25

THE AI EXTREME RISK MITIGATION PHILANTHROPIC SECTOR

A philanthropic ecosystem at the forefront of AI

Siméon Campos and Daniel S. Schiff

Artificial intelligence (AI), like previous advanced information technologies, offers new opportunities and challenges for philanthropic efforts. As a tool, it offers a variety of new methods for designing new interventions, evaluating organizational effectiveness, or improving workflows and organizational infrastructure. As a substantive field, AI represents a complex and dynamic set of social challenges that need to be addressed whatever the objective of a non-profit organization, foundation, or public sector actor.

In this chapter, we explore the implications of AI for actors and organizations focused on philanthropy, in the broadest sense. In Section 1, we introduce two notions of AI in the context of philanthropy: the implications of AI-as-a-*tool* for philanthropic actors and organizations, and AI as a substantive *field* and *target* of philanthropic action. We then focus on the latter topic by providing a history, overview, and analysis of the AI philanthropic sector. In particular, we discuss a loose set of communities focused on assessing and mitigating extreme risks related to AI, which are often associated with attention to AI governance and AI safety. While the appropriate definitions and boundaries between subcommunities are themselves a focus here, for simplicity, we refer to the movement as the "AI extreme risk mitigation philanthropic sector" (AIERMPS), defined as the set of actors and organizations that seek to mitigate extreme AI-related risks through and with private philanthropy.

In Section 2, we review the founding of the AIERMPS and provide an overview of the core elements of its culture, strategy, and organizational landscape. We discuss highly unique aspects of its community, such as its strong intellectual entanglement with transhumanism, rationalism, and utilitarianism. These aspects of the AIERMPS, combined with its close relationship with major philanthropic donors, help to explain its rise in leading private sector organizations during the new AI spring of the 2010s. In turn, the engagement of these actors in leading AI research labs and policy conversations foregrounded its rise from a niche to a mainstream movement in the 2020s, following the public popularization of generative AI.

In Section 3, we turn to familiar and unique challenges faced by the movement. For example, while many philanthropic movements face persistent disagreements about structure or strategy, the AI extreme risk mitigation movement faces a unique challenge in that some of its concerns may only bear out in the future and thus remain inaccessible to study, limiting the opportunity for

DOI: 10.4324/9781003468615-29

useful feedback on effective strategies. In addition, philanthropic efforts are unusually focused on technical, rather than social issues, such as addressing the inexplicability of black box AI models and promoting the alignment of these systems with (unsettled) human values. We examine related structural challenges associated with the community, such as the tensions between trying to advance highly capable AI through leading research labs and technology companies and doing so as part of advancing safety efforts.

Based on these analyses, Section 4 highlights what the AI philanthropic community could learn from related fields and social movements. We consider the significance of centering AI itself as a technical issue with relatively low public salience compared to climate change or animal welfare, which were instead framed early on as social problems. Indeed, this difference between mass public movements and an elite-led project has implications for advocacy, fundraising, influence efficacy, and policymaking efforts, along with aspects of the community like its demographic profile. We also consider an issue common to philanthropic communities: internal division, particularly in the AI domain, between actors focused on so-called short-term issues (i.e., AI ethics) and long-term ones (i.e., AI safety). We discuss how other social movements have worked to manage conflicts and build coalitions. While the AI philanthropic movement may be most similar to other elite-driven efforts, such as the open-source Internet community or the nuclear safety community, we conclude that it has much to learn from other enduring social movements.

1 An overview of the AI philanthropic landscape

1.1 AI-as-a-tool

To a significant degree, AI continues historical conversations in the philanthropic community related to the use of technology and analytics. Associated concepts include data, ICT systems, advanced analytics, big data, automated decision systems, and e-government. Organizations in the non-profit and public sectors, and even wings of the private sector focused on corporate responsibility have sought to leverage these tools to advance their efforts.

For example, the philanthropic movement has considered the use of advanced analytics to predict the profile or giving behavior of donors (Eiland et al., 2021; Mittal & Srivastava, 2021; Sulaeman, 2018), to enhance fundraising efforts (Key, 2001; Vequist, 2014), and to guide potential donors or government organizations in the selection of effective charities themselves (Ramirez & Saraoglu, 2009; Singer, 2019; Stern, 2013). This emphasis on effectiveness and efficiency became increasingly popularized through several movements, including the turn to evaluation planning and logic modeling (Kaplan & Garrett, 2005), the early effective altruism movement, and the evidence-based philanthropy and policy movements of the 2000s and 2010s (Braverman et al., 2004; Fiennes, 2017; Johnson, 2018; Pawson, 2002).

Government agencies and private foundations alike increasingly require the use of logic models, evaluation plans, cost-benefit analysis, cost-effectiveness analysis, randomized controlled trials, or other high-quality quasi-experimental studies in areas ranging from health care, education, and public welfare to innovation policy and global development (de Souza Leão & Eyal, 2019; Pawson, 2002; White, 2019). A shared logic across these efforts is the desire to ensure that philanthropic or public money is effectively achieving its goals, while a secondary logic is to do more with less, especially in the face of reduced federal funding or stressed fundraising (Gore, 1993; Osborne, 1993).

For these reasons, the use of data, analytics, and now, AI, is an ongoing but largely unfinished ambition of the philanthropic sector, pursued in the hope that it can improve program efficiency,

streamline operations, enable more effective intervention design, and enhance evaluation (Henriksen & Blond, 2023; Madeo, 2022; Shapiro & Cody, 2015; Voida, 2014). Moreover, while some of these advances are particularly salient to the philanthropic sector, still other operational advances due to informatics and AI are increasingly being adopted across the public and private sectors in general, such as the use of AI to improve recruitment hiring, performance management, legal and compliance functions, procurement, financial management, worker training, and more (Noordt & Tangi, 2023; Wirtz et al., 2019; Zuiderwijk et al., 2021). In short, much of AI's promise to the philanthropic sector is similar to its potential for productivity gains more broadly, simply applied to the set of organizations focused on philanthropy. However, the philanthropic sector has historically been resource-constrained, with below-market wages, precarious work, and scrutiny over operations that often limit its ability to retain technical talent and experiment with new innovations. It is no surprise, then, that sectors like financial services and telecommunications are better positioned to innovate and have become more robust adaptors of AI, while the philanthropic sector lags behind.

A characteristic of AI compared to other analytical tools is that there are some unique capabilities associated with AI that enable relatively novel ways of adopting AI-as-a-tool. These efforts leverage the powerful predictive capacity of AI, emanating from its ability to find subtle patterns in large data sets. Most commonly referred to as "AI for good" or "AI for social good," the actors and organizations aligned with this movement focus on creative applications of AI to solve large-scale social problems. For example, in the environmental space, AI has been used to promote forest ecosystem restoration, track wildlife diversity, and identify appropriate locations for renewable energy installations (Guo et al., 2023; Isabelle & Westerlund, 2022; Schiff et al., 2021; Schwartz et al., 2021). In health care, AI has been lauded for its application to medical imaging, medical diagnosis, robotic surgery, and even the creation of new drugs and vaccines (Isbanner et al., 2022; Morley et al., 2020; PricewaterhouseCoopers, 2017).

In public services and social welfare, AI has been used to triage access to housing services, provide mental health care, translate government documents for non-native language speakers, or answer questions via chatbots, among many other "for good" applications in finance, education, and other sectors (Chui et al., 2018; Cowls et al., 2019, 2021; Herzog et al., 2021). This movement is also associated with efforts such as applying AI to achieve the sustainable development goals (AI4SDGs) or goals related to equity (AI4Equity), development, human rights, well-being, and more (Cath et al., 2020; Mazzi et al., 2023; Schiff et al., 2020; Stahl et al., 2023; Wakunuma et al., 2022). Despite criticisms related to ethical washing, corporate capture, or narrow technological solutionism (Holzmeyer, 2021), "AI for good," which represents a family of philanthropic uses of AI-as-a-tool, may be the most prominent manifestation of AI in the philanthropic space.

