Side of Litigation as a Social Movement Strategy*, 96 IOWA L. REV. BULL. 61, 62 (2011) (observing that lawyers often "deradicalize and subtly reshape social movements"); Scott Cummings, *Law and Social Movements: An Interdisciplinary Analysis*, *in* HANDBOOK OF SOCIAL MOVEMENTS ACROSS DISCIPLINES 233, 263 (Conny Roggeband & Bert Klandermans eds., 2017) (noting that legal framings can "sanitize" issues to "comport with mainstream values," perhaps because, as CLS scholars have long argued, law favors the status quo and legal victories do not transform structural relations); Scott L. Cummings & Deborah L. Rhode, *Public Interest Litigation: Insights From Theory and Practice*, 36 FORD. URB. L.J. 603, 612 (2009) (noting that legal actions can dissipate grassroots activism, thereby reducing a "movement's transformative potential").

103. SCOTT L. CUMMINGS, LAWYERS AND MOVEMENTS (forthcoming Oct. 1, 2024) (describing the scholars of "movement liberalism" who recommend that social-change lawyers take a more limited, conventional client-centered advocacy role in support of grassroots organizations).

104. *See* Alan K. Chen, *Rights Lawyer Essentialism and the Next Generation of Rights Critics,* 111 MICH. L. REV. 903, 905–06 (2013) (describing "rights lawyer essentialism" as a common but inaccurate portrayal of civil rights attorneys as "elitist, singularly minded litigation hawks who care little for their clients or the subtleties of the dialectic political process."); *see also* ALAN K. CHEN & SCOTT L. CUMMINGS, PUBLIC INTEREST LAWYERING: A CONTEMPORARY PERSPECTIVE 518 (2013) (providing examples of how cause lawyers conceive of litigation "not in isolation, but as part of a comprehensive set of tools that are useful in advancing social reform").

105. *See* CUMMINGS, *supra* note 103 (observing that movement lawyering is experiencing a "revival" after "decades of dormancy").

106. *Id*.

107. *Id*.

108. *Id*. (arguing that lawyers can and should contribute to all aspects of this integrated model).

of resistance and struggle.[109] Law, when coordinated with action in other strategic domains, can be a powerful tool within a larger campaign for reform.[110]

This literature on social-change lawyering has focused primarily on lawyers within progressive grassroots movements—generally where a community turns to law in an effort to find legal voice and remedies.[111] Some scholars have extended this inquiry to other contexts, including movements for animals, the environment, and conservative causes, where the grassroots element is, sometimes (but not always), less salient.[112] The existential advocates are a step further in this direction away from traditional understandings of social movements. If their primary constituency is future generations, it is not possible for the existential advocates to develop a grassroots movement where the most affected community would organize to form a collective voice.[113]

Yet, the existential advocates resemble a social movement in some key respects. The effort to enfranchise a voiceless population of future generations is similar to the classic concern for marginalized communities in the civil rights tradition. Moreover, some participants in this study view their activism as an extension of their past contributions to progressive movements for social justice. These advocates frame mitigating existential risk as an effort to counteract discrimination against future persons and to advance equality.[114] As one participant explained: "Our assumptions are quite simple. We want to treat everyone as equal, not just in space, but also in time."[115] Participants in this study are engaged with the same strategic questions debated throughout the literature on movement lawyering—how to enhance de facto impact and how to best include the voices and represent the interests of key constituencies. With these commonalities in mind, this article will refer to the community working on existential risk as a nascent "movement." This allows for an inquiry into whether, and to what extent, insights from socio-legal literature apply to the novel context of existential risk mitigation. It also allows for a strategic analysis of whether the existential

109. *Id.*

110. *Id.*

111. *See* Tomiko Brown-Nagin, *Elites, Social Movements, and the Law: The Case of Affirmative Action*, 105 COLUM. L. REV. 1436, 1508 (2005) (noting that movements are generally made up of "socially marginal citizens" responding to oppression and inequality); Scott L. Cummings, *Law and Social Movements: Reimagining the Progressive Canon*, 2018 WIS. L. REV. 441, 451–60, 470–78, 487–94 (2018) (describing the "contemporary progressive legal canon" rooted in concern for marginalized groups, inequality, and a struggle over resources); David A. Snow, Sarah A. Soule & Hanspeter Kriesi, *Mapping the Terrain*, *in* THE BLACKWELL COMPANION TO SOCIAL MOVEMENTS 3 (1st ed. 2004).

112. *See*, *e.g.*, Camila Bustos, *Movement Lawyering in the Time of the Climate Crisis*, 33 PACE ENV'T L. REV. 1 (2022); HELENA SILVERSTEIN, UNLEASHING RIGHTS: LAW, MEANING, AND THE ANIMAL RIGHTS MOVEMENT (2009); SOUTHWORTH, *supra* note 97.

113. *See* Toby Ord, Remarks at the Stanford Existential Risks Initiative Virtual Conference (Apr. 17, 2021) (noting that, in the context of social movements focused on the interests of current people, the populations who "bear the most of the relevant costs" seek to vocalize their grievances and "campaign and push for change").

114. *See e.g.*, Interview with Anonymous Participant, *infra* note 115 and accompanying quote.

115. Interview with Anonymous Participant.

advocates should attempt to broaden their community into something more closely resembling a grassroots social movement.

II. RESEARCH DESIGN

The research for this Article consisted of a three-year study including ethnography, interviews, analysis of online materials, and first-hand experience in the field. The ethnography was conducted at the Legal Priorities Project ("LPP") over a period of five months in 2021 and 2022, during which fifty-nine virtual meetings were observed. Ethnography is an anthropological method of participant observation, where the researcher is invited to join a community while taking detailed analytic and descriptive notes on norms, interactions, and other cultural dynamics, as well as the researcher's own experiences of membership in the community.[116] Rather than testing hypotheses, ethnographers tend to inductively generate "grounded theory," wherein the study's key theoretical observations emerge from the thematic coding and analysis of fieldnotes.[117] This requires deep immersion in the culture under study. LPP was highly receptive to this methodology. They permitted me to observe weekly all-hands and small-group meetings, research workshops, online discussions, and internal documents.

Following this period of formal ethnography, I have continued to regularly engage with LPP and the larger legal community working on existential risk. This has included visits to key sites of the movement in the United States and Europe. I have attended more than twenty conferences and workshops in this field while regularly following online discussions (e.g., Slack workspaces, Facebook groups, blogs, Discord channels, podcasts, newsletters, and the Effective Altruism Forum). A range of published online materials are also referenced throughout this article, including white papers, research agendas, and curricula for courses and reading groups.

Over these three years of data collection, I conducted fifty-three semi-structured interviews. The interview participants included LPP team members, affiliates, and summer research fellows, as well as legal and political advocates at other organizations working on existential risk. Some interviews were held in person, but most were remote (via video-conferencing calls) with participants who were physically located in all continents of the globe except Antarctica.[118]

The interviews began by asking participants for a biographical account of how they became interested in the topic of existential risk. This was followed by questions about how they perceive existential risk, the community working on this issue, and the organizational cultures they have encountered, including

116. *See* KAREN O'REILLY, ETHNOGRAPHIC METHODS (2012) (suggesting that ethnographers combine "emic" understandings, from the perspective of the study's subjects, with the researcher's own "etic" understandings rooted in their research questions and interests).

117. *See* JULIANNE S. OKTAY, GROUNDED THEORY (2012).

118. This geographical observation assumes the convention of categorizing seven continents as North America, South America, Africa, Australia, Asia, Europe, and Antarctica.

considerations of epistemics, dissent, emotions, and identities (the cultural traits discussed in Part V). The interviews then transitioned to a discussion of legal strategy. Participants who are lawyers were asked about how their professional identities as lawyers comport with their identities as activists or members of a movement. In addition to these formal interviews, I engaged in hundreds of hours of informal conversations with members of this community, which are not quoted or paraphrased here but nevertheless inform the empirical analysis.

With qualitative methods, the researcher serves as the research instrument in the field. In contrast to, for example, collecting public data or distributing a survey, the qualitative researcher directly asks participants their questions and observes participants in their settings. The entire data collection process is filtered through the researcher's own perceptions and biases. This fact, along with a long line of critical literature about the power dynamics of representing research subjects, particularly considering the historical association between ethnography and colonialism, has led to a call for qualitative scholarship that is accompanied by a reflexive account of the researcher's interests, goals, and positionality. In this spirit, I briefly provide relevant self-reflections. More auto-ethnographic notes can be found throughout the presentation of findings (Parts III through V).

Over the course of the study, I transitioned from an outside observer to a more active participant in strategic discussions in this field. Following the conclusion of the ethnography, I have continued to attend LPP meetings on a regular basis, contributing to a range of the organization's strategic discussions. I generally agree with the notion that existential risk matters and that this community is taking useful steps toward an effective response, although I make recommendations for diversifying and scaling up this movement in Part VI. Becoming a "member" is not uncommon for an ethnographer, nor is it necessarily discouraged, although it can predispose the researcher to portray the community under study in a more favorable light.[119] This tendency is somewhat diminished by the existential advocates' strong cultural commitments to inviting dissent and criticism (discussed in Part V).

In the LPP meeting where I formally requested permission to conduct the ethnography, I expressed my intention to contribute to socio-legal theory and to provide feedback to the organization and the larger movement around existential risk. I explained that the project would draw comparisons to how law has been deployed in other movements for large-scale change, including movements for civil rights and animal protection that I have examined in other projects. LPP members embraced this project from the start and, to their credit, encouraged me to provide candid and critical feedback. I presented preliminary findings to the

119. *See* Hsun-Yu Sharon Chuang, *Complete-Member Ethnography: Epistemological Intimacy, Complete Membership, and Potentials in Critical Communication Research*, 14 INT'L J. QUALITATIVE METHODS 1, 1–2 (2015) (discussing the empirical value of becoming a "member" in ethnographic studies). *Cf.* Margaret D. LeCompte & Judith Preissle Goetz, *Problems of Reliability and Validity in Ethnographic Research*, 52 REV. EDUC. RSCH. 31, 46 (1982) (detailing the empirical and ethical risks when researchers develop strong social relationships with participants).

LPP team several times and invited participants to comment on an earlier draft of this article. This collaborative approach, in combination with assurances of confidentiality,[120] helps to promote trust and transparency in the interaction between researcher and participants. My relationship with this community has been consistently warm and respectful. This sense of rapport likely colors some of my interpretations, although my ultimate sense of responsibility as an empirical researcher is to provide an accurate description of this community, including its shortcomings and points of tension.

III. THEORY OF EFFICACY: THE PRIORITIES METHODOLOGY

This Part begins the presentation of empirical findings with a summary of how these advocates approach the question of efficacy. As reviewed in Part I, all movements for large-scale change face the question of how to achieve lasting de facto impact. The existential advocates take a distinct approach to this question, which I label the "priorities methodology," in keeping with the terminology of the field (e.g., the naming of the "Legal Priorities Project"). Their methodology has two steps. Section A explores how these advocates select goals by undertaking a global search for the cause areas where they predict that they can do the "most good" in the world. This search is continual and sometimes leads these advocates to entirely shift their focus to prioritize different cause areas. Section B explores how these advocates, upon selecting a particular objective, then make decisions about strategies and tactics. In short, they use a "reverse engineering" framework to keep a focus on maximizing their overall, counterfactual impact toward the chosen end goal. This approach requires keeping an open mind about what strategic tools would be most useful, even if these tools sometimes have little to do with a lawyer's legal expertise. This zealous pursuit of impacting the most people (and other sentient beings) to the greatest extent is unusual within the history of social-change lawyering. But it also raises points of tension when applied in practice.

A. SELECTING GOALS

How do social-change lawyers decide to focus on a particular issue or cause? The answer from socio-legal scholarship is relatively intuitive. Lawyers, like other activists, are attracted to causes because of some combination of identity, values, ideology, and the availability of resources and opportunities.[121] Often, grassroots communities express a demand for remedy or reform, which the lawyers then work to translate into legal action.[122] Lawyers also pursue their own

120. Identifying information is used in this article only where specific permission was granted by the identified participant.

121. *See generally* STUART A. SCHEINGOLD & AUSTIN SARAT, SOMETHING TO BELIEVE IN: POLITICS, PROFESSIONALISM, AND CAUSE LAWYERING (2004).