1.2 AI-as-a-domain

However, AI's implications for philanthropy extend beyond its use as a tool. There is now a prolific body of work and research considering the impact of AI on essentially every social and economic sector. Across civil society, private sector organizations, academia, and government, actors have been thinking through the challenges and risks that AI poses for discrimination, accountability, manipulation, surveillance, labor displacement, arms races, power imbalances, inequality, and much more (Attard-Frost et al., 2022; Coeckelbergh, 2020; Prunkl & Whittlestone, 2020; Schiff et al., 2022). These issues cut across individuals, communities, and populations, economic and social sectors, professional roles and disciplines, and time frames.

As a consequence, AI ethics has developed into a subfield in its own right, with branches in philosophy, policy, sociology, science and technology studies, economics, information science, communications, history, and other disciplines. The movement draws on predecessors in engineering and computing ethics, robot ethics, machine ethics, and numerous other traditions, and was driven in the 2000s–2020 by concerns about autonomous weapons, labor displacement, and algorithmic bias, among other issues. Sometimes associated with the acronym FEAT or FATE (fairness, accountability, transparency, and ethics), the ethics movement has especially advanced a focus on bias and discrimination, transparency of AI systems and governance, power and inequality, and more (Floridi & Cowls, 2019; Howard et al., 2019).

The AI ethics movement has had unusual success for an emerging technology policy domain in receiving policy attention; it is overwhelmingly common for government policy documents, as well as private sector documents focused on AI strategy, to include large sections discussing AI's ethical implications (Schiff, 2023). Academic and industry conferences, professionals in newly created job roles in the public and private sectors, and a suite of new non-profit and for-profit organizations focused on trustworthy, ethical, and responsible AI have advanced a range of policy problems and solutions to address challenges in this field (Benjamins, 2020; Maas, 2023; Perry & Uuk, 2019), with proposals as narrow as the ethical development of future computer scientists to as broad as the creation of global institutions or major cultural and economic reforms.

While this review simplifies a vast movement of actors concerned with the philanthropic implications of AI, it is helpful to distinguish this community from a related but distinct movement, the AIERMPS. The AIERMPS is more commonly associated with concepts such as AI safety, AI alignment, or extreme AI risk. This field is chiefly concerned with the study, practice, and governance of AI systems to ensure that they remain aligned with human goals and do not pose an unacceptable (or even existential) threat to human flourishing. Like the AI ethics community, the AI safety or extreme risk community is composed of researchers, organizations, and donors, as well as a growing cohort of policy actors. However, the AIERMPS differs somewhat from the AI ethics community. Indeed, it focuses substantially on technical aspects of AI safety, including the robustness, interpretability, and security of AI systems, and emphasizes severe to catastrophic risks, which are sometimes considered speculative by critics of the community. The movement has also encouraged its members to focus explicitly on AI policymaking, fundraising, and technical research to advance adequate regulation, accountability and transparency measures, independent auditing, export controls, and other regulatory regimes that could mitigate AI-induced risks.

To some extent, the AI ethics and AI extreme risk mitigation communities share similar goals and strategies (Baum, 2018; Stix & Maas, 2021). Actors in both subcommunities may be concerned about privacy, manipulation, deception, autonomy, misinformation, trustworthiness, independent auditing, negative environmental impacts, labor displacement, human rights, and so on. However, the AI extreme risk mitigation community is particularly notable in its focus on catastrophic, extinction-level, or "existential" risks that could arise from AI-enabled threats to physical and financial infrastructure, manipulation of humans, biological weapons, cyber warfare or terrorism, and rogue AI systems (Hendrycks et al., 2023). Key concepts associated with this community include transformational AI, AI safety, long-termism, transhumanism, rationalism, effective altruism, artificial general intelligence (AGI), artificial super intelligence (ASI), and more. Importantly though, these broad strokes descriptions of these two communities inevitably oversimplify and mischaracterize many individuals and organizations active in these philanthropic spaces. Nevertheless, they are useful for providing a rough understanding of the unique (and non-unique) aspects of the AI extreme risk mitigation community.

To briefly summarize, we have so far presented a discussion of the relationship between AI and philanthropy, including the many uses of AI-as-a-tool that can be employed by philanthropic actors for operational improvement or for "social good" use cases, followed by a discussion of the philanthropic movements focused on AI-as-a-domain or cause area. In the next section, we turn to a particular subcommunity, the AIERMPS, given its important role in the overall AI landscape and its relative lack of scholarly and analytical attention.

2 The AI extreme risk mitigation philanthropic sector: culture, strategy, and organizational landscape

2.1 The founding of the sector: a strong cultural identity and shared principles

The AI extreme risk mitigation philanthropic sector (AIERMPS), which we again define as the set of actors and organizations that seek to mitigate AI-related extreme risks through and with private philanthropy, has coevolved since its inception and to date with a very specific culture that strongly shapes its identity (Lazar & Nelson, 2023). The first major component of this culture is transhumanism, a movement defined by its "taking a decidedly positive view of the prospect of a 'post-human" future' (Birnbacher, 2009). Two of the first research organizations dedicated to studying AI risks have close ties to this component.

Founded in 2000 by Eliezer Yudkowsky, originally as the "Singularity Institute for AI," the Machine Intelligence Research Institute (MIRI) was one of the first organizations to work on AI extinction risks, even before AI gained public attention in the 2010s. While its initial focus was on *accelerating* the development of artificial general intelligence (AGI), that is, an AI more capable than humans at any cognitive task, MIRI began working in 2005 on countering the risks that such an AI might pose to humanity. This turn came after staff realized that a superintelligence could potentially cause significant harm to humans, up to and including extinction (Nast, 2015). This formative insight, highlighting the potentially extreme harms of advanced AI systems, would become the core mission of the field of AI safety and the AIERMPS.

Starting modestly with philanthropy from a small cohort of private donors, the AIERMPS has grown steadily over time, especially as AI has attracted more attention thanks to new players such as the research-focused organizations (now industry-affiliated organizations) DeepMind and OpenAI, and thanks to major AI discoveries in the 2010s. At this time, the first large academic center dedicated to mitigating potential risks from AGI, The Future of Humanity Institute (FHI), was founded in Oxford in 2003 by Nick Bostrom, with the mission to study existential risks (including but not limited to those posed by AI) (Ó hÉigeartaigh, 2017). After more than a decade of conversations with members of MIRI on mailing lists covering a wide range of topics, including transhumanism (Taillandier, 2021), Bostrom published the book *Superintelligence* in 2013, drawing largely on a range of ideas that had emerged from members of MIRI and FHI (Bostrom, 2014). This substantially contributed to making AI risks more well-known.

Pioneers in the field, FHI and MIRI have developed idiosyncratic views that focus substantially on the technical challenge of aligning AGI with human values. First, they consider the robust alignment of AGI with core human values to be an extremely difficult technical problem (Bostrom, 2014; Yudkowsky, 2016). Second, they consider it likely that at some point, AI progress will accelerate sharply (Yudkowsky, 2013), manifesting in an exponential increase in AI's capabilities, an event called the "singularity." This rapid increase in capabilities can be analogized to the same way that DeepMind's AlphaZero (one of the first superhuman-level AI Go players) went from subpar to substantially better than world champions at Go after only a few dozen hours of training (Silver

et al., 2017). Some core MIRI staff believe that the same rapid increase in AI capabilities could happen with respect to all tasks currently performed by humans, once the field of AI develops sufficiently powerful and general AI systems. Finally, a criticism of this movement has been that while bringing a fresh perspective to the field, only a minority of FHI and MIRI members have backgrounds in technical AI research, which carries the risk of imparting views and concepts that may lack relevance when applied to concrete AI systems.

However, by writing most of the early influential articles and shaping core conceptualizations of AI risks, MIRI and FHI have had a long-lasting impact on the AIERMPS, on the field of AI safety, and on AI more broadly. Even prominent industry actors such as DeepMind, a startup created in 2010 and now an industry leader and subsidiary of Google (Alphabet), have been closely linked with MIRI since its inception. Indeed, Eliezer Yudkowsky introduced DeepMind's founders, Demis Hassabis and Shane Legg, to MIRI's lead donor, Peter Thiel. Thiel subsequently became the first investor in DeepMind (Metz, 2022), the first AI company with the stated mission to develop AGI. Beyond these personal and financial connections, MIRI and FHI's intellectual conceptualizations of AI risks have also had a lasting influence. For instance, while new technological advances, especially related large language models, have led to the evolution of some concepts and frameworks applied to current AI systems (Ngo et al., 2023), many of the concepts that MIRI and FHI emphasized, such as "agents," "corrigibility," and "alignment," are still widely discussed and actively used in the AI safety research field (Byrnes, 2021) as well as in policy, evidenced by the creation of international AI safety workshops and the U.S. NIST AI Safety Institute.

During the late 2000s and the early 2010s, large segments of the transhumanist component of the culture evolved into a new subculture, significantly shaped by MIRI and its leader Eliezer Yudkowsky, now commonly referred to as "rationalists" (Matthews, 2023). Built around a set of concepts, ways of thinking, and idiosyncratic preferences around discourse, all dedicated to improving rationality, MIRI and the surrounding community have fostered a remarkable subculture. For example, a unique component of this subculture is that much of it has developed, and continues to be expressed and practiced, through a shared blog called *LessWrong*, based on a core set of writings by Yudkowsky known as "The Sequences."

This culture values norms that differ from those more commonly used in everyday society, such as its openness to inconvenient truths, high standards for what constitutes acceptable or productive discourse, and a preference for systematic and quantitative calculation to determine true or ethically sound positions. Because of its influence on the AIERMPS, it is partly responsible for the criticism that AI safety is characterized by a "near-monoculture" (Lazar & Nelson, 2023). Beyond the AI risk mitigation philanthropic sector, this culture has also affected AI industry leaders (Matthews, 2023), in large part because the culture has become very prominent in Silicon Valley and technology circles, where most of the top AI companies working to develop AGI have been founded. Thus, intellectual, social, historical, and even geographical factors have played a role in shaping the AIERMPS.

Finally, another critical dimension of this community with a special relationship to philanthropy is the community focused on evidence-based philanthropy. In the early 2010s, and increasingly over time, the rationalist culture has been increasingly influenced by the rapidly growing effective altruism (EA) movement. Coined in 2011 by a group of Oxford academics, the EA movement emerged from the convergence of a focus on rationalism (how to think better), altruism (how to organize philanthropy and charitable giving more efficiently, including making career decisions), and futurist concerns (how to ensure that human civilization thrives) (Chivers, 2019).

Rooted in the prominent framework of normative moral philosophy, utilitarianism, and centered around a small but powerful core of principles—rationality and altruism—the EA movement

has itself become increasingly influential and even dominant in the AI risk philanthropy landscape. Among the reasons for its influence was a shift away from the original views of MIRI and FHI, which were more closely linked to the academic field of AI, and instead toward influence through industry and policymaking, including cultivating major financial backing to enable these goals. This ambition has been reinforced by the arrival of other major players in the field, to whom we turn to next.

To summarize, the early years of the AIERMPS are characterized by the emergence and convergence of several communities steeped in academic and social philosophies and subcultures surrounding rationalism, transhumanism, futurism, and later, evidence-based philanthropy. While the associated ideologies and cultures have evolved over time, many of the core principles have remained highly operative and even determinative in the AIERMPS.

2.2 A contesting power: the rise of AI industry in AI safety

The year 2015 marked a significant turning point in AI extreme risk mitigation philanthropy, with the entry of new philanthropic players less tied to the initial cultural epicenter of the AIERMPS, MIRI. One of the major drivers of these new resources was the increased interest of tech billionaire Elon Musk, which brought new funding and visibility to the ecosystem. Musk began his philanthropy in the AIERMPS by providing a $10 million gift to the newly founded Future of Life Institute (FLI), a non-profit organization co-founded by MIT professor Max Tegmark that is dedicated to mitigating risks arising from advanced AI systems (Kosoff, 2015). Additionally, Musk played a leading role in the founding of the (then entirely non-profit) organization OpenAI, donating $1 billion to the organization in conjunction with donations from Peter Thiel, Sam Altman, and other high-profile donors (Markoff, 2015). This influx of resources empirically dwarfed existing philanthropic efforts related to AI and AI safety.

The stated goal of OpenAI (now a subsidiary of Microsoft), like its already existing rival DeepMind, was to build AGI, which it defines as "highly autonomous systems that outperform humans at the most economically valuable work" and to ensure that it "benefits all of humanity" (OpenAI, 2018). Initially true to its name, OpenAI initially placed openness at the center of its philosophy, intending to open source its technology and even collaborate with competitors to foster beneficial rather than harmful AGI. However, it later changed its mind on its approach to openness, at least in part to growing concerns about risks arising from putting such a powerful technology in the hands of anyone, including malevolent actors.

Consequently, 2015 and 2016 marked the beginning of a power shift within the AIERMPS away from the MIRI-centered academic ecosystem toward a set of new actors from the effective altruism movement who held different views on how best to address AI safety and mitigate AI risks (Christiano, 2022; Karnofsky, 2012). Some prominent members of the former MIRI-centered ecosystem joined and led the core organizations of this group. Notably, this new landscape of actors operated under comparatively less pessimistic assumptions than did the original MIRI cluster. For instance, individuals in the community and their organizations were less pessimistic about the difficulty of making AI safe, believing that the fundamental technical problems were not as intractable. Accompanying this stance was their view that AI capabilities would evolve more gradually rather than suddenly, making incremental research and learning about effective governance more feasible. This set of positions would become the lynchpin of many subsequent disagreements in the field over the next decade about the viability of different strategies, organizations, and philanthropic priorities.

Embodied by three key individuals, research scientists Dario Amodei and Paul Christiano and philanthropist Holden Karnofsky, this new worldview increasingly shaped the trajectory of the

AIERMPS and the AI industry beginning in 2015 (Amodei et al., 2016; Karnofsky, 2012). The trajectory of these leading individuals also strengthened the AIERMPS's important connections with AI industry leaders (Coldewey, 2021; Piper, 2023). This became particularly important because the vast majority of AI-related research and development occurs in the private sector rather than the public sector and even within a handful of leading companies, making them the epicenter of AI's trajectory.

OpenAI was influenced by that worldview, at least for the first few years of its existence. This led them to prioritize and produce empirical research, which became one of the main drivers in bringing AI safety into the mainstream academia, with seminal research such as *Concrete Problems in AI Safety*, by Amodei, Christiano, and other leading researchers. With a focus on technical research, working closely with the AI industry to steer the trajectory of AI toward its definition of safety, and access to the best existing AI models and industry-leading resources, Christiano and Amodei have played an influential role in the nascent AGI industry. Both were present at OpenAI's founding dinner (Brockman, 2016), and each led the AI safety team at OpenAI at different points in time.

After 2020, they both left to start new institutions: Amodei left OpenAI to create a competing AI industry player, Anthropic (responsible for the Claude series of models), and Christiano left to create the Alignment Research Center, a research non-profit focused on research at the intersection of AI safety theory and deep learning AI systems. This center also later incubated an auditing organization, which aims to evaluate systems created by AI industry players. Both institutions became increasingly influential in the field of AI, starting in 2021, reinforcing the ties between parts of the AIERMPS and leading AI industry actors.

Importantly, the research and industry arms of the movement have been significantly fueled by its growing philanthropic arm. As CEO of Open Philanthropy, a major philanthropic foundation that gives tens to hundreds of millions of dollars annually to a wide range of causes, Holden Karnosfky developed an AI risk philanthropy program beginning in 2016. Under his leadership, Open Philanthropy quickly became the main grantmaker of the AIERMPS, cumulatively administering $330 million in grants for the field by 2023 (Open Philanthropy, 2023b).