122. *See generally* CHEN & CUMMINGS, *supra* note 104.

political values and ideologies, sometimes taking little direction from clients and constituencies.[123] These traditional approaches can produce powerful results, as reflected in a long line of successful campaigns for civil rights and other causes.[124] Yet, they can also lead to well-intentioned but ineffective efforts, including where lawyers fail to consider alternative cause areas where they could more effectively advance justice, equality, democracy, well-being, and other conceptions of "the good." This concern has led some scholars to call on public-interest lawyers to develop more rigorous and formalized processes for prioritizing cause areas and objectives.[125]

The priorities methodology developed by participants in this study provides one answer to this call. Rather than beginning with a favored cause, these advocates take the unusual starting place of "cause neutrality," meaning that they begin with an openness to working on any conceivable issue—anything in the world.[126] This is a novel starting place for a group of public-interest lawyers. They then undergo a systematic process for selecting the cause area where they predict that they can have the greatest moral impact. As discussed below, this methodology is drawn from "Effective Altruism," a theoretical framework most often applied in philanthropy that seeks to maximize "doing good" by combining well-meaning intentions with an evidence-based approach to effectiveness.[127]

The priorities methodology, as applied by participants in this study, begins with deliberations over first principles of morality. Although participants often spoke of a utilitarian effort to maximize well-being, most participants described a great degree of normative uncertainty and a reluctance to fully embrace utilitarianism—with some describing themselves as "2/3 utilitarian."[128]

---

123. *See generally* SCHEINGOLD & SARAT, *supra* note 121.

124. *Id.*

125. *See* Cummings & Rhode, *supra* note 102, at 637, 639 (observing that public-interest law generally lacks formal processes for "identifying objectives and establishing priorities among them" and that "few organizations operate with explicitly articulated theories of change or specific measures of performance").

126. For a discussion of cause neutrality, see WILLIAM MACASKILL, DOING GOOD BETTER: EFFECTIVE ALTRUISM AND A RADICAL NEW WAY TO MAKE A DIFFERENCE, 202 (2015).

127. *See id.* at 14; Benjamin Todd, *Can One Person Make a Difference? What the Evidence Says*, 80,000 HOURS (Apr. 2016), https://80000hours.org/career-guide/can-one-person-make-a-difference [https://perma.cc/HD86-Y2M5] ("People often wonder how they can 'make a difference,' … [but] the key question is: What are some of the best ways to make a difference?") (emphasis omitted). For context about Effective Altruism, the community around these ideas has grown to include a number of non-profit organizations focused on research (e.g., the Global Priorities Institute, Rethink Priorities), career advising (e.g., 80,000 Hours, Training for Good), assessing effective philanthropy (e.g. GiveWell), grantmaking (e.g., Open Philanthropy, Longview Philanthropy), and education and outreach (e.g., the Centre for Effective Altruism, which hosts several conferences every year, runs the Effective Altruism Forum, and supports other Effective Altruist organizations).

128. *See, e.g.*, WINTER ET AL., *supra* note 25, at 1, 14–16, 85, 89 (citing grounds for protecting future generations primarily in utilitarianism but also in deontology, virtue ethics, and other traditions of moral and political philosophy); *Tyler Cowen on Stubborn Attachments, Prosperity, and the Good Society*, ECONLIB, at 43:25 (Aug. 7, 2017), https://www.econtalk.org/tyler-cowen-on-stubborn-attachments-prosperity-and-the-good-society/#audio-highlights [https://perma.cc/U6UD-K829] (attributing the notion of being 2/3 utilitarian to Cowan although he notes that he uses this term in a somewhat "tongue in cheek" manner).

Their next step is to determine which cause areas are amenable to the greatest impact toward chosen moral goals. The core criteria of this methodology are importance, neglect, and tractability (the "INT" analysis).[129] In this framework, a cause is prioritized not only because it affects many people and to a great degree (importance), but also because it is feasible to reduce this risk without imposing morally-offsetting costs (tractability), and because the issue is not already receiving adequate attention such that any interventions would be subject to diminishing returns (neglect).[130]

For example, in the early period of this study, LPP was focused on risks from engineered pandemics because of the potential to produce widespread illness and death affecting current and future generations (importance), the evidence that litigation could be useful in addressing issues such as unsafe laboratory practices (tractability), and the lack of governmental attention to the issue (neglect).[131] In this early period, the LPP team tended to view existential threats from AI as highly important and neglected but less tractable.[132] More recently, LPP has shifted their attention to AI as new opportunities for legal and political action have arisen, thus enhancing tractability.[133] These shifts in priorities emerge from lengthy collaborative discussions in LPP's "theory of change retreats," including a three-day event that I attended during the period of this study.[134]

Existential risk has recently emerged as a core focus of the Effective Altruism community,[135] prioritized due to the importance of the issue for a great number of current and future lives, in addition to evidence of severe neglect and at least some degree of tractability.[136]

The concern for future generations in this analysis follows from a commitment to "moral circle expansion" and an effort to treat all sentient beings equally—whether or not these beings are especially proximate or similar to ourselves. In the early years of Effective Altruism (the 2000s and early 2010s), the community

---

129. *See* MACASKILL, *supra* note 126 (outlining the "INT" analysis and proposing five key questions to assess these criteria: "How many people benefit, and by how much? Is this the most effective thing you can do? Is this area neglected? What would have happened otherwise? What are the chances of success, and how good would success be?"); DirectedEvolution, *Review of ITN Critiques*, EFFECTIVE ALTRUISM F. (Oct. 9, 2019), https://forum.effectivealtruism.org/posts/MtCAsPMftvJqRBYzr/wip-summary-review-of-itn-critiques [https://perma.cc/TE64-SA9S].

130. *See* MACASKILL, *supra* note 126 (explaining the neglect criterion by reference to the diminishing returns of investing in causes that are already receiving a great deal of attention, and noting the potential for counterfactual impact when otherwise few resources would be spent on an issue).

131. Legal Priorities Project, Internal Documents (on file with author).

132. *Id.*

133. Legal Priorities Project, Internal Documents (on file with author)

134. This is the term that the Legal Priorities Project has chosen to call their collaborative discussion retreats.

135. Note that the community of existential advocates examined in this article can be roughly viewed as a subset of the larger Effective Altruism community.

136. *See* Benjamin Todd, *The Case for Reducing Existential Risk*, 80,000 HOURS (Oct. 2017), https://80000hours.org/articles/existential-risks [https://perma.cc/7HVB-MRYU].

was focused on expanding the moral circle across spatial boundaries, leading to a focus on effective global poverty and health interventions.[137] Members of this community cite empirical evidence suggesting that it currently costs roughly two to five thousand dollars, on average, to "save a life."[138] Saving lives is measured in various ways, including the consideration of life years or examples like preventing a child from dying of a preventable illness.[139]

Just as this community seeks to expand the moral circle across geographic boundaries, they also seek to expand across the boundaries between species in recognition that non-human animals are suffering on an immense scale.[140] And in just the past few years, this community has taken moral circle expansion across temporal boundaries with a focus on future generations.[141] Although it is difficult to know how our actions affect the long-term future (in other words, it is difficult to know the total or final impact of any actions we might take), existential risk is arguably less susceptible to this concern in one key respect—existential threats persist into the future because they, by definition, lock in a condition (human extinction or permanent dystopia).

Although existential risk is prioritized by many Effective Altruists at the moment, the priorities methodology can, by design, yield different answers. One of the foundations of this framework is a willingness to update and change directions upon new information.[142] As one participant explained, "If you were to prove that one of the core [Effective Altruist] cause areas actually wasn't an issue, like that AI safety was not a big risk, you would be celebrated."[143]

B. SELECTING STRATEGIES

Having selected a cause area where advocates predict that they can have the greatest impact (such as existential risk or particular sources of existential risk), their next step is to select strategies in an effort to optimally advance the prioritized cause. This step involves a process of reverse engineering from end goals, as visually represented in Figure 2.

---

137. *See* Dylan Matthews, *How Effective Altruism Went from a Niche Movement to a Billion-dollar Force*, VOX (Aug. 8, 2022), https://www.vox.com/future-perfect/2022/8/8/23150496/effective-altruism-sam-bankman-fried-dustin-moskovitz-billionaire-philanthropy-crytocurrency [https://perma.cc/7E8K-CAKL] (tracing the history of Effective Altruism).

138. *See Why is it so Expensive to Save Lives?*, GIVEWELL (Dec. 2021), https://www.givewell.org/cost-to-save-a-life [https://perma.cc/UF6D-JRZP] (offering the example of health interventions in Guinea, which can be estimated to save one life per $4,500 donated).

139. *See* Derek Thompson, *The Greatest Good,* THE ATLANTIC (June 15, 2015), https://www.theatlantic.com/business/archive/2015/06/what-is-the-greatest-good/395768/[https://perma.cc/HZW6-3US7].

140. *See* Matthews, *supra* note 137.

141. *Id.*

142. *See* WINTER ET AL., *supra* note 25 ("[W]e offer a rigorous yet flexible, and potentially ever-evolving methodological framework for deciding which problems to work on and how to tackle them.").

143. *See* MACASKILL, *supra* note 126 ("[W]e genuinely just want to do what's best for the world, so if we're wrong about anything—even if it's the thing we've been dedicating our lives to—we should want to know.").

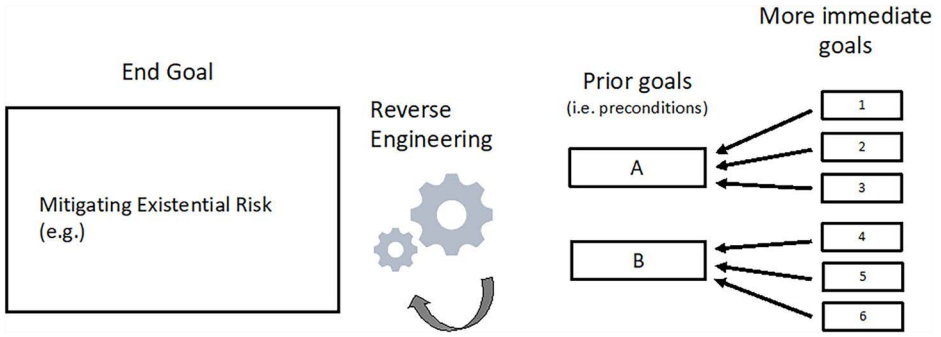

FIGURE 2. The Priorities Methodology for Selecting Strategies

As observed in this study, strategies are selected through a process that I label "end goal primacy." This involves working backward from their cause (the "end goal") to more specific and immediate goals that are preconditions to achieving the end goal.[144] This approach is derived from "Theory of Change," a framework drawn from philanthropy and impact investing, which recommends visually mapping out the reverse engineering process.[145]

In an internal document observed in this study, LPP created a Theory of Change map with an end goal to advance a state of the world in which "humanity's long-term potential is safeguarded."[146] The report noted that this goal is "far too vague to guide our decision-making" but that it helps the organization "link every subsequent step" to their end goal.[147] For example, one causal chain through this document begins with the goal that research on existential risk is known and valued by policy-makers.[148] This is preceded by building relationships with policy makers and encouraging LPP affiliates to pursue policy careers, which is preceded by building relationships with organizations on the frontlines

144. LPP's website describes their approach as "mission-oriented," whereby they "work backwards from these long-term goals to prioritize the projects that [they] think are most promising and impactful." *Open Positions*, LEGAL PRIORITIES PROJECT, https://www.legalpriorities.org/open-positions.html [https://perma.cc/Y7SC-27BX] (last visited Sept. 18, 2023).

145. *See* Edward T. Jackson, *Interrogating the Theory of Change: Evaluating Impact Investing Where it Matters Most*, 3 J. SUSTAINABLE FIN. & INV. 95, 100 (2012) (recommending visual mapping to identify "underlying logic, assumptions, influences, causal linkages, and expected outcomes of a development program or project").