This new funding enabled the field to grow substantially and in many directions: training young researchers, organizing conferences, providing fellowships and internships, supporting academic research, fostering independent researchers, sustaining organizations such as MIRI and FHI, and enabling the creation of a variety of new organizations and initiatives tied to the ecosystem. Indeed, it is difficult to overstate the importance of this large and sustained influx of funding, which has allowed Open Philanthropy to express its strategic preferences and ideology through numerous channels. Open Philanthropy's strategic views and influence on the AIERMPS also served to further connect this philanthropic community to the industry, allowing it to have an ongoing influence on the development of AI. Ultimately, these efforts, fostered over only a little more than a decade, positioned the AIERMPS to be at the forefront of AI when it exploded into the public eye in the 2020s.

The 2010s then represented an important period in the evolution of the AIERMPS. With important leaders in research, industry, and philanthropy reaching new levels of relevance, and armed with increased funding and the support from prominent academic experts and a growing pool of philanthropists, the AIERMPS was ready for the mainstream.

2.3 From niche to mainstream

After gaining an influential position in the laboratories and boardrooms of leading AI industry players, the AIERMPS became even more central thanks to a major shift in public and political

awareness of AI. Namely, the release of AI systems such as DALL-E 2, ChatGPT, Microsoft Bing, and Google Bard represented an important shift that first came to prominence in November 2022 with the public release of ChatGPT by OpenAI. Whereas AI systems had largely been discussed in specialized circles, the general public and policymakers have now become aware of the implications and risks of AI (Miyazaki et al., 2023).

For instance, the uptake of generative AI tools for academic misconduct and embedded in popular search platforms increased the salience of AI for students, parents, and workers alike. The public and decision-makers were increasingly exposed to high-profile AI-related failures involving data leaks, offensive content, misinformation, and bias, leading to organizational bans on large language models and increased regulatory attention. As an example of such an early focus event documented in leading publications, a New York Times article detailed a chatbot's attempt to persuade a journalist to leave his wife (Roose, 2023b). These and other incidents heightened awareness of AI-related concerns, though matched by the contemporaneous eagerness with which numerous organizations sought to adopt AI systems into their workflows.

This growing concern among experts and the general public was amplified by three events, all of which led to major news cycles and intensified debates over the course of several months:

- An open letter for an "AI pause" was signed by major actors, such as Elon Musk and Turing Prize winner Yoshua Bengio (Future of Life Institute, 2023);
- Another prominent AI expert and Turing Prize winner, Geoffrey Hinton, known as the "Godfather of AI," resigned from his position at Google, citing escalating concerns about AI risks as the core reason (Metz, 2023);
- A one-sentence statement from the Center for AI Safety, stating only that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war," was signed by many of the world's leading AI figures, including the CEOs of the top AI companies, hundreds of professors, and top AI researchers (Roose, 2023a).

This major shift in the landscape heralded and accompanied significant changes for the AIERMPS as a whole. First, the movement shifted a large portion of its focus to AI policy and governance, compared to its predominant focus on technical AI safety research. For example, major funders such as Open Philanthropy rapidly ramped up their grantmaking in this area (Open Philanthropy, 2023a). The sea change also led to a large influx of interested talent, stakeholder discussions, and public and policymaker attention, causing a surge of both opportunities and demands for the small number of experts in the field. Finally, a massive increase in investment in AI by leading technology companies contributed to a power shift from the non-profit and academic AI sector to the industrial sector. Due to the massive inequality of resources which allowed only the largest companies access to the data and infrastructure needed to develop AI models as well as the capacity to offer extremely high wages, industry players were far better positioned to attract top talent, develop leading AI models, and have access to information that enabled cutting-edge research (Mickle, 2023).

In response to this newly acquired publicity, a growing segment of individuals associated with the AIERMPS have endorsed strategies more reminiscent of traditional public-facing activism (Meaker, 2023). Their objective is to solicit public and policy attention to counterbalance the expanding influence of the AI industry, which continues to advance AI, despite arguably heightened risks. An "AI pause" is one of the core themes proposed and supported by these segments of the movement, with calls for a moratorium on AI development until developers of AI systems

can guarantee that their next AI systems will be developed and deployed safely. The contest between "accelerationist" and "decelerationist" philosophies is ongoing, and reflects the role of the AIERMPS at a moment in time when AI advances, governance, and public attention make these issues perhaps historically urgent and reflective of a contingency point.

In summary, this section discussed the history of the AIERMPS and its evolution since the early 2000s. From humble beginnings with a passionate and unique subculture based on shared principles and focused on theoretical work, it evolved into a sector with significant resources, important industry ties, and a more applied research and philanthropic approach, before rapidly reinventing itself in light of a massive increase in interest in AI risk mitigation. Familiarity with the ideological and cultural evolution of the AIERMPS is critical to understanding the ongoing challenges facing the movement and its possible trajectory, which we detail in the next section.

3 Challenges faced by the AI extreme risk mitigation philanthropic sector

3.1 A sector bound to disagree despite a strong willingness to agree

The AIERMPS faces several major difficulties in its efforts to mitigate AI risks. One such difficulty is that many catastrophic risks—and extinction risks in particular—are associated with events that are rare. Indeed, an extinction event is, by definition, unique and irreversible. As such, there may be few or no instances from which to learn lessons about mitigation successes or failures, limiting the possibility of feedback loops that could improve the movement's prospects for learning and improving its philanthropic strategies.

Even in the *absence* of such a catastrophic event, it is difficult to attribute this potential success to specific interventions or to even determine whether the interventions are causally responsible for mitigating these risks. Hence, philanthropists in the AIERMPS, who focus on evidence-based grantmaking, must rely on indirect indicators to measure whether a given funding initiative or strategy is effective. Such indicators often derive from and rely on strategic views about which trajectories are likely to increase or decrease the likelihood of catastrophic risks. Not unlike other social movements, then, differences in underlying philosophy can lead to wide disagreements among grantmakers, despite a shared culture centered on rationality.

Lightcone Infrastructure is an example of an organization with a central role in the AIERMPS that decided to shut down one of its core programs because of serious concerns about whether it had had positive or negative impacts, given the opportunities it had to help accelerate AI progress (Habryka, 2023). The AIERMPS frequently refers to this concern as "capabilities externalities," implying that advancing AI progress also entails the creation of externalities such as increased AI risks. Thus, whether to advance (accelerate) or alternatively limit (decelerate) AI progress is itself a recurring question, debated as grantmakers and strategic leaders assess the value of various research and funding directions.

A related complicating factor is that many of the strategies, often deemed necessary to ensure AI's safe development, rely critically on predictions of how actors will behave when these risks are elevated. This additional degree of contingency built into speculative scenarios leads many actors in the AIERMPS to disagree about whether leading AI industry actors such as OpenAI or Anthropic are beneficial or negative overall. One part of the community argues that these organizations are much more concerned about risk than other existing industry actors like Meta, owing to their strong tradition in AI safety. The other side argues that these organizations have ironically been the main culprits responsible for accelerating AI progress and AI risks, and are additionally skeptical that their efforts in technical safety offset this effect (Matthews, 2023).

This state of affairs constrains, for example, the ability of actors in the movement to robustly judge whether an industry actor is beneficial or harmful. Strategic planners may need to rely on information that may be absent, inherently available only in the future, or otherwise difficult to measure, and thus ripe for disagreement, such as:

- The likely technical trajectories of AI development, including what types of advances are needed for AI to become increasingly or decreasingly risky;
- The quality of the company's contributions to AI safety research;
- A judgment of the company's culture, including the extent to which employees care about and focus on mitigating AI risks;
- The true intentions of the company's CEO and leadership; and
- Whether the company or regulation is likely to be captured by industry motives.

This situation implies that many disagreements about grantmaking strategy are difficult to resolve, due to the lack of empirical evidence and clear metrics for success or failure. Contrast the challenges here with the environmental movement, for example, which has many metrics for measuring outcomes such as wildlife diversity, water quality, or carbon dioxide emissions.