146. Legal Priorities Project, Internal Report (on file with author).

147. *See id.* Other organizations working on the long-term future of humanity have also applied elements of Theory of Change. For example, the Simon Institute for Longterm Governance published a diagram that lists an end goal of "long-term human flourishing" with preconditions to improve "long-term institutional fit" through improving decision-making process and integrating longtermist concerns in "dominant societal narratives," "institutions," and "policy agendas." *See* Felix Haas, *Our Theory of Change*, SIMON INST. FOR LONGTERM GOVERNANCE, https://www.simoninstitute.ch/blog/post/our-theory-of-change/ [https://perma.cc/GK49-MMSX] (last visited Sept. 18, 2023).

148. *See* Legal Priorities Project, *supra* note 146.

of policy outreach and advocacy.[149] The key point here is that more specific and immediate goals are defined after higher goals. If new information alters higher goals, this model requires a willingness to reconsider an organization's immediate and day-to-day strategic priorities.

When designing strategies, participants drew insights from the "integrated advocacy" literature on law and social change.[150] This literature recommends blurring the distinction between law and polic"y such that advocates coordinate their strategies and frames across the traditional law/policy divide.[151] For the participants in this study, this integration is consistent with the Effective Altruist notion of a "portfolio approach" and an "alliance mentality," emphasizing the overall impact of collective efforts toward a particular end goal, rather than the isolated impact of the actions taken by any particular individual or organization.[152] An internal LPP report on impact litigation directly references the integrated advocacy literature and emphasizes that litigation efforts should be paired with complementary actions in the domains of legislation, policy, public opinion, academic research, and education.[153]

The scholars of integrated advocacy view the blurring of law and policy as a corrective to the long-standing concern that lawyers in social movements tend to over-emphasize the value of law and courts.[154] In the most unsympathetic portrayals, social-change lawyers seem to fall under a "myth of rights" while discouraging legislative advocacy, grassroots organizing, and other forms of activism that might prove more effective.[155] But the myth of rights does not seem to hold much appeal for the existential advocates. Some participants in this study noted that, until very recently, lawyers in Effective Altruism were more drawn to the opposite myth: that, for Effective Altruist causes, as one participant put it, "litigation is useless, law is useless."[156] Law students interested in Effective Altruism had been advised to take high-paying positions so that they could "earn

149. *Id.*

150. *See* CUMMINGS, *supra* note 103 (observing that the "new convention" among public interest lawyers stresses the importance of "multidimensional" and "integrated" advocacy where lawyers are "strategically sophisticated" and work with a wide range of allies to "advance political goals in multiple venues through coordinated tactics in the face of persistent opposition").

151. *See id.* (reviewing the interdisciplinary literature on the "interaction between law and politics" and recommending that law be "coordinated with politics through an integrated strategy that maximizes the potential for sustainable social change").

152. *See* MACASKILL, *supra* note 126 ("The fact that we each act as part of a wider community warrants a 'portfolio approach' to doing good—taking the perspective of how the community as a whole can maximize its impact.").

153. *See* Legal Priorities Project, Internal Impact Litigation Report (on file with author).

154. *See supra* Part I.B.

155. *See* SCHEINGOLD & SARAT, *supra* note 121 (noting that social-change lawyers tend to be "attracted to courts as fly paper," mesmerized by a "myth of rights").

156. Interview with anonymous participant.

to give" [157] to effective charities rather than seeking out impactful legal work relating to prioritized cause areas.

This trend has changed over the past few years. The Effective Altruism community now appears to see much more value in legal and political efforts.[158] Although the legal efforts in this space are receiving a great deal of funding and support, some hesitations about the role of the law remain.[159] These persistent hesitations about law may help this community avoid falling into a myth of rights. For example, these advocates are considering efforts to create justiciable rights for future generations in domestic constitutions, common law doctrines, international law (e.g., extensions of human rights across time), and intergovernmental agreements. But participants were consistently mindful that simply establishing such rights would not necessarily mean they would be enforced in transformative or otherwise meaningful ways.

A central commitment of the priorities methodology is to undertake a counterfactual analysis, assessing the marginal expected value of any action under consideration.[160] This involves weighing how one option compares to alternatives, including the possibility of doing nothing. This approach draws from the Effective Altruist norm of asking "what would have happened otherwise?"[161]

In my observations at LPP, this counterfactual approach was consistently evident in debates over planned legal interventions. As described in an LPP report, any strategic action should be compared to "feasible alternatives and counterfactuals," including consideration of "the chances the legislature or executive will take up this issue at some point anyway."[162] For example, when considering whether to provide formal comments on the U.N. Declaration on Future Generations, LPP engaged in a lengthy decision-making process about whether their input would be counterfactually impactful when considering what would happen otherwise and how the same "people hours" could be spent on other projects. This emphasis on counterfactuals is represented by the term, "no action," in Figure 3.

157. Interview with anonymous participant.

158. *See e.g.*, Legal Priorities Project, *supra* note 144.

159. *See supra* notes 121–25 and accompanying discussion.

160. *See* MACASKILL, *supra* note 126, at 57 (describing "marginal utility" as a "fundamental piece of scientific reasoning" that motivates the methodology of Effective Altruism).

161. *Id.* at 204.

162. Legal Priorities Project, Internal Litigation Report (on file with author).

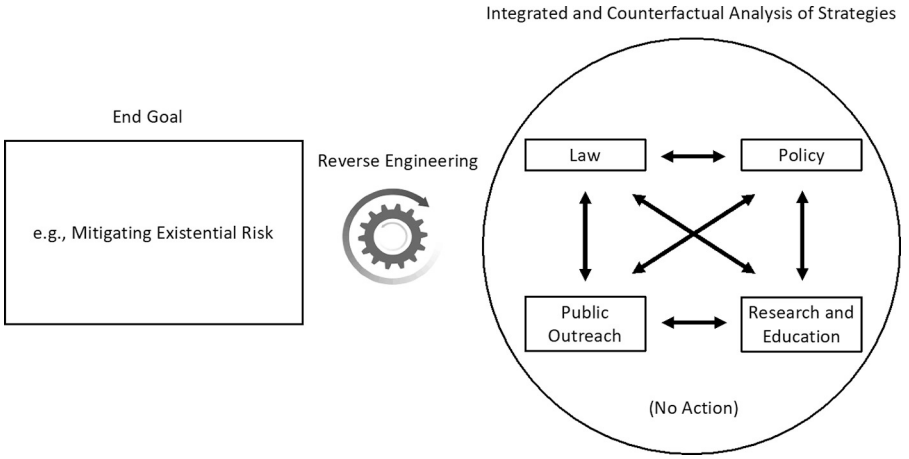

FIGURE 3. Integrated Advocacy

These advocates place a particular value on strategies that challenge the cognitive biases that stand in the way of mitigating existential risk. This is evident in discussions of specific strategies as well as high-level discussions about whether to focus more attention on law (e.g., litigation) or policy (e.g., legislative and regulatory advocacy). A participant working in the policy space emphasized that their work is often stymied by cognitive biases that inhibit recognition of existential risk, noting that "99.5% of policymakers don't actually realistically think in the next 100 years human beings could possibly go extinct . . . ."[163] Another participant working in this space emphasized that elected officials are focused on the near-term demands of their constituents and have "no time" because they have people "hounding them . . . every second of the day" and an incessant "huge box of materials to go through."[164] As another participant put it, "Most people in [the] political realm are working on 24-hour review cycle . . . Long-term [for those in the political realm] is a year out from now . . . But with respect to 100 years that would be a really hard case to make."[165]

Participants working on the legal dimension of these issues cited further law-specific biases, including the notion that law tends to have a conservative, backward-looking precedential orientation that might be unreceptive to the "new legal techniques" that would help to address existential risks to future generations.[166] Moreover, participants worried that judges might struggle with the probabilistic nature of existential risk given their lack of "formal training in quantitative subjects."[167]

163. Interview with anonymous participant.
164. Interview with anonymous participant.
165. Interview with anonymous participant.
166. *See* Albiston, *supra* note 102.
167. *See* WINTER ET AL., *supra* note 25 (noting that law tends to rely on non-quantitative reasoning, as evident in standards framed as "beyond a reasonable doubt," "probable cause," and "balancing tests").

But participants also saw promise in the judiciary, which might be inclined to view future generations as a disenfranchised group. As one participant noted, courts uphold the "liberal values" of democracies and the protection of the "political minority."[168] Several participants expressed optimism that judges, owing to their general reliance on "abstract values," might be receptive to expanding human rights "independently of time."[169] It is also important to note that these advocates are taking some actions to reduce cognitive biases, including holding educational workshops for policymakers on existential threats and the distinct decision-making challenges around low-probability/high-impact risks.

This section has described the ex-ante phase of the priorities methodology: predicting how planned actions might optimally advance an end goal. But the effort to maximize impact also demands continual assessment of actions that have been taken. This involves a commitment to empiricism, Bayesian updating, and developing metrics and experimental methods when possible.[170]

These advocates draw on different conceptions of impact found in the scholarship on law and social change.[171] As applied by participants in this study, this includes questions of whether to pursue broad and holistic approaches to existential risks (e.g., establishing rights and legal standing for future generations, criminalizing activities that generate existential threats, and influencing regulatory policy that affects multiple branches of government) or narrower and more cause-specific legal challenges (e.g., enforcing regulations on a laboratory developing a dangerous technology in an unsafe manner). They also follow the socio-legal distinction between direct effects (e.g., how a court order is enforced) and indirect effects (e.g., how a court order reshapes political discourse). As described in LPP internal documents, some legal actions may be especially valuable for their indirect effects.[172] For example, where a litigation victory or loss serves an educational function (e.g., to reveal the urgency of a problem), a motivational function (what socio-legal scholars call "internal effects" where a legal victory helps stimulate growth in the movement), a symbolic function (e.g., legitimating ideas in the movement, and gaining control over issue definition), and a resource mobilization function (e.g., attracting funding and increasing bargaining power).[173] These indirect effects may be particularly important in the context of non-binding international soft law, which participants consider a key avenue for

168. Interview with anonymous participant.

169. Interviews with anonymous participants.

170. An internal LPP report makes the case that "specific measures of performance" are necessary to assess "progress. . .made towards our goals in an observable, measurable way" and to "identify where theory and practice fall apart and thus where we should rework our [Theory of Change]." Experimental projects in this field include a forthcoming study of how judges may interpret legal arguments relating to existential risk cause areas.

171. *See generally* Albiston, *supra* note 102.

172. Legal Priorities Project, Internal Litigation Documents (on file with author).

173. *Id.*

addressing the global dimension of existential risks.[174] As one participant noted, even if soft law does not have the teeth of strong enforcement mechanisms, it can help "inform the common language that people use . . . how they perceive the future . . . and whether or not they consider extreme risks."[175]

IV. THEORY OF ACCOUNTABILITY: REPRESENTING FUTURE AND CURRENT GENERATIONS

The previous Part described the priorities methodology as a formula for enhancing social-change efficacy. This section describes how these advocates think about the ethical dimensions of their advocacy, particularly their practices of inclusiveness and accountability.

These ethical dimensions have been a salient concern in the literature on social-change lawyering. Even the most celebrated civil rights advocates have been charged, in some scholarly accounts, with failing to hear or heed the voices of the populations most affected by an issue.[176] Recent scholarship acknowledges this "elite critique," but tends to offer an increasingly collaborative portrait of activist lawyers working alongside movements and seeking to secure the remedies desired by affected constituencies.[177]

These debates about the accountability of lawyers find new expression among the existential advocates who conceive of their primary constituency, essentially their clients, as the multitudes of people who could exist in the future.[178] Some participants described themselves as part of the tradition of civil rights, protecting future generations in an effort to, as one participant put it, "give voice to people who have been underrepresented."[179] But how can these lawyers be accountable to a population that does not yet, and might not ever, exist? Moreover, accountability in this context is complicated by the question of how to include the various perspectives of currently living people who are affected by existential risk and the costs of interventions. This raises the central puzzle explored in this Part: How do these advocates conceive of their accountability to current people (who are capable of speaking) and future people (who are potentially the greater affected population but incapable of speaking)?

---

174. Participants frequently acknowledged the limitations of international soft law. For example, one participated cited the failures of international health regulations, which they described as "the most advanced form of regulation in the international system," but noted that these regulations were "not respected during Covid." Interview with anonymous participant.

175. Interview with anonymous participant.

176. *See* Bell, *supra* note 101 (noting that the lawyers of the Civil Rights Movement chose to prioritize racial integration in schools while Black Southerners expressed a clear preference to focus on educational quality and other approaches to addressing racial subordination).