In addition to these disagreements, another core aspect of the culture of the AIERMPS, which is also grounded in its focus on evidence-based philanthropy, presents a further challenge. Because of its origins in academic philosophy and its preference for economic methods that favor causal inference, the AIERMPS places significant emphasis on understanding counterfactual impact, including the opportunity costs of philanthropy. The concept is simple: if Alice receives a grant to pursue a line of research, the impact of the grant is not merely Alice's research output, but rather the difference in the advancement of that line of research *compared* to the situation in which Alice had not pursued this line of research. In such an alternative scenario, a grant could have gone to Bob instead, or Alice might have pursued another more or less promising research direction. As an example, if Alice is pursuing a research direction that ten other research teams are also pursuing, then the true impact of her research might be far more minor, given that if she hadn't pursued her research, other teams might have found her results anyway.

Core to the culture of effective altruism (Gabriel, 2017), this reasoning adds yet another layer of uncertainty to the evaluation of the impact of philanthropic interventions. An example widely discussed in the AIERMPS is that of Open Philanthropy, one of the core actors in the sector described previously, which granted $30 million to OpenAI. Those in the community who argue that this grant was harmful tend to say that it contributed substantially to the success of OpenAI overall, and thus helped OpenAI to accelerate AI progress writ large. Thus, by contributing to the success of its grantee, Open Philanthropy has accelerated AI progress, which, in the eyes of this community, unacceptably increases AI risks. In response to this criticism of Open Philanthropy, others presented a counterfactual: if Open Philanthropy had not given $30 million, OpenAI would not have failed, but would simply have raised funds slightly earlier or from other actors. In effect, this would have made little difference to the overall progress of AI, while Open Philanthropy's strategic giving might have increased its "say" in OpenAI's approach including an increased focus on AI safety (Moskovitz, 2023). This discussion illustrates how considering counterfactual impact, in addition to other measures of impact, can make it especially hard for actors in the AIERMPS to reach agreement on even core decisions.

Other considerations such as technical and geopolitical perspectives constitute areas where disagreement can affect grantmaking strategies pursued by various actors of the AIERMPS. One such persistent disagreement, another defining feature of the strategic decisions considered by the

AIERMPS, is the assessment of how difficult or easy it is to make powerful AI systems "safe." One perspective, with increasing following, is that AI safety research should *leverage* the development of large language models until they are sufficiently advanced to automate AI safety research itself (Karnofsky, 2022). As a consequence, this cluster of the community is comparably more enthusiastic about the safety strategies favored by AI industry players such as OpenAI or Anthropic. This contrasts with other segments of the AIERMPS who believe that AI safety is a much harder problem, that the use of large language models will not succeed and may backfire, and that the most desirable policy to pursue is to enforce a pause in advanced AI development until a technical solution to AI risks is in sight (Yudkowsky, 2023).

A first set of challenges faced by this community, then, relates to the numerous compounding difficulties in assessing the efficacy of their strategies. Determining an effective philanthropic strategy for AIERMPS's causes is no easy task. This is because there is little actual empirical evidence available, as the events in question are definitionally rare and because strategic decisions rely on forward-looking predictions which are contingent on uncertain technical advances and human behavior. Even such core questions as whether advancing or halting the progress of AI are more effective in risk mitigation remain controversial. While many social movements face disagreements, some of the factors faced by the AIERMPS are indeed unique.

3.2 A philanthropic sector that needs strong adaptation abilities

Another difficulty the AIERMPS had to face was that its originating community preceded most of the other key institutional actors that would later address AI risks, including those from academia. In practice, the leading actors had to determine what strategic directions were most likely to advance an otherwise non-existing field, and were thus limited in their ability to draw on surrounding infrastructure and perspective. This required, for example, making key direction-setting decisions such as funding individuals or teams in the absence of established signals of credibility, due to the absence of specialized curriculum, established university programs and credentials, or proven success in the field.

This led the AIERMPS to pursue a wide range of approaches, including creating a large infrastructure to enable a new research, policy-engaged, and philanthropic field. Activities have included funding online courses on AI risks (sometimes taken by students without background in the typical prerequisites), establishing forums and platforms dedicated to sharing AI safety research (using formats without the traditional checks and rigors of academic research, such as formal peer review), or funding non-profit organizations led by individuals from non-traditional academic paths to train cohorts of young researchers and build the community. One example is SERI MATS, a non-profit organization providing six-month training programs which pair candidates with experienced mentors to teach them AI safety research skills.

A well-known defining characteristic of AI since the 2010s is the pace at which the technology and the surrounding landscape evolve. This subsequently urges the AIERMPS to frequently change its grantmaking strategy and be continually forward-looking. For instance, while the technology underlying ChatGPT called *transformers* (Vaswani et al., 2023) was brand new and still a proof of concept in 2017, it is now widely deployed and considered one of the most powerful and standard technologies in AI. Similarly, the AI policy landscape has radically changed after the release of ChatGPT in November 2022.

For the AIERMPS, the only way to make sure that their giving can effectively address the dynamic challenges of AI development, governance, and risk mitigation at any given time is to remain up to date on the latest developments of the technology, of the fast-evolving landscape,

and to employ grantmakers with relevant technical and policy backgrounds. In practice and due to associated uncertainties, this can limit the capacity of grantmaking organizations to robustly evaluate a wide number of possible grant projects. Such a problem is particularly exacerbated by how difficult it is to hire and retain grantmakers with unusually high technical skills (e.g., advanced machine learning skills), another unique feature of the AIERMPS compared to other philanthropic movements. This dynamic may also make it more difficult for donors to engage in long-term giving commitments, as any specific research direction, policy solution, or funding initiative could be made obsolete by a new breakthrough.

In addition to the pace of AI, the field of AI safety research is notable for being at an extremely early stage, leading some to call the field "pre-paradigmatic" in that there is no strong agreement on the most crucial areas to work on (Hernandez-Orallo et al., 2020). On the one hand, grantmakers want their efforts to help mitigate risks relevant to the current state of the technology. This has driven much grantmaking on large language models, such as research on the explainability of the most advanced AI systems. On the other hand, the lack of clear plans for solving key technical safety problems also suggests a need to provide grants to help explore radically new approaches that may be largely speculative and even ineffective. Both are risky philanthropic approaches. Improper calibration toward "safer" or "riskier" grantmaking could undermine the core goals of the AIERMPS; striking an appropriate balance is similarly difficult. Informed grantmaking in this context hence requires grantmakers with a large range of technical (as well as social, political, and geopolitical) knowledge, in order to be able to evaluate numerous scientific areas. It also leads grantmakers to rely on external advisors and technical experts to properly evaluate proposals in highly narrow fields, introducing further complexity.

There are still remaining issues for philanthropic strategy when a subfield is very new. Some areas that are considered promising are pursued by no more than a few dozen individuals, which could severely undermine the movement's effectiveness if the donor strategy turns out to be less than optimal. This risk applies, for example, to a highly specialized branch of AI safety research called *Infrabayesianism* (Kosoy, 2020) or Open Agency Architecture (Dalrymple, 2024). Compounding this challenge, limited early investment or the inability to even initiate work in potentially promising areas can preclude the ability of funders to understand whether these trajectories are promising.

In combination, the unique cultural and ideological characteristics of the AIERMPS, its novelty, the high uncertainty surrounding key factors that affect the impact of grants, the pace of development of AI, and several other key factors mean the AIERMPS has had to face a significant number of challenges in its philanthropic strategy. This has led it to foster significant adaptability and flexibility to react to dynamic developments as they arise. Beyond its idiosyncratic challenges, the AIERMPS has also had quite unusual relations with other actors in the AI field, in contrast to how other social movements have evolved. We discuss these unique relationships and the associated opportunities and limitations they presented in the next section.

4 What the AI extreme risk mitigation philanthropic sector could learn from other fields and social movements

4.1 Commonalities and differences with other social movements

Early on, the AIERMPS framed AI risk mitigation primarily in terms of technical problems that researchers needed to solve, despite its understanding of AI risk as, at its core, a societal issue that

could affect everyone. As Ó hÉigeartaigh summarized in his 2017 overview of the field, much technical AI research

> has focused on translating some of the more foundational questions raised by early work at FHI and MIRI and elsewhere into crisp technical research problems that can be worked on today. This includes approaches involving fundamental mathematical frameworks for agent decision-making and behavior, as well as research programs exploring how some of the behaviors that would be of concern in long-term systems may manifest in the near-term systems we are building currently.