177. See CUMMINGS, *supra* note 103.

178. It is worth noting that many participants in this study would frame the issue as a concern not only for future humans but also for future non-human animals and possible sentient AI and other beings.

179. *See* ORD, *supra* note 1 (calling for efforts to bring "the representation of future generations into national and international democratic institutions").

Accountability to clients and constituencies is a well-developed topic in the literature on the professional responsibility of lawyers.[180] First-order representation issues, which relate to duties to clients, raise an inherent tension. Lawyers may be justified in some degree of paternalism, owing to their expertise, but clients also have justified interests in their autonomy and control over objectives.[181] In the context of social-change lawyering, these tensions can be heightened where attorneys prioritize the broader impact of a case over securing remedies for the particular client.[182] Participants in this study anticipated first-order issues, noting that their litigation plans may tend to focus on relatively remote (future) impacts of a case, which may be orthogonal or opposed to the interests of a represented client.[183]

Participants were generally more troubled by representational issues of the second order, regarding duties to constituencies, causes, and a sense of "public accountability." These second-order duties are largely absent from the rules of professional conduct, except in aspirational prefatory language where the U.S. *Model Rules of Professional Conduct* describes the lawyer as a "public citizen having special responsibility for the quality of justice," who should "seek improvement of the law" and engage in "civic influence."[184] This concern for the public good is especially central to the practice and theory of "cause lawyering," which refers to lawyers who work on behalf of causes and social movements.[185] The image of the lawyer as a hired-gun for clients is replaced in this tradition by the lawyer's own "political or moral commitment."[186] In a wide range of movement lawyering contexts, this approach has raised difficult questions about how, and to what extent, affected constituencies should be included in strategic decision-making processes.[187]

---

180. *See e.g.*, Maute, *infra* note 181.

181. *See* Judith Maute, *Allocation of Decisionmaking Authority Under the Model Rules of Professional Conduct*, 17 U.C. DAVIS L. REV. 1049, 1081 (1984) (discussing the tension between paternalistic and client-service visions of the legal profession); Stephen L. Pepper, *The Lawyer's Amoral Ethical Role: A Defense, A Problem, and Some Possibilities*, AM. BAR FOUND. RES. J. 613 (1986) (defending the "standard conception" of the lawyer role, which demands zealous partisanship on behalf of the client's interests while holding in abeyance the lawyer's own moral assessments of client objectives);

Marcy Strauss, *Toward a Revised Model of Attorney-Client Relationship: The Argument for Autonomy*, 65 N.C. L. REV. 315 (1987); Mark Spiegel, *Lawyering and Client Decisionmaking: Informed Consent and the Legal Profession*, 128 U. PA.

L. REV. 41 (1979).

182. *See* Carle *supra* note 97; Susan D. Carle & Scott L. Cummings, *A Reflection on the Ethics of Movement Lawyering*, 31 GEO. J. LEGAL ETHICS 447 (discussing conflicts that arise between the client's interests in a speedy and beneficial remedy and the lawyer's interests in advancing a larger cause or reform).

183. An internal LPP report notes a desire to "avoid the type of victimization" that has occurred in some civil rights litigation where plaintiffs' interests have been overridden or "misrepresented." Legal Priorities Project, Internal Report (on file with author).

184. MODEL RULES OF PROF'L CONDUCT pmbl. (2023) [hereinafter MODEL RULES].

185. SCHEINGOLD & SARAT, *supra* note 121, at 4.

186. *Id.*

187. *See* CUMMINGS, *supra* note 103.

The participants in this study seem to fit squarely within the cause-lawyering tradition, drawing from their moral imperative to "do the most good." LPP members regularly deliberated over second-order representational issues.

In one meeting, I suggested potential labels for LPP's theory of accountability, such as "first-principle accountability." This framing stresses moral commitment, although participants noted that the term lacks recognition of their emphasis on methods for maximizing impact. One participant suggested "internal accountability," as opposed to "external accountability" where "you have direct feedback from living constituencies."[188] This participant explained that aiming to protect future generations, who cannot provide input, means that the existential advocates must find accountability through the integrity of their own moral reasoning and their commitment to maximizing impact.[189]

Regarding their focus on future generations, I proposed the term, "astronomical value accountability," in reference to the arguably immense value associated with future persons—with a future potential of perhaps $10^{a}16$ "human lives of normal duration," in addition to the likelihood that technological developments could greatly increase this number if we do not destroy ourselves along the way.[190] Participants were unpersuaded by this label, in part because it seems to imply "strong longtermism," the view that our greatest priority today should be impacting the far future.[191] Most participants expressed uncertainty and skepticism about this brand of longtermism. A less "strong" version of "future generation accountability" might be a more accurate term, reflecting the general sense of compassion and concern that participants expressed for people who will one day exist and have meaningful lives if we can avoid an existential catastrophe. But this term fails to capture the concern that participants consistently expressed for current living people who are affected by existential risk—and the overlapping issue of sub-existential catastrophic risk.

Within this discussion, some participants were skeptical of any suggestion that they "represent" future generations, because this would imply a fiduciary relationship that seems impossible given the inability to receive communications from the parties being represented.[192] Participants referenced an "epistemic challenge" when seeking to know what future generations might want and need. All social-change lawyers face some degree of epistemic challenge when seeking to

188. Anonymous participant, Remarks at a Legal Priorities Project Internal Meeting.

189. *Id.*

190. *See* Bostrom & Ćirković, *supra* note 66, at 8 (discussing scientific predictions that complex life on earth may last for .9 to 1.5 billion years, although it is possible that space travel could enable us to survive until the last stars burn out, around 100 trillion years for now, or even until black holes disappear, which might be $10^{a}100$ years from now, or even longer).

191. *See* Greaves & MacAskill, *supra* note 95 (defining strong longtermism as "the view that impact on the far future is the most important feature of our actions today") (emphasis omitted).

192. *See* MACASKILL, *supra* note 7 ("Though we cannot give genuine political power to future people, we can at least give consideration to them. By abandoning the tyranny of the present over the future, we can act as trustees . . .").

understand what an affected constituency might desire, especially when representing loosely affiliated groups with complex power dynamics.[193] This epistemic problem is arguably even more difficult when working to protect the long-term future. History reveals vast changes in human values over centuries and millennia, which suggests that any values we lock in today, no matter how beneficial they seem to us, might be deemed undesirable by future generations.[194] Participants responded to this challenge by suggesting that future generations would want us to protect their "basic needs," as defined in the lower third of Maslow's hierarchy, such as the necessities of survival and a baseline of well-being.[195] Another common response among these advocates was to emphasize the principle of optionality, which recommends avoiding locked-in and irreversible effects that reduce the autonomy of future generations. The mitigation of existential risk would seem to fit both of these responses—by avoiding extinction and permanent dystopia, these advocates hope to secure basic needs and preserve optionality.

These advocates also deliberated extensively over their accountability to people alive today. Some de-emphasized this notion of current-person accountability. As one participant explained, to follow the prevailing norms and values of "present humans" may be "almost irresponsible" because humans are generally subject to cognitive biases against recognizing low-probability risks and the moral interests of such a distant population as future generations.[196] Another participant argued that conceiving of accountability primarily to future rather than current persons is more inclusive than prevailing theories of democracy, which can be critiqued for "only taking into consideration the interests and preferences of people currently alive."[197]

But most participants, including those who greatly value the interests of future generations, also emphasized the importance of being accountable to current persons. These advocates are working to mitigate threats that could cause a great deal of suffering and loss of life to people alive today.[198] Participants regularly discussed how these harms, if realized, would likely be distributed unequally, along the familiar axes of global inequality.[199] Moreover, these advocates have

193. ANN SOUTHWORTH, BIG MONEY UNLEASHED: THE CAMPAIGN TO UNLEASH BIG MONEY IN AMERICAN POLITICS (forthcoming Dec. 2023) (noting that the rules of professional conduct are silent on how movement lawyers should represent "groups with poorly defined decision-making processes, or loose coalitions in which power among the different groups within the coalition is unequal . . . [where] it can be difficult to decide who speaks for the group or coalition, and it can be hard to reconcile the competing claims"); *see* Tarsney, *supra* note 95.

194. *See* MACASKILL, *supra* note 7.

195. *See* Saul McLeod, *Maslow's Hierarchy of Needs*, 1 SIMPLY PSYCH. 1 (2007).

196. Interview with anonymous participant.

197. Interview with anonymous participant.

198. *See* discussion, *supra* Part I.

199. An internal LPP report emphasized the inequalities often associated with catastrophic events, as some harms "may be readily avoided and mitigated locally by those with resources." Legal Priorities Project, Internal Report (on file with author).

proposed new governing structures that would bring a broader community of current persons into the effort to protect future generations.[200]

Some participants worried that if they were only accountable to future generations, the distant and abstract nature of this population might tend to undermine the notion of accountability altogether, leading to a self-interested and biased application of the priorities methodology. In response to this concern, participants sometimes referenced current-person accountability as a reference point to help internalize their sense of future-person accountability. For example, they noted that $5,000 spent on an event focused on protecting future generations could have instead been spent to save a child's life who would otherwise die from a preventable illness. This creates an urgent sense that efforts aimed at the future should be subjected to a great deal of scrutiny because lives are at stake.

These discussions of current-person accountability often centered on the question of who seems to have a seat at the table and who seems to be excluded in the community of existential advocates. This was usually framed as an issue of racial, geographic, and gender representation. Although this movement started with an Oxford-based conversation among mostly white men, the community has (by some measures) diversified over time.[201]

Relative to many other organizations in this field, LPP appeared to be quite diverse in terms of international representation. In addition to their wide-ranging geographic backgrounds, LPP adopts a cultural value to "think globally,"[202] which is reflected in their strategic discussions (e.g., planning legal interventions according to their cross-jurisdictional impacts), as well as informal conversations (e.g., remarking on cultural differences as revealed through anecdotes from day-to-day life or different reactions to global news events).

The international nature of this movement was a near constant theme in my observations during the ethnography and interviews, whether I was traveling in

---

200. John & MacAskill, *supra* note 88 (proposing citizens' panels as a "novel representative, deliberative, and future-oriented body" with "an explicit mandate to represent the interests of future generations").

201. *See also* Vaidehi Agarwalla, *2019 Ethnic Diversity Community Survey*, EFFECTIVE ALTRUISM F. (May 11, 2020), https://forum.effectivealtruism.org/posts/2T3cGecjHfbEPXeEc/2019-ethnic-diversity-community-survey [https://perma.cc/YU3D-MBZD]; *Anonymous Contributors Answer: How Should the Effective Altruism Community Think About Diversity?*, 80,000 HOURS (Apr. 27, 2020), https://80000hours.org/2020/04/anonymous-answers-diversity [https://perma.cc/N9VW-JWKR] (discussing measures of diversity in the Effective Altruism community). *Compare* Neil Dullaghan, *EA Survey 2019 Series: Community Demographics & Characteristics*, EFFECTIVE ALTRUISM F. (Dec. 5, 2021), https://forum.effectivealtruism.org/posts/wtQ3XCL35uxjXpwjE/ea-survey-2019-series-community-demographics-and [https://perma.cc/M86R-ZSYW] (reporting a survey of the Effective Altruist community showing that respondents disproportionately identify as male (seventy-one percent), white only (eighty-seven percent), and young (median age of twenty-eight, mean thirty-one, and seventy-eight percent younger than thirty-five)), *with* David Moss, *EA Survey 2020: Demographics*, EFFECTIVE ALTRUISM F. (May 20, 2021), https://forum.effectivealtruism.org/posts/ThdR8FzcfA8wckTJi/ea-survey-2020-demographics [https://perma.cc/AZ34-5MQL] (reporting a survey of the Effective Altruist community showing that respondents still disproportionately identify as male (seventy point-five percent), white only (seventy-five point-nine percent), and young (median age of twenty-seven, mean twenty-nine, and eighty percent younger than thirty-five)).

202. *Open Positions*, *supra* note 144 (noting that their "staff is spread around the world").

different countries or engaging in remote meetings held at odd hours to accommodate different time zones. This movement's ability to operate on a global level is facilitated by the rise of remote work and the related technological tools that have been developed and expanded during the COVID-19 pandemic. While this makes for a diverse community in some respects, participants still emphasized major representational deficits. The community continues to overrepresent the Global North and white men.