This technical view of the issue might also be partially driven in part by the "tendency to valorize corporate-driven tech solutions" that Broad (2018) notes regarding the effective altruism movement, as well as the movement's general origins in elite and intellectual research communities rather than activist or advocacy circles.

This focus contrasts with other social movements such as climate change or animal welfare, which were instead framed early on as predominantly social problems (McCright & Dunlap, 2000; Singer, 1975), affecting the set of interventions and approaches that are considered viable to solve these problems. This focus also impacts which kinds of actors are deemed relevant (or irrelevant) in shaping the movement's aims. If progress requires advanced computer science research rather than, say, cutting down on one's consumption of meat or recycling in the workplace or public protest, the space for public engagement is de facto diminished. Moreover, the types of concerns raised were not historically promoted to the mass public in the way that other social movements were, meaning that AI risk remained an issue of low public salience until it exploded into public attention.

Thus, despite the continued rise in interest since the late 2010s of the AIERMPS in the social aspects of AI risks, such as how to govern AI (Ó hÉigeartaigh, 2017), the movement has not yet seen the development of a social movement comparable in size or magnitude to the climate change or animal welfare movements. The largest existing AI safety social movement to date is still very small, with overwhelming participation from experts rather than the general public (Meaker, 2023), and little evidence of penetration into the general public's consciousness.

This technical coloration of the movement prevents the AIERMPS from using strategies that other social movements have used to achieve comparable goals. While a comprehensive review of relevant strategies, commonalities, and differences is beyond the scope of this chapter, some examples help to illustrate the point. For instance, like the climate change movement and the animal welfare movement, the AIERMPS faces a problem in the 2020s that essentially surrounds the behavior of a few corporations. Hence, the AIERMPS could likely learn from how other movements have approached these dynamics.

In *Ethics Into Action*, written in 1998, Singer shows how Henry Spira, one of the pioneers of the animal welfare movement, achieved significant social change among corporations despite having few resources by using both (external) adversarial and (internal) cooperative strategies. He used these strategies to respectively acquire bargaining power and to use it through interactions with the corporations to achieve concrete outcomes. By mixing concrete threats to the public relations of organizations that mistreated animals, such as *McDonald's* or *Revlon*, with frequent interactions with employees from these companies, Spira was able to increase reputational pressure on these firms through targeted ads against the companies and demonstrations until they made the concrete changes Spira demanded, ultimately improving animal welfare.

Another example of how the lack of a strong social movement in AI safety limits the field's ability to achieve social change is its inability to tap one of the core mechanisms of the animal welfare and climate change movements known as the radical flank effect (Evans, 2023; Lange, 1990; Simpson et al., 2022; Singer, 1998). The radical flank effect describes the effects that radical activists have in increasing the likelihood of counterparties to negotiate alternatively with more moderate activists, or by changing the public issue framing around which strategies are considered moderate or radical. Critically, this effect has been found to have both negative and positive implications and is still much debated in the academic community. However, it arguably reflects another element missing from the AERMPS movement.

Still, other ways in which the AIERMPS's unique composition and history limit its effectiveness may revolve around voting, lobbying, fundraising, and overall resilience to threats. Limiting fundraising to a small number of wealthy, dedicated donors renders movements dependent on those donors and the risks they pose, as evidenced by the Sam Bankman-Fried scandal (Kim, 2022), while alternatively cultivating a broader base of funding support might increase resilience. Similarly, without broad-based public support, individuals are unlikely to engage in protests, walkouts, create civic clubs or student groups, call their political representatives, and so on.

A shorthand way of making this critique is that the movement remains centered on elite and esoteric perspectives aimed at rationally identifying optimal strategies. However, the animal welfare, environmental, civil rights, and gay rights movements have arguably not achieved such levels of success through rational efficacy alone. Additional exploration and serious incorporation of multidisciplinary perspectives could help the AIERMPS learn from these insights.

4.2 Contingent coalitions with natural enemies and contingent tensions with natural allies

On top of those distinctive aspects, the AIERMPS has pursued unusual coalition-building strategies to advance its goals. One of these unusual characteristics is that the social networks of those who are developing (potentially harmful) AI systems and those who are trying to *prevent* those harms are closely tied together (Lazar & Nelson, 2023). By analogy, imagine dedicated environmental activists and oil executives working as close friends (Alexander, 2022). As discussed previously, many of the most successful AI companies are closely tied in multiple ways to prominent members of the philanthropic sector and, more broadly, to the rationalist community. While there may be benefits to such ties, for instance in terms of policy learning, they may also contribute to limiting the ability of core organizations of the AIERMPS to fund interventions that would be seen as too hostile to AI industry leaders. The potential for conflicts of interest is real.

Nevertheless, this link has allowed the AIERMPS to form highly atypical coalitions that have arguably had a major impact on public discourse. Today, several of the leading AI organizations are managed by figures in contact with the movement, and many have prominent mission statements that explicitly call for the development of responsible or safe AI, with dedicated teams focused on AI safety. As an example, the prominent AI extinction risk statement produced by the Center for AI Safety generated a major news cycle in large parts thanks to the signatures from essentially all of the CEOs of leading AI firms (Roose, 2023a). This suggests that the AIERMPS, despite its nature as an elite and technical community, has fostered a growing ability to build unusual but powerful coalitions capable of shaping policy and public discourse toward their perspective on AI risks.

This can be better understood in light of Van Dyke and Amos (2017) who explain which factors are crucial to coalition-building. Among these factors, the AIERMPS has cultivated:

- Strong social ties with prominent "bridge builders," such as the Future of Life Institute, which has organized major conferences with a broad range of actors (Ó hÉigeartaigh, 2017);
- A shared ideology and culture across many organizations (Chivers, 2019);
- Increased political opportunities due to the rise in interest in the issue, which, according to political opportunity theory, raises the chances of successful coalitions; and
- A significant amount of philanthropic resources available to be deployed, allowing organizations in the movement to dedicate some resources to coalition-type activities.

Paradoxically, while the AIERMPS has maintained unusually strong ties with industry actors, it has largely failed to form coalitions with the AI ethics communities, whose interests and goals arguably make them much more natural allies than industry players. Some cultural and ideological differences have turned into conflicts that may have calcified actors and made coalition-building harder.

The original seeds of the disagreement, as Prunkl and Whittlestone (2020) explain, arose from disagreements over which issues were more important when attempting to minimize the harms (or risks) from AI. To simplify, those closer to the "AI ethics" side argue that AI safety concerns are overblown, even pseudoscientific, and a distraction from what they view as more pressing problems, such as bias or privacy concerns arising from already existing AI systems. In contrast, some closer to the "AI safety" perspective suggest that the problems they focus on are catastrophic or existential in magnitude, and dwarf so-called "short-term" AI ethics concerns in importance.

This disagreement became cemented by episodes where lack of mutual support and contestation over public attention and funding have created resentment. Two recent examples illustrate this conflict:

- Some prominent voices in the AI ethics community blamed Hinton for leaving Google in 2023, due to his concerns about AI safety, when he had not similarly reacted two years earlier when prominent AI ethics advocate Timnit Gebru was pushed to resign from Google after having been asked to not publish one of her papers related to the problems of large language models (Chan, 2023).
- Despite having founded the Distributed Artificial Intelligence Research Institute (DAIR), an organization focused on preventing harms of AI, and aiming to launch a "Slow AI" movement in 2022 (Strickland, 2022), she along with other prominent AI ethics researchers severely criticized statements on the AI pause proposed by an organization from the AI safety movement, due to their emphasis on AI extinction risks (Sætra & Danaher, 2023).

To resolve these disagreements and conflicts and create room for greater impact overall, various proposals have been offered. Stix and Maas (2021) emphasize many avenues for positive collaborations, such as the study of available policy levers that would help achieve changes that both camps find amenable, or jointly pushing for mechanisms to maintain the integrity of public discourse in the face of AI systems. Prunkl and Whittlestone (2020) emphasize how the division between "long term" (AI safety) and "short term" (AI ethics) risks overemphasizing these differences, including their associated time scales. Instead, they propose four dimensions that could better identify disagreements in prioritization and thus foster the potential for collaboration.

More systematic discourse and shared analysis could help the communities identify common ground on which AI capabilities to focus on, when to focus on current or future impacts, whether to focus on more or less uncertain issues, and whether to focus on extreme risks or risks at all scales. As Sætra and Danaher (2023) identify, the movements could endeavor to build bridges to achieve common goals and avoid a situation where "neither short- nor long-term risks are managed and mitigated," which would represent a failure of both communities. For the AIERMPS to achieve its goals effectively then, it may need to continually revisit both its unusual alliances with industry actors and its disagreements with actors in adjacent communities.

5 Conclusion

This chapter began with a review of the status of AI and philanthropy, articulating a distinction between AI when used as a tool to advance numerous aspects of philanthropic practice, and AI when considered as a domain or cause area. Here, we focus on the latter, presenting a history and evaluation of the prominent and increasingly important community focused on extreme AI risks.

We reviewed its unusual intellectual and philosophical origins in rationalism, effective altruism, and technical safety research before discussing its transition to public relevance. Some of the movement's unique features may have played a role in its recent successes, such as close alliances between leading AI industry actors and AI safety researchers. However, the movement's largely elite nature and difference from typical broad-based social movements also pose limitations and threats to its viability, including distancing it from potentially natural allies.

This chapter only begins to articulate some important characteristics of the movement. Substantial research, including historical analysis, interviews, studies in management and political science, and so on, is needed to unpack many related issues, understand possible trajectories, and provide analysis to evaluate, achieve, and perhaps modify the movement's aims and efficacy. As a starting point, we suggest greater research is needed to understand the social and intellectual history, political coalitions, and trade-offs involved with the movement, as well as the movement's positionality in broader philanthropic and AI circles.

For individuals who are members or observers of the movement, this chapter echoes calls for multidisciplinary engagement, learning from outside perspectives, and learning from the successes and failures of other social movements. We suggest here, echoing numerous commentors, that direct engagement with other social movements, critical scrutiny of current alliances, and efforts to build bridges across coalitions could be prudent, along with deeper engagement with the general public. While the impact of the AI extreme risk philanthropic sector is yet to be fully understood, it is likely to be monumental.

References

Alexander, S. (2022, August 8). Why not slow AI progress? *Astral Codex Ten*. https://www.astralcodexten.com/p/why-not-slow-ai-progress

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

Attard-Frost, B., De los Ríos, A., & Walters, D. R. (2022). The ethics of AI business practices: A review of 47 AI ethics guidelines. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00156-6

Baum, S. D. (2018). Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI & SOCIETY*, *33*(4), 565–572. https://doi.org/10.1007/s00146-017-0734-3

Benjamins, R. (2020, May 22). A new organizational role for Artificial Intelligence: The responsible AI champion. *Think Big*. https://business.blogthinkbig.com/a-new-organizational-role-for-artificial-intelligence-the-responsible-ai-champion/

Birnbacher, D. (2009). Posthumanity, transhumanism and human nature. In B. Gordijn & R. Chadwick (Eds.), *Medical Enhancement and Posthumanity* (pp. 95–106). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8852-0_7

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies* (First edition). Oxford University Press.

Braverman, M. T., Constantine, N. A., & Slater, J. K. (2004). *Foundations and Evaluation: Contexts and Practices for Effective Philanthropy*. John Wiley & Sons.

Broad, G. M. (2018). Effective animal advocacy: Effective altruism, the social economy, and the animal protection movement. *Agriculture and Human Values*, *35*(4), 777–789. https://doi.org/10.1007/s10460-018-9873-5

Brockman, G. (2016). *My Path to OpenAI*. https://blog.gregbrockman.com/my-path-to-openai

Byrnes, S. (2021, December 14). Consequentialism & corrigibility. *AI Alignment Forum*. https://www.alignmentforum.org/posts/KDMLJEXTWtkZWheXt/consequentialism-and-corrigibility

Cath, C., Latonero, M., Marda, V., & Pakzad, R. (2020). Leap of FATE: Human rights as a complementary framework for AI policy and practice. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 702–702. https://doi.org/10.1145/3351095.3375665

Chan, W. (2023, May 5). *'I Didn't See Him Show Up': Ex-Googlers Blast 'AI Godfather' Geoffrey Hinton's Silence on Fired AI Experts*. Fast Company.

Chivers, T. (2019). *The AI Does Not Hate You: The Rationalists and Their Quest to Save the World*. Weidenfeld & Nicolson.

Christiano, P. (2022, June 19). Where I agree and disagree with Eliezer. *LessWrong*. https://www.lesswrong.com/posts/CoZhXrhpQxpy9xw9y/where-i-agree-and-disagree-with-eliezer

Chui, M., Harrysson, M., Manyika, J., Roberts, R., Chung, R., Nel, P., & Heteren, A. van. (2018). *Applying AI for Social Good*. McKinsey Global Institute. https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press. ISBN: 9780262357074

Coldewey, D. (2021). Anthropic is the new research outfit from OpenAI's Dario Amodei. *Yahoo News*. https://www.yahoo.com/now/anthropic-ai-research-outfit-openais-175923024.html

Cowls, J., King, T., Taddeo, M., & Floridi, L. (2019). *Designing AI for Social Good: Seven Essential Factors* (SSRN Scholarly Paper ID 3388669). Social Science Research Network. https://doi.org/10.2139/ssrn.3388669

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). A definition, benchmark and database of AI for social good initiatives. *Nature Machine Intelligence*, *3*(2), Article 2. https://doi.org/10.1038/s42256-021-00296-0

Dalrymple, D. (2024). Safeguarded AI: Constructing safety by design. *Aria*. https://www.aria.org.uk/wp-content/uploads/2024/01/ARIA-Safeguarded-AI-Programme-Thesis-V1.pdf

de Souza Leão, L., & Eyal, G. (2019). The rise of randomized controlled trials (RCTs) in international development in historical perspective. *Theory and Society*, *48*(3), 383–418. https://doi.org/10.1007/s11186-019-09352-6

Eiland, J., Hammonds, C. M., Ponos, S. M., Weigand, S. M., & Scherer, W. T. (2021). Developing models to predict giving behavior of nonprofit donors. *2021 Systems and Information Engineering Design Symposium (SIEDS)*, 1–6. https://doi.org/10.1109/SIEDS52267.2021.9483771

Evans, E. M. (2023). Animal advocacy and the "good cop-bad cop" radical flanking of laboratory research. *Sociological Inquiry*, *93*(3), 662–686. https://doi.org/10.1111/soin.12521

Fiennes, C. (2017). We need a science of philanthropy. *Nature*, *546*(7657), Article 7657. https://doi.org/10.1038/546187a

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Future of Life Institute (2023). Pause giant AI experiments: An open letter. *Future of Life Institute*. https://futureoflife.org/open-letter/pause-giant-ai-experiments

Gabriel, I. (2017). Effective altruism and its critics. *Journal of Applied Philosophy*, *34*(4), 457–473. https://doi.org/10.1111/japp.12176

Gore, A. (1993). *Creating a Government That Works Better & Costs Less: The Report of the National Performance Review*. The Review.