Diversity is valued in this community primarily for the sake of inclusivity and a desire to open the discussion around existential risk to a wider array of voices. But it is also possible that by expanding current-person representation, this movement may grow more effective. A more heterogeneous community may yield information that would otherwise be overlooked, and this may serve the movement's goal of maximizing impact.[203] Participants noted that cultivating a broader range of voices may be essential to persuading lawmakers and policymakers to support important actions relevant to existential risk.

To take an example from the international policy efforts, participants observed that some Global South diplomats appear somewhat resistant to proposals relating to existential risk, both because such proposals seem to divert attention away from issues currently affecting their constituencies and because of general distrust of the Global North countries where theories of existential risk have originated.

One response to this issue may be found in LPP's support for the development of new existential risk initiatives and Effective Altruism fellowships in the Global South.[204] The executive director of LPP is a law professor at Instituto Tecnológico Autónomo de México, where he has developed existential risk programming. Another LPP member from Kenya recently led an eleven-week fellowship with a group of around sixty law students in Nairobi. Although this participant reported that some fellows initially found longtermism "almost ridiculous . . . given the problems we have today of poverty and hunger . . . and corruption," his survey at the end of the term showed that the fellows' "minds were changed drastically" and they rated concerns for "far future people" as the cause area deserving of the greatest concern.[205] This participant observed that several fellows from this program went on to other positions with a focus on mitigating

---

203. *See* Holden Karnofsky, *Worldview Diversification*, OPEN PHILANTHROPY (Dec. 13, 2016), https://www.openphilanthropy.org/research/worldview-diversification [https://perma.cc/3XL5-REU9]; *see also* Luke Freeman, *"Big Tent" Effective Altruism is Very Important (Particularly Right Now)*, EFFECTIVE ALTRUISM F. (May 19, 2022), https://forum.effectivealtruism.org/posts/SjK9mzSkWQttykKu6/big-tent-effective-altruism-is-very-important-particularly [https://perma.cc/6VL2-QJP6] (defining the "Big Tent" approach of EA community building as one that "encourages 'a broad spectrum of views among its members'").

204. The first Effective Altruism conferences in Latin America and India were held in January 2023. *See Our First Conferences in Latin America and India*, CTR. FOR EFFECTIVE ALTRUISM (Jan. 27, 2023), https://www.centreforeffectivealtruism.org/blog/latin-america-and-india [https://perma.cc/4D9H-ZT4Q].

205. Interview with anonymous participant.

existential risk.[206] If this movement can shift to Global South leadership, participants noted that this may enhance the movement's ethics, inclusion, and accountability, as well as helping to create a more persuasive demand for legal and political action.

V. THE CULTURE OF EXISTENTIAL ADVOCACY

Having discussed the theories of efficacy (maximizing de facto moral impact) and accountability (with a focus on representing the interests of future generations), this Part examines how the priorities methodology is operationalized in the daily culture of social-change advocacy. One might assume that this methodology could serve as an ideal but would find little expression in practice. This model demands a commitment to evidence-based reasoning, while setting aside other considerations, biases, and incentives. It also demands a commitment to working on behalf of a future population that is invisible and difficult to even imagine. Yet, in spite of these challenges, participants seemed remarkably adherent to a set of scientific truth-seeking norms that support their methodology and their focus on future generations. As categorized in the findings presented below, these norms relate to uncertainty (Section A), deliberative rationality (Section B), dissent (Section C), and limiting group identity (Section D).

A. THE UNCERTAINTY NORM

Throughout the interviews and ethnographic observations in this study, one of the most ubiquitous cultural themes was uncertainty. The notion of existential risk is fundamentally a matter of uncertain risk assessment regarding possible future events. This differentiates existential advocates from social-change lawyers in many other contexts, where the occurrence of at least some degree of harm is viewed as a certainty, e.g., discrimination against some class of current living persons. For the existential advocates, their goal is to reduce the probability of a particular class of events. One participant noted that if their efforts helped to reduce the chance of an existential catastrophe from fifteen to ten percent that would be a major win for the movement.[207] Success looks like something (a catastrophe) not happening. Failure could similarly look like nothing has happened, in a sense, if a failure would involve a catastrophic loss of life such that the advocates would not live to experience the outcome. Moreover, it can be difficult to know how much risk reduction has been achieved and how much can be attributed to the work of advocates.

Even more fundamentally, participants expressed a great deal of uncertainty, both normative and empirical, about how much global priority should be placed on existential risk. In my ethnographic observations, I was struck by how often

206. *Id*.
207. Interview with anonymous participant.

participants engaged in lengthy discussions on the question of whether or to what extent existential risk matters.

Uncertainty was a central theme when participants applied the priorities methodology, which can be viewed as an effort to make best guesses about the expected impact of strategic actions. This methodology requires maintaining a sense of uncertainty so that one can continually reassess priorities rather than becoming overly attached to particular goals or strategies. The notion of "updating" views is pervasive in the vocabulary of this community, as participants often described changed perspectives, and either increasing or decreasing confidence in their views, upon receiving new information.

Thus, rather than perceiving uncertainty as an obstacle, this community tends to view it as something to embrace as a cultural norm.[208] As one participant put it, a core objective of this community is to "normalize uncertainty" both within existential risk organizations and in their outward-facing advocacy and educational efforts.[209] Members of this community regularly remind one another to maintain a sense of uncertainty, often admonishing peers who seem to overstate confidence in their claims. One participant explained this cultural norm by observing, "anyone who comes across as overly certain will be greeted with suspicion."[210] This norm was often framed as a matter of "epistemic humility," which is a foundational concept in Effective Altruism.[211] For example, posts on the Effective Altruism Forum generally begin with an "epistemic status," as recommended by the Forum guidelines, describing the degree of confidence the author has in their empirical and normative claims.[212]

In order to pursue strategic action under these conditions of uncertainty, and these norms that continually draw attention to uncertainty, the existential advocates look to expected value theory, decision theory relating to low-probability/high-impact events, and Bayesian reasoning.[213] But participants acknowledged that these rational foundations of their "uncertainty culture" are difficult to realize in the face of the general cognitive tendency to seek out more firm conclusions. This tendency may be heightened when engaging in the persuasive work of

---

208. *See, e.g.*, Keiran Harris, *Effective Altruism in a Nutshell,* 80,000 HOURS (Oct. 18, 2021), https://80000hours.org/2021/10/effective-altruism-in-a-nutshell/ [https://perma.cc/DK88-EMGP] (emphasizing the importance of uncertainty and humility within Effective Altruism).

209. Interview with anonymous participant.

210. Interview with anonymous participant.

211. *See e.g.*, Lizka, *infra* note 212.

212. Lizka, *Epistemic Status: An Explainer and Some Thoughts*, EFFECTIVE ALTRUISM F. (Aug. 31, 2022), https://forum.effectivealtruism.org/posts/bbtvDJtb6YwwwtJm7/epistemic-status-an-explainer-and-some-thoughts [https://perma.cc/5782-UKAJ] (suggesting that the epistemic status at the beginning of an EA Forum post should cover the author's biases, qualifications, reasons for believing their arguments, amount of effort spent on the post, feedback received in the writing process, intended audience, and confidence in the claims made).

213. *See* Bostrom, *Existential Risk Prevention as Global Priority*, *supra* note 6, at 15 (introducing a "maxipok" principle, which aims to maximize the probability of not having an existential disaster, and thus having an at least "ok" outcome); MACASKILL, *supra* note 7, at 53–62.

advocacy. In such contexts, participants noted that they often deemphasize the language of uncertainty, effectively code switching. As one participant reflected: "if you speak the language of [Effective Altruism] epistemics within policy culture, they will not like it . . . you need to speak a different dialect with them."[214]

Participants also stressed that too much emphasis on uncertainty can have some negative organizational effects. Uncertainty can inhibit action, where strategic discussions digress and become unduly complicated or lead to excessive changes in direction. As one participant explained, "in the end, one has to commit to a certain path at least for some time . . . [rather than] changing your trajectory every few weeks."[215]

B. THE DELIBERATIVE RATIONALITY NORM

The priorities methodology is rooted in the notion of combining the head with the heart—using the head to guide the heart to do the most good. This approach emphasizes the importance of deliberative "System 2" cognition.[216] While strong emotions and automatic, instinctive cognition can be useful when thinking about what issues are morally significant, participants worried that drawing too heavily on "System 1" could lead to focusing on the most familiar and personally relevant cause areas, while overlooking opportunities for greater impact.

It is conceivable that strong emotions would drive efforts to mitigate existential risk, such as fear about apocalyptic scenarios, anger toward particular entities that exacerbate risks, or hope for a utopian future. But these framings were uncommon and disfavored among participants. Some even reflected on their own "missing mood,"[217] wherein their concern for the "long-term future of billions and billions of people"[218] is not met by a commensurate emotional response. This missing mood may be a product of the cognitive biases that fundamentally limit our ability to comprehend the scale of existential risks (as discussed supra Part I). Even for the participants in this study who work with existential risk on a daily basis, it is difficult, as one participant put it, to "emotionalize uncertainty" regarding existential events and to generate strong feelings for "people who don't exist yet," because "you can't meet them" and "you can't have a graphic documentary about them."[219] One participant contrasted this missing mood with their work in animal welfare, where advocates would frequently "cry together" during

214. Interview with anonymous participant.

215. Interview with anonymous participant.

216. *See generally* DANIEL KAHNEMAN, THINKING, FAST AND SLOW 20–24 (2011) (describing "system 1" cognition as emotional and instinctive ("fast") and "system 2" cognition as more rational and deliberative ("slow")).

217. Interview with anonymous participant.

218. Interview with anonymous participant.

219. Interview with anonymous participant.

meetings,[220] and in social justice contexts, where this participant remarked that they "miss the . . . bleeding heart of being more emotionally drawn to the cause."[221]

But the "missing mood" characterization of this community should not be exaggerated, as some participants described their work in more emotional terms. For example, one participant explained that they were drawn to working on existential risk after undergoing extensive psychotherapy to get more "connected to [their] emotions" and to have more "emotional capacity for compassion with broader groups and issues," which led to thinking about how they could "care about other people in a meaningful way" and then thinking about whether there are "more and less effective ways of doing this."[222] Two participants reported drawing a poignant source of motivation from the 2020 entreaty of a suicide note written by a Harvard Law School Effective Altruism member, who wrote, as family members shared with the press: "Please look after each other, the animals, and the global poor for me."[223] One participant described their admiration for this late colleague's deeply felt and expansive compassion and "beautiful heart."[224]

The role of emotional reasoning in this field was sometimes framed as a matter of rationality. I observed an internal LPP debate over whether to continue using the term "rational" in the public description of their organizational culture, which, at the time, read: "Rational: We use evidence and careful analysis to tackle the world's most pressing problems . . . ."[225] Some members worried that this term may imply a simplistic rational/emotional dichotomy and a judgment toward others for being "irrational,"[226] although one participant noted that the term accurately describes "an interest in being good with thinking, good with statistics," and an effort to encourage "Bayesian thinking" and "ways of reducing cognitive bias."[227]

---

220. As this participant noted, "seeing animals suffer is emotionally extremely charged, and seeing animals happy is also emotionally charged." Interview with anonymous participant.

221. Another participant similarly noted, "I think it is infrequent that I'm super connected to the emotions of 'oh no, how horrible would it be if this [catastrophic event] really happened?' Compared to being involved in the social justice movement where I can see these things and experience them in my whole life . . . ." Interview with anonymous participant.

222. Interview with anonymous participant.

223. *See* Meagan Flynn, *Rep. Raskin and his Wife on Their Late Son: 'A Radiant Light in this Broken World*,' THE WASH. POST (Jan. 4, 2021), https://www.washingtonpost.com/local/md-politics/rep-jamie-raskin-and-wife-sarah-share-moving-tribute-remembering-their-son-tommy-raskin/2021/01/04/0ef01b30-4ee3-11eb-83e3-322644d82356_story.html [https://perma.cc/5NFW-D7G3] (noting that the deceased was a committed and vocal board member of Harvard Law School Effective Altruism).

224. Interview with anonymous participant.

225. Legal Priorities Project, Internal Document (on file with author).

226. *See* James M. Jasper, *Emotions and Social Movements: Twenty Years of Theory and Research*, 37 ANN. REV. SOCIOLOGY 1, 2 (2011) (challenging the rationality/emotion dichotomy with the observation that "feeling and thinking are parallel" and are "composed of similar neurological building blocks").

227. Anonymous participant, Remarks at a Legal Priorities Project Internal Meeting.

This terminological debate reflects a deep tension regarding the role of emotions in motivating this community. Effective Altruism is often said to require "not going with your gut" and instead "taking a step back, figuring out your values, and determining where you can have the most impact."[228] In *Doing Good Better*, William MacAskill recommends "combining the heart and the head."[229] This formulation was perhaps an easier fit with MacAskill's focus, at the time, on current-generation issues of global health and animal welfare. Combining the heart and the head might be more difficult in the context of existential risk, which raises the less emotionally salient issue of protecting future generations.

Some participants noted that they sought to resolve this tension by finding "outlets for the heart" in their pro bono and philanthropic contributions outside of their full-time daily efforts to reduce existential risk. One such participant explained that they felt more "emotional connection" to their donations to effective global poverty interventions, because "my heart is there [with global poverty], and my head is with longtermism."[230]

C. THE SUPPORTIVE DISSENT NORM

The two cultural underpinnings of the priorities methodology already discussed—uncertainty and deliberative rationality—are further supported by norms of dissent and heterogeneous discourse. Dissent is an express pillar of the Effective Altruism framework.[231] In the effort to determine what priorities and strategies can be expected to maximize impact, dissenting views can help reveal uncertainties and provide information for deliberative processes. The dissent norm was often framed as a matter of limiting "value alignment" in the community of existential advocates, which would lead to excessive groupthink about priorities.[232] Although members of the Effective Altruist community often discuss the extent to which someone is "EA aligned," many participants in this study seemed uncomfortable with the homogeneity implied by this terminology—two participants noted that "alignment" has even worse connotations in European languages where it implies something like marching in formation.[233]

---

228. *See* MACASKILL, *supra* note 7, at 10 (arguing that "relying on good intentions alone to inform your decisions is potentially disastrous").

229. *Id.* (noting that the Effective Altruist notion of "combining the heart and the head" can mean that the heart inspires altruistic pursuits and the head informs those pursuits by turning "good intentions into astonishingly good outcomes").

230. Interview with anonymous participant.

231. *See* William MacAskill, *Effective Altruism and the Current Funding Situation*, 80,000 HOURS (May 16, 2022), https://80000hours.org/2022/05/ea-and-the-current-funding-situation/ [https://perma.cc/GC3P-2NY2] (noting that the Effective Altruist culture of dissent and "independence of thought" is reflected in the observation that several of the "most-upvoted Forum posts" are "critical" or "critically self-reflective" in nature).

232. *See generally* CarlaZoeC, *Objections to Value-Alignment Between Effective Altruists*, EFFECTIVE ALTRUISM F. (Jul. 15, 2020), https://forum.effectivealtruism.org/posts/DxfpGi9hwvwLCf5iQ/objections-to-value-alignment-between-effective-altruists [https://perma.cc/97WW-YJGE].

233. Interview with anonymous participant.

Some of the non-profit organizations in this field provide rewards for evidence that tends to discredit conclusions about the organization's current priorities.[234] This includes "red team challenges," where prizes are awarded to the best criticism and counterarguments regarding common conclusions in the Effective Altruism community.[235] In the meetings that I observed, participants would often suggest a round of counterarguments. LPP conducted surveys, held discussions, and wrote a report on the topic of honest feedback and "normalizing being more critical toward each other's work."[236] These norms seemed to influence how participants saw their own work. An LPP member noted that they were hired for a research-focused position "with the goal of advancing longtermism" but explained that "advancing" was meant in a scientific sense, which includes developing counterarguments and identifying uncertainties, and thus "pursuing the truth, be it favorable to longtermism or not . . . through science and research."[237]

Given this commitment to continually expressing misgivings and objections, it is perhaps surprising that this culture is also marked by what LPP has labeled a "warm, kind, and supportive" environment, which was very much how this culture appeared through the lens of this qualitative study. For example, weekly all-hands meetings began with a round of "achievements and gratitude" where attendees shared personal updates, (e.g., openly disclosing details about health, personal challenges, childcare, pets, family, and favorite TV shows including multiple references to baking shows that were especially appreciated because the contestants are "so kind to each other").[238] In this round of personal updates and at other times, participants openly expressed appreciation for various things in their lives including assistance and feedback they received from their colleagues. LPP members also held regular sessions focused on mental health, team-building activities, "informal hangouts," and one-on-one "watercooler" meetings. When members spoke during online meetings, they were regularly greeted with supportive emojis, (e.g., hands clapping, party hat, hearts, and crying laughing).

This norm of kindness and mutual support seemed to play a crucial role in the culture of dissent. Support for a colleague's comment was most often directed toward their reasoning or articulation of different positions, not necessarily their

---

234. *See, e.g.*, Crosson, *infra* note 235.

235. *See, e.g.*, Cilliam Crosson, *Apply for Red Team Challenge [May 7 – June 4]*, EFFECTIVE ALTRUISM F. (Mar. 18, 2023), https://forum.effectivealtruism.org/posts/DqBEwHqCdzMDeSBct/apply-for-red-team-challenge-may-7-june-4 [https://perma.cc/FL2B-JNSD]; Stanford Existential Risks Initiative, *Fireside Chat at Stanford Existential Risks Conference*, YOUTUBE (Feb. 27, 2022), https://youtu.be/6JJvIR1W-xI [https://perma.cc/M7YM-AA29] (describing the norm within Effective Altruist discussions to include the question, "And the thing I'm getting wrong about all of this?"); Holden Karnofsky, *Learning by Writing*, LESSWRONG (Feb. 22, 2022), https://www.lesswrong.com/posts/ii4xtogen7AyYmN6B/learning-by-writing [https://perma.cc/6QWZ-U3YT].

236. In public-facing materials, LPP has described their cultural commitment to "discuss ideas openly and honestly, letting the best ideas win. Honest criticism and feedback are both welcome and expected."

237. Interview with anonymous LLP member.

238. Interview with anonymous LLP member.

conclusion, which generally remained subject to uncertainty and debate even at the end of a discussion. Some participants suggested that these warm interactions, when paired with a norm of valuing counterarguments, can encourage members to view disagreement as a means to find the best ideas rather than as a personal criticism. Members of the Effective Altruist community sometimes joke about their own linguistic norms, laughing in a self-effacing manner when they say "I claim . . . I *sub*-claim" as they present arguments. But this language reveals a serious discursive effort to keep a focus on the content of the claims, which are freely open for debate, rather than focusing on the person making the claims.

It is possible that these efforts to avoid offending one another could lead to a "too nice" dynamic, as one participant noted, such that members of the community could grow unwilling to express dissent out of fear of breaking a norm of amicability.[239]

Some commentaries have expressed the opposite concern, suggesting that Effective Altruism, and the existential risk community in particular, are inhospitable to certain lines of counterargument, (e.g., regarding the value of using the priorities methodology or regarding the conclusion that existential risk and AI risk in particular should be prioritized).[240] Some have suggested that a monoculture has taken hold in spite of rhetorical efforts to maintain a heterogeneous idea space.[241] But, overall, as observed in this study, this culture of what I reference in my field notes as "supportive dissent" was one of the most striking features of existential advocacy.[242]

Over the first few weeks of immersing myself in the existential risk community, I observed repeatedly in my notes that this community's norms of dissent and questioning assumptions and evidence far surpassed what I experience in academia—which is often thought of as a bastion of skeptical inquiry. LPP members and other participants in this study showed a consistent willingness to revisit and debate even the most fundamental concepts that motivate their work, as well as the meta-level question of finding the right mix of agreeableness and dissent.

---

239. Interview with anonymous participant.

240. *See*, *e.g.*, Carla Zoe Cremer & Luke Kemp, *Democratising Risk: In Search of a Methodology to Study Existential Risk* (2021); *see also* Centre for Effective Altruism, *Against Naïve Effective Altruism*, YOUTUBE (Nov. 20, 2017), https://www.youtube.com/watch?v=-2oRgxxafXk [https://perma.cc/56B6-9ETG] (describing the risk that a naïve understanding of Effective Altruism could lead to being overly concerned with "signaling contrarianism" and "appearing cool and wise by holding weird beliefs," and as a result being overly dismissive of common sense).

241. *Id.*

242. While writing this article, I learned that this term, "supportive dissent," is very similar to what some Effective Altruists call, "supportive scepticism in practice." *See* Michelle Hutchinson, *Supportive Scepticism in Practice*, EFFECTIVE ALTRUISM F. (Jan. 15, 2015), https://forum.effectivealtruism.org/posts/CkikpvdkkLLJHhLXL/supportive-scepticism-in-practice [https://perma.cc/X6DU-B7WJ]; *see also* Lizka, *Guide to Norms on the Forum*, EFFECTIVE ALTRUISM F. (Apr. 28, 2022), https://forum.effectivealtruism.org/posts/yND9aGJgobm5dEXqF/guide-to-norms-on-the-forum [https://perma.cc/57B6-GM2K] (advising that "when you criticize someone's point, consider doing so supportively").

D. THE EPISTEMIC IDENTITY NORM

The participants in this study seemed highly reluctant to identify as part of a larger group of existential advocates. Even full-time employees of organizations in this space were quick to express caveats and reservations when describing themselves as "Effective Altruists," "longtermists," or "advocates for existential risk mitigation."[243] As one participant explained, identifying too strongly with such labels could imply a "movement [with] core tenets you have to follow," whereas they conceived of themselves belonging to a "scientific community" marked by curious inquiry, dissent, and uncertainty.[244] This notion of belonging to a scientific community seems designed to limit a sense of group solidarity around a shared identity.[245] Participants suggested that a strong sense of solidarity would be undesirable because it would create "social pressure to conform," a preference for in-group members, and, as one participant put it, an "us vs. them . . . antipathy" toward out-groups.[246]

Similar concerns have been raised by sociologists who find that members of social movements tend to develop a strong preference for in-groups, while undergoing "identity work" to "preserve or to enhance their egos" by aligning their views with the goals of the movement.[247] While this solidarity process may make

---

243. Note that "existential advocacy" is a novel term of analysis used in this article. This is not a term circulating among advocates in this field.

244. Interview with anonymous participant.

245. One participant noted that Effective Altruism does not particularly lend to identification because it is a "very diverse movement" full of "different views and priorities" and a strong commitment to debate.

246. *See* Benjamin Todd (@ben_j_todd), TWITTER (Aug. 8, 2021, 3:54PM), https://twitter.com/ben_j_todd/status/1424458937286512647 [https://perma.cc/E4DA-RWSG] ("Turning [E]ffective [A]ltruism into an identity has been powerful, but has had many downsides," and this includes that it "[c]reates social pressure to conform."); Lizka, *Against "Longtermist" as an Identity*, EFFECTIVE ALTRUISM F. (May 13, 2022), https://forum.effectivealtruism.org/posts/FkFTXKeFxwcGiBTwk/against-longtermist-as-an-identity [https://perma.cc/8EXH-ASMG] (arguing that identifying as a longtermist . . . can "make it harder to change your mind based on new information" and can lead to confusion where the "group identity" encourages viewpoints that one would "otherwise not have adopted"); Helen, *Effective Altruism is a Question (not an ideology)*, EFFECTIVE ALTRUISM F. (Oct. 16, 2014), https://forum.effectivealtruism.org/posts/FpjQMYQmS3rWewZ83/effective-altruism-is-a-question-not-an-ideology [https://perma.cc/QA9F-E4VE] (arguing that rather than identifying as an Effective Altruist one should state that one is an "aspiring Effective Altruist," a "member of the Effective Altruism movement," or "interested in Effective Altruism"); *see also* Cullen O'Keefe, *What Would a Longtermist Flag Look Like?*, EFFECTIVE ALTRUISM F., (Mar. 24, 2021), https://forum.effectivealtruism.org/posts/efd4B2LLd3DXGivSP/what-would-a-longtermist-flag-look-like [https://perma.cc/2JPX-AFEY] (suggesting the possibility of creating a flag for longtermism, as has proven useful in "mature and successful movements," but raising the concern that a flag might encourage partisan psychology and "ideological loyalty," and that it could "stymie open and honest discourse about longtermism, including criticism thereof").

247. HADLEY CANTRIL, THE PSYCHOLOGY OF SOCIAL MOVEMENTS 162, vii (1941) (observing that, upon joining a social movement, "the individual is now an in-group member of a rather highly selected gathering," and that such identification with the movement can cause the individual to "lose themselves in some cause that seems strange or esoteric to the observer"); Jasper, *supra* note 226, at 13 (suggesting that movements tend to "minimize affective loyalties to anyone outside the group and maximize them to the group or its leaders," and that movements tend to frame in-groups positively relative to out-groups); David A. Snow & Doug McAdam, *Identity Work Processes in the Context of Social Movements: Clarifying the Identity/movement Nexus*, *in* SELF, IDENTITY, AND SOCIAL MOVEMENTS (2000) (Sheldon Stryker, Timothy J. Owens & Robert W. White eds.)

movements more effective in some respects, it may also tend to limit available "argument pools" due to dynamics of "enclave deliberation."[248] Moreover, participants in this study commonly cited psychological research on "moral tribes" and "righteous minds" widely read in the Effective Altruism community, suggesting that when groups band together around certain viewpoints they tend to falter in their truth-seeking and moral reasoning.[249] Yet, a recurring theme in my field notes was the surprising (to me) absence of judgment and derision toward out-groups. I had anticipated that a community that is working to set optimal priorities might view the rest of the world as setting the "wrong" priorities (e.g., failing to address existential risk) and thus would tend to define themselves in opposition to these other groups. But over the course of this study, such judgment was rarely expressed and did not appear to be a strong source of motivation.

The members of this community seek to limit, not entirely reject, group identification. Participants acknowledged some benefits of aligning personal and collective identities—similar to the "identity/movement nexus" cited by sociologists as a crucial element in the formation and growth of social movements.[250] Some participants described their affinity with fellow existential advocates in terms of a common "life plan,"[251] a set of background readings, and a general commitment to "thinking about making a fair world" through an evidence-based methodology.[252] This creates a basis for some degree of shared identity within the movement, although this identity is rooted primarily in a commitment to a methodology and an epistemic culture rather than any particular conclusions

---

(summarizing a body of literature suggesting that social movements are especially appealing to those with a "spoiled" identity who are engaged in a "collective search for identity").

248. CASS SUNSTEIN, HOW CHANGE HAPPENS (2019) (observing that when one conceives of oneself "as part of a group having a degree of connection and solidarity . . . group polarization is all the more likely, and it is also likely to be more extreme . . ." and noting that these dynamics "tend to suppress dissent and thus to lead to inferior decisions").

249. *See generally* JOSHUA GREENE, MORAL TRIBES: EMOTION, REASON, AND THE GAP BETWEEN US AND THEM (2013) (discussing the human tendency to rely on "tribal gut reactions" and our "automatic" rather than "manual mode"); JONATHAN HAIDT, THE RIGHTEOUS MIND: WHY GOOD PEOPLE ARE DIVIDED BY POLITICS AND RELIGION (2012) (detailing the psychological bases for the human tendency toward group-based morality, which leads to a failure to appreciate the perspectives of out-groups).

250. Snow & McAdam, *supra* note 247, at 50 (discussing mechanisms of identity convergence in social movements, including amplification (building on existing identities that are already congruent with the movement), consolidation (revealing the compatibility of identities that seemed inconsistent), extension (making existing identities more pervasive or salient), and transformation (forming new identities)); DAVID SNOW, COLLECTIVE IDENTITY AND EXPRESSIVE FORMS (2001) (describing how leaders of the mid-twentieth century Civil Rights Movement drew on church identity, in addition to racial identity, to mobilize and foster commitment to the cause, while also transforming identities of activists in the Freedom Summer); SUNSTEIN, *supra* note 248 (noting that group polarization has been useful to spur movements, citing a number of examples including feminism, the Civil Rights Movement, and nationalism); Eric Neyman, *Can Group Identity be a Force for Good?*, UNEXPECTED VALUES (July 4, 2021), https://ericneyman.wordpress.com/2021/07/04/can-tribalism-be-a-force-for-good [https://perma.cc/MRK3-EDDH].

251. *See, e.g.*, WINTER ET AL., *supra* note 25, at 2 (noting that the LPP project is the "main life project" for most members of the team).

252. Interview with anonymous participant.

about what cause areas or strategies should be prioritized. As one participant reflected: "We unite around our epistemic norms."[253] These accounts of what I referred to in my field notes as "epistemic identification" were near ubiquitous in my interviews and ethnographic observations.

This approach to group identity may not be the most powerful basis for motivation and recruitment. One participant described their relatively "weak" identification as an Effective Altruist and contrasted this with their experience in movements where they attended protest demonstrations for criminal justice reform and other "leftist" causes, which fostered feelings of "certainty and a sense of belonging and having a clear enemy."[254] Without a clear enemy or even a clear out-group, it is perhaps more difficult to decipher one's collective identity, that is, to demarcate where one's identity begins and ends. Moreover, epistemic identification may fail to provide the same "warm glow," as this participant put it, of belonging and collective voice.[255] But this same participant, when reflecting back on their engagement in other social movements, noted that what was missing was a "focus on reasoning," which they were grateful to find in the existential advocacy community.[256]

Other participants similarly observed that they had generally been disappointed by the lack of evidence-based reasoning and high epistemic standards in their experiences in academia, legal practice, government service, and in society generally. As already noted, these participants did not tend to denigrate these other contexts. Instead, they emphasized their sense of appreciation for the existential risk community (and the related Effective Altruist community). Some participants offered biographical accounts of first discovering these communities and "finding a home" when they realized the shared commitment to, as one participant put it, "take epistemics seriously."[257]

In sum, this analysis suggests a tension within the existential advocate identity, which offers a sense of belonging but also demands that one resist overly identifying with the group. Several participants noted that the shared experience of this tension can foster a sense of solidarity, whereby members identify with one another on the basis of the shared norm against over-identification. This form of solidarity is grounded in a contradiction (group identity based on resisting group identity) and so likely faces some inherent limitation. But perhaps this is to a desired degree. Some participants suggested that maintaining this identity tension is crucial to the cultural underpinnings of the priorities methodology. Too little identification with the movement might lead to a lack of motivation. But too

253. Interview with anonymous participant.
254. Interview with anonymous participant.
255. *Id*.
256. *Id*.
257. Interview with anonymous participant.

much identification might tend to undermine the cultural commitments to uncertainty, dissent, and deliberative decision-making.

VI. DISCUSSION

In his 2004 book, *Catastrophe: Risk and Response*, Richard Posner argued that the legal profession has failed to play its crucial role in mitigating catastrophic risks, including risks on the existential scale, and that this was largely due to the profession's lack of scientific methodology and mathematical reasoning.[258] This criticism took particular aim at lawyers, who, owing to their "culture of advocacy and doctrinal manipulation," tend to assume a truth favorable to their client and then seek to persuade others of that truth rather than engaging in a scientific process of evaluating which claims are most likely to be true.[259] Posner contrasts the scientific orientation "toward knowledge," as demonstrated through experimental research, with the law's orientation "toward action," "bending the rules . . . fitting them to goals," and asserting certitude, by, for example, declaring "my client is innocent, and that's the truth."[260] Posner advised law schools to recruit more students with STEM backgrounds and provide basic STEM education within the law curriculum.[261] This would then help produce a class of "catastrophic risk lawyers" who appreciate the prerequisite scientific methods, probabilistic claims, and decision theory.[262] These lawyers would recognize that low-probability harms can be deserving of legal attention and concern when the potential magnitude of the harms is sufficiently large.[263]

The participants in this study are living evidence of the exceptional case that Posner hoped law schools would foster—a group of lawyers who are addressing global catastrophic risks through an express commitment to scientific reasoning and truth-seeking.[264] These advocates generally lack the STEM backgrounds emphasized by Posner. Instead, they find a commitment to scientific reasoning in the priorities methodology drawn from the framework of Effective Altruism.

Posner's analysis would seem to suggest a culture clash between the notion of learning to "think like a lawyer" and what MacAskill calls learning to "think like

---

258. POSNER, *supra* note 25, at v–vii (observing that the law lags "dangerously behind an accelerating scientific revolution" because its "conventional methods for resolving science-laden legal disputes" are inadequate in the face of "increasing scientific complexity).

259. *Id*. This concern is reflected in the laments of legal empiricists that the profession disregards empirical research. *See* James D. Greiner, *The New Legal Empiricism & Its Application to Access-to-Justice Inquiries,* 148.1 DAEDALUS 64 (2019) (arguing that the legal profession is "not evidence-based in the scientific sense" and instead tends to "rely on gut intuition and instinct, not on rigorous evidence," and advocating for the "new legal empiricism," which has the potential to "transform the U.S. legal profession into an evidence-based field").

260. POSNER, *supra* note 25, at 201–02 ("The idea of subjecting a legal proposition to a decisive experiment . . . horrifies the lawyer.").

261. *Id*.

262. *Id*. at 201–09.

263. *Id*.

264. *Id*. at 201–02.

an Effective Altruist."[265] Similarly, scholars of professional responsibility may find the priorities methodology a surprising fit with the legal profession. Although lawyers are obligated to advance the public interest, the core mandate of their professional role centers on zealous advocacy for clients.[266] The tradition of cause lawyering (as discussed supra Part IV), places a greater focus on legal reform and social change. But even within this tradition, the notion that lawyers would find their deepest sense of accountability in a formal methodology for maximizing impact is anomalous. How have these lawyers become what we might label "maximizing lawyers" or "prioritarian lawyers," zealously committed to scientific reasoning and net moral impact?

One answer may be that the scientific model used by these advocates is a function of privileged identities. This community is disproportionately white, male, and elite educated. It is remarkably diverse in terms of representing different nationalities, but there is a tilt toward the Global North.[267] The prevalence of these privileged identities may tend to give this community less appreciation for current suffering and oppression in the world, leading to a focus on other populations who will exist in the future.[268] These privileged identities could also help to explain some of the cultural traits outlined in this article—efforts to enhance rational deliberation while limiting group identity and emotional reasoning.

If the prevalence of these privileged identities creates some biases or blind spots in the application of the priorities methodology, then it is all the more important that this movement continues its diversification efforts. As discussed in Part IV, diversity is valued by advocates in this field because it may lead to greater inclusivity and legitimacy, as well as more accurate assessments of prioritized causes and strategies. Recent scholarship on public interest law has noted that the pervasiveness of white leadership in many civil rights campaigns has led these movements to overlook key injustices and other crucial considerations.[269] Similarly, the existential risk community would likely benefit from more perspectives of people who experience different forms of injustice and who live relatively dystopic lives today, analogous to the permanent dystopian scenarios contemplated by scholars of existential risk.

---

265. *See generally* MACASKILL, *supra* note 7.

266. *See* Pepper, *supra* note 181; William H. Simon, *The Ideology of Advocacy: Procedural Justice and Professional Ethics*, 1978 WIS. L. REV. 29, 30 (1978); Richard Wasserstrom, *Lawyers as Professionals: Some Moral Issues*, 5 HUM. RTS. 1, 23 (1975).

267. *See supra* Part V.D.

268. It is important to note here that participants in this study seemed remarkably attentive to issues of present-day global health and poverty, which continue to be the cause areas where the most philanthropic dollars are spent in the Effective Altruism community. *See* Benjamin Todd, *How are resources in effective altruism allocated across issues?*, 80,000 HOURS (Aug. 9, 2021), https://80000hours.org/2021/08/effective-altruism-allocation-resources-cause-areas/ [https://perma.cc/FJ28-YZWY].

269. *See* Atinuke O. Adediran & Shaun Ossei-Owusu, *The Racial Reckoning of Public Interest Law*, 12 CAL. L. REV. 1, 1–2 (2021) (calling for greater scrutiny of the racial composition of U.S. public interest law as it impacts marginalized communities).

The effort to create a rigorous methodology for maximizing impact may also be influenced by the nature of existential risk as a cause area, which is an unusual topic to serve as the subject of a movement for legal and social change. The empirical literature on social-change lawyering has primarily focused on grassroots social movements that address matters of public controversy and pervasive cultural norms (e.g., how we relate to one another across lines of race, gender, and sexuality).[270] In contrast, the advocates examined in this article are addressing an issue that might not be particularly affected by the general population's daily norms and values. Moreover, even if these advocates were to attempt to form a broader social movement, it may be exceedingly difficult to mobilize large populations around low-probability/high-impact harms, particularly where such harms are framed as a threat primarily to a distant population of future generations. This lack of what socio-legal scholars call "mobilizing frames" may help to explain why these advocates are attracted to the priorities methodology.[271] If this community remains a relatively small (although hopefully diversifying) group of advocates dedicated to working full-time on the complex strategic considerations around existential risk, this may be the sort of community that is well-positioned to take a scientific approach to setting priorities and designing strategies.

While this notion of maintaining a small expert-based movement holds a great appeal for many of the existential advocates, this community has recently stepped into a very high-profile public spotlight. Before 2022, existential risk had already been the subject of public-facing books, videos, blogs, and podcasts,[272] as well as in-depth New Yorker profiles of the leading scholars.[273] But 2022 saw the field truly, and quite suddenly, enter the mainstream of popular and political debate. This was the product of three events in particular. First, in August 2022, the release of William MacAskill's book, *What We Owe the Future*, was a New York Times Bestseller and received high praise across leading news outlets, including a Time Magazine cover story.[274] Second, just three months later, a major donor to

---

270. *See supra* Part I.B.

271. *See* Snow et al., *supra* note 58.

272. *See generally* Nick Bostrom, *The End of Humanity*, YOUTUBE (Mar. 26, 2013), https://www.youtube.com/watch?v=P0Nf3TcMiHo [https://perma.cc/5LE4-HK4J]; MACASKILL, *supra* note 7; ORD, *supra* note 1; Kurzgesagt – In a Nutshell, *The Last Human – A Glimpse into the Far Future*, YOUTUBE (June 28, 2022), https://www.youtube.com/watch?v=LEENEFaVUzU [https://perma.cc/A7LH-FKPL]; FUTURE OF LIFE INST., https://futureoflife.org/the-future-of-life-podcast [https://perma.cc/585V-WAZR] (last visited Sept. 23, 2023); *see also* VOX: FUTURE PERFECT, https://www.vox.com/future-perfect [https://perma.cc/6EZB-LCMA] (last visited Sept. 23, 2023).

273. *See* Raffi Khatchadourian, *The Doomsday Invention*, NEW YORKER (Nov. 23, 2015), https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom [https://perma.cc/V4ZF-Z4CG]; Gideon Lewis-Kraus, *The Reluctant Prophet of Effective Altruism*, NEW YORKER (Aug. 15, 2022), https://www.newyorker.com/magazine/2022/08/15/the-reluctant-prophet-of-effective-altruism [https://perma.cc/27SB-8WPN]; Corinne Purtill, *How Close is Humanity to the Edge?*, NEW YORKER (Nov. 21, 2020), https://www.newyorker.com/culture/annals-of-inquiry/how-close-is-humanity-to-the-edge [https://perma.cc/5RWJ-RXLL].

274. *See, e.g.*, MacAskill, *supra* note 95; *Three Sentences that Could Change Your Life,* THE EZRA KLEIN SHOW (interview with William MacAskill) (Aug. 9, 2022), https://www.nytimes.com/2022/08/09/opinion/

the field of existential advocacy, Sam Bankman-Fried, was charged with corporate fraud on a massive scale.[275] Much of the public outcry about Bankman-Fried took aim at his philanthropic and political efforts, with many journalists and academics ridiculing existential risk as a tech billionaire fantasy—or, even worse, as a misconceived public relations effort to bolster the business interests of the greatest fraudster of a generation.[276] Third, the November 2022 release of ChatGPT brought public awareness to transformative advances in artificial intelligence, leading to a new appreciation for the strange and potentially dangerous world of unprecedented technology we may now be entering.[277] Other events that have recently brought public attention to existential risk in the past few years include: COVID-19, which generated near ubiquitous discussion of global pandemics; the invasion of Ukraine, which has raised fears of autonomous weapons and nuclear war; and extreme fires, floods, and temperatures, which has brought a new urgency to extreme climate change scenarios.[278]

A non-profit research organization was recently founded to investigate whether the existential risk community (and related communities associated with Effective Altruism) should pursue broader movement-building strategies.[279] Drawing on empirical research and analogous case studies, the researchers strongly recommended the formation of social movement organizations, which would organize protest demonstrations and other actions intended to shape public opinion and put pressure on key decision-makers.[280] Beginning in the spring of 2023, a small protest movement had held more than a dozen demonstrations

---

ezra-klein-podcast-will-macaskill.html [https://perma.cc/S52Z-GAAD]; Bajekal, *supra* note 23; *What We Owe The Future*, COMEDY CENT.: THE DAILY SHOW WITH TREVOR NOAH (interview with William MacAskill) (Sept. 9, 2022), https://www.cc.com/video/8fl6g9/the-daily-show-with-trevor-noah-william-macaskill-what-we-owe-the-future [https://perma.cc/N5X3-DXSX]. *But see* Christine Emba, *Opinion: Why 'Longtermism' Isn't Ethically Sound*, WASH. POST (Sept. 5, 2022), https://www.washingtonpost.com/opinions/2022/09/05/longtermism-philanthropy-altruism-risks [https://perma.cc/8QGV-6CRJ]; Alexander Zaitchik, *The Heavy Price of Longtermism*, NEW REPUBLIC (Oct. 24, 2022), https://newrepublic.com/article/168047/longtermism-future-humanity-william-macaskill [https://perma.cc/37DW-QM6K]; Setiya, *supra* note 46.

275. *See* De Vynck, *supra* note 47.

276. *See* Szalai, *supra* note 48.

277. *See* Kelsey Piper, *ChatGPT Has Given Everyone a Glimpse at AI's Astounding Progress*, VOX (Dec. 15, 2022), https://www.vox.com/future-perfect/2022/12/15/23509014/chatgpt-artificial-intelligence-openai-language-models-ai-risk-google [https://perma.cc/F8KM-5M3l] (noting that ChatGPT, a chatbot drawing from a large language model, is "the general public's first hands-on introduction to how powerful modern AI has gotten").

278. *See, e.g.*, Amy Maxmen, *Has COVID Taught us Anything About Pandemic Preparedness?*, NATURE (Aug. 13, 2021), https://www.nature.com/articles/d41586-021-02217-y [https://perma.cc/K9ZP-SATX] (noting increased awareness of global pandemic risks in the wake of COVID-19 but a general lack of response among lawmakers); *see also* Bridget Williams & William MacAskill, *Investing in Pandemic Prevention is Essential to Defend Against Future Outbreaks*, BULL. ATOMIC SCIENTISTS (Nov. 2, 2022), https://thebulletin.org/2022/11/investing-in-pandemic-prevention-is-essential-to-defend-against-future-outbreaks [https://perma.cc/76VV-P3B7].

279. SOCIAL CHANGE LAB, https://www.socialchangelab.org/ [https://perma.cc/NQV4-A6AT] (last visited Sept. 23, 2023).

280. *Id.*

around AI safety with a core focus on existential risks.[281] It is not yet clear where this will develop into a broader movement. What is clear is that the cat is out of the bag and a significant degree of public engagement around the issue of existential risk is now unavoidable.

The academic literature on law and social change seems to generally support the notion of expanding the community of existential advocates. As discussed in Part I.B, scholars in this field overwhelmingly recommend "integrated advocacy," noting that legal activists in other social movements have found greater efficacy and accountability by embedding their legal work within larger movements.[282] This approach is thought to help promote lasting social change by creating a collective demand for reform, which motivates favorable legal and political decision-making, and diminishes backlash to legal victories.[283]

Just how far should the existential risk community take this received wisdom from the literature? Should they recruit new members with the widest possible net, expanding their movement to create a larger collective voice? Should they focus on public opinion, seeking to foster a world where people generally understand the notion of existential risk and support interventions as opposed to dismissing the issue as a matter of science fiction or billionaires' whims? Or should this community stay relatively narrow, working to put experts in conversation with powerful decision-makers (e.g., lawmakers and judges)?

These are questions for further research beyond the scope of this article, but one consideration bears directly on this article's core findings: efforts to expand this movement may tend to compromise the culture underlying the priorities methodology. The "mobilizing frames" that social scientists have identified as key ingredients for building a broad social movement are seemingly point-by-point the exact opposite of the cultural commitments of the existential advocates as detailed in this Article.

For example, broad expansion of this movement may require developing a stronger sense of group identification. The movement to mitigate existential risk has been so resistant to group identity that it does not even have a label for itself. As already noted, the term "existential advocates" was coined in this Article, but the term is not generally used by members of this movement—although some identify, often with considerable hesitation, as "longtermists" or "Effective Altruists."

Moreover, most social movements depend on some group taking "identity ownership" over an issue, which can help with recruiting a large population and developing a unified voice.[284] It is conceivable that the existential advocates could frame themselves as a youth movement acting on behalf of future

---

281. *PauseAI Protests*, PAUSEAI, https://pauseai.info/protests [https://perma.cc/VG4Z-7LYA] (last visited Apr. 19, 2024).

282. *See* CUMMINGS, *supra* note 103.

283. *See supra* Part I.B.

284. *See* Doug McAdam, *Social Movement Theory and the Prospects for Climate Change Activism in the United States*, 20 ANN. REV. POL. SCI. 189 (2017).

generations. Just as we have seen in response to climate change, this could involve widespread student walkouts, sit-ins, and other protests. But the participants in this study tend to strongly resist these identity-based dynamics, fearing that an emphasis on collective identities might undermine the movement's commitments to dissent, uncertainty, and deliberative rationality.

Efforts to persuade broader populations may require presenting existential risk mitigation in familiar terms, meeting people where they are, putting less emphasis on uncertainty, and making issues more emotionally salient. This affective dimension could be amplified by stirring up fear and other strong emotions around existential risk. Some advocates have explored this approach by producing short films about existential risk scenarios, (e.g., depicting swarms of small autonomous drones that execute people in great numbers).[285] Most of the participants in this study seemed to disfavor this approach on the grounds that, as discussed above in Part V.C., motivations rooted in powerful emotions may cause the movement to drift toward a focus on near-term and smaller-scale catastrophic risks.

At least at this early stage of their movement, the existential advocates seem very effective at avoiding this drift. They are remarkably consistent in their methodology and their cultural commitments, assessing nearly all strategic decisions according to how much each option is expected to reduce our overall level of existential risk. This ability to "keep your eyes on the prize," in the refrain that echoed through the Civil Rights Movement,[286] is a defining feature of the social-change lawyering observed in this study. As this community scales up and pursues more direct legal interventions, their culture will likely need to adapt and compromise. But maintaining some core of advocates who are relatively uncompromising in their commitment to the priorities methodology could be vital to this movement's success. As humans, we carry cognitive biases that limit our ability to assess existential risk, an issue that is uncertain, unprecedented, large-scale, and seemingly remote—affecting future generations who are an abstraction well beyond our usual moral circles.[287] Moreover, legal and political decision-makers are incentivized to focus on issues of immediate concern to voters, markets, and interest groups. Creating a legal movement dedicated to overcoming, or at least challenging, these biases and incentives is no small task. These advocates approach this task with a framework of continual evidence-based analysis backed by a culture of scientific reasoning. In this way, they aim to keep their focus on representing the current and future generations whose well-being and existence may hang in the balance.

---

285. *See, e.g.*, FUTURE OF LIFE INST., *Slaughterbots* (Nov. 13, 2017) https://futureoflife.org/video/slaughterbots/ [https://perma.cc/8SFT-BKLE].

286. This phrase is attributed to a lyric from the traditional African American spiritual, "Gospel Plow." *See* DUKE ELLINGTON & MAHALIA JACKSON, *Keep Your Hand on the Plow*, *on* LIVE AT NEWPORT 1958 (Columbia Records 1958).

287. *See* John & MacAskill, *supra* note 88, at 5; SUNSTEIN, *supra* note 88; WINTER ET AL., *supra* note 25, at 21.