Guo, Y., Dong, Y., Wei, X., & Dong, Y. (2023). Effects of continuous adoption of artificial intelligence technology on the behavior of holders' farmland quality protection: The role of social norms and green cognition. *Sustainability*, *15*(14), Article 14. https://doi.org/10.3390/su151410760

Habryka, O. (2023). Shutting down the Lightcone offices. *LessWrong*. https://www.lesswrong.com/posts/psYNRb3JCncQBjd4v/shutting-down-the-lightcone-offices

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An Overview of Catastrophic AI Risks* (arXiv: 2306.12001). arXiv. https://doi.org/10.48550/arXiv.2306.12001

Henriksen, A., & Blond, L. (2023). Executive-centered AI? Designing predictive systems for the public sector. *Social Studies of Science*, 03063127231163756. https://doi.org/10.1177/03063127231163756

Hernandez-Orallo, J., Martınez-Plumed, F., Avin, S., & Whittlestone, J. (2020). AI paradigms and AI safety: Mapping artefacts and techniques to safety issues. In *European Conference on Artificial Intelligence* (2020). Santiago de Compostela, Spain. https://ecai2020.eu/papers/1364_paper.pdf

Herzog, P. S., Naik, H. R., & Khan, H. A. (2021). *AIMS Philanthropy Project: Studying AI, Machine Learning & Data Science Technology for Good*. Indiana University Lilly Family School of Philanthropy and Indiana University School of Informatics and Computing, IUPUI, Indianapolis. https://hdl.handle.net/1805/25177

Holzmeyer, C. (2021). Beyond `AI for Social Good' (AI4SG): Social transformations-not tech-fixes-for health equity. In *Interdisciplinary Science Reviews* (Vol. 46, Issues 1–2, SI, pp. 94–125). Routledge Journals, Taylor & Francis. https://doi.org/10.1080/03080188.2020.1840221

Howard, A., Borenstein, J., & Gosha, K. (2019). *NSF-Funded Fairness, Ethics, Accountability, and Transparency (FEAT) Workshop Report* (10139705). Georgia Institute of Technology. https://par.nsf.gov/biblio/10139705

Isabelle, D. A., & Westerlund, M. (2022). A review and categorization of artificial intelligence-based opportunities in wildlife, ocean and land conservation. *Sustainability*, *14*(4), Article 4. https://doi.org/10.3390/su14041979

Isbanner, S., O'Shaughnessy, P., Steel, D., Wilcock, S., & Carter, S. (2022). The adoption of artificial intelligence in health care and social services in Australia: Findings from a methodologically innovative national survey of values and attitudes (the AVA-AI study). In *Journal of Medical Internet Research* (Vol. 24, Issue 8). JMIR Publications, Inc. https://doi.org/10.2196/37611

Johnson, P. D. (2018). *Global Philanthropy Report: Perspectives on the Global Foundation Sector*. https://policycommons.net/artifacts/1847356/global-philanthropy-report/2593720/

Kaplan, S. A., & Garrett, K. E. (2005). The use of logic models by community-based initiatives. *Evaluation and Program Planning*, *28*(2), 167–172. https://doi.org/10.1016/j.evalprogplan.2004.09.002

Karnofsky, H. (2012). Thoughts on the Singularity Institute (SI). *LessWrong*. https://www.lesswrong.com/posts/6SGqkCgHuNr7d4yJm/thoughts-on-the-singularity-institute-si

Karnofsky, H. (2022). How might we align transformative AI if it's developed very soon? *AI Alignment Forum*. https://www.alignmentforum.org/posts/rCJQAkPTEypGjSJ8X/how-might-we-align-transformative-ai-if-it-s-developed-very

Key, J. (2001). Enhancing fundraising success with custom data modelling. *International Journal of Nonprofit and Voluntary Sector Marketing*, *6*(4), 335–346. https://doi.org/10.1002/nvsm.159

Kim, W. (2022, November 15). Sam Bankman-Fried's arrest is the culmination of an epic flameout. *Vox*. https://www.vox.com/the-goods/23458837/sam-bankman-fried-ftx-sbf-downfall-explained

Kosoff, M. (2015). Elon Musk donates $10 million to the future of life institute. *Business Insider*. https://www.businessinsider.com/elon-musk-donates-10-million-to-the-future-of-life-institute-2015-1

Kosoy, V. (2020). Infra-Bayesianism. *AI Alignment Forum*. https://www.alignmentforum.org/s/CmrW8fCmSLK7E25sa

Lange, J. I. (1990). Refusal to compromise: The case of earth first! *Western Journal of Speech Communication*, *54*(4), 473–494. https://doi.org/10.1080/10570319009374356

Lazar, S., & Nelson, A. (2023). AI safety on whose terms? *Science*, *381*(6654), 138–138. https://doi.org/10.1126/science.adi8982

Maas, M. (2023). International AI institutions: A literature review of models, examples, and proposals. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4579773

Madeo, E. (2022). Fundraising in the higher education context: A topical and theoretical literature review. *Education Society and Human Studies*, *3*, 1–22. https://doi.org/10.22158/eshs.v3n2p1

Markoff, J. (2015). Artificial-intelligence research center is founded by Silicon Valley investors. *The New York Times*. https://www.nytimes.com/2015/12/12/science/artificial-intelligence-research-center-is-founded-by-silicon-valley-investors.html

Matthews, D. (2023, July 17). The $1 billion gamble to ensure AI doesn't destroy humanity. *Vox*. https://www.vox.com/future-perfect/23794855/anthropic-ai-openai-claude-2

Mazzi, F., Taddeo, M., & Floridi, L. (2023). AI in support of the SDGs: Six recurring challenges and related opportunities identified through use cases. In F. Mazzi & L. Floridi (Eds.), *The Ethics of Artificial Intelligence for the Sustainable Development Goals* (pp. 9–33). Springer International Publishing. https://doi.org/10.1007/978-3-031-21147-8_2

McCright, A. M., & Dunlap, R. E. (2000). Challenging global warming as a social problem: An analysis of the conservative movement's counter-claims. *Social Problems*, *47*(4), 499–522. https://doi.org/10.2307/3097132

Meaker, M. (2023). Meet pause AI, the protest group campaigning against human extinction. *Wired UK*. https://www.wired.co.uk/article/pause-ai-existential-risk

Metz, C. (2022). *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World*. Penguin Publishing Group.

Metz, C. (2023, May 1). 'The godfather of A.I.' leaves Google and warns of danger ahead. *The New York Times*. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

Mickle, T. (2023). Big Tech rebounds and preps for transformative A.I. investments. *The New York Times*. https://www.nytimes.com/2023/08/05/technology/tech-nvidia-chips.html

Mittal, P., & Srivastava, V. K. (2021). A review of supervised machine learning algorithms to classify donors for charity. *International Journal of Advanced Research in Computer Science*. https://ijarcs.info/index.php/Ijarcs/article/view/6685/5388

Miyazaki, K., Murayama, T., Uchiba, T., An, J., & Kwak, H. (2023). Public perception of generative AI on Twitter: An empirical study based on occupation and usage (arXiv: 2305.09537). *arXiv*. http://arxiv.org/abs/2305.09537

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine*, *260*, 113172. https://doi.org/10.1016/j.socscimed.2020.113172

Moskovitz, D. (2023). What happened to the OpenPhil OpenAI board seat? *EA Forum*. https://forum.effectivealtruism.org/posts/CmZhcEpz7zBTGhksf/what-happened-to-the-openphil-openai-board-seat

Nast, C. (2015, November 16). The doomsday invention. *The New Yorker*. https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom

Ngo, R., Chan, L., & Mindermann, S. (2023). The alignment problem from a deep learning perspective. https://doi.org/10.48550/arXiv.2209.00626

Noordt, C., & Tangi, L. (2023). The dynamics of AI capability and its influence on public value creation of AI within public administration. *Government Information Quarterly*, 101860. https://doi.org/10.1016/j.giq.2023.101860

Ó hÉigeartaigh, S. (2017). *The State of Research in Existential Risk* (SSRN Scholarly Paper 3446663). https://papers.ssrn.com/abstract=3446663

OpenAI. (2018, April 9). OpenAI charter. *OpenAI*. https://openai.com/charter

Open Philanthropy (2023a). New roles on our global catastrophic risks team. *Open Philanthropy*. https://www.openphilanthropy.org/research/new-roles-on-our-gcr-team/#4-ai-governance-and-policy-aigp

Open Philanthropy. (2023b). Potential risks from advanced artificial intelligence. *Open Philanthropy*. https://www.openphilanthropy.org/focus/potential-risks-advanced-ai/

Osborne, D. (1993). Reinventing government. *Public Productivity & Management Review*, *16*(4), 349–356. https://doi.org/10.2307/3381012

Pawson, R. (2002). Evidence-based policy: In search of a method. *Evaluation*, *8*(2), 157–181. https://doi.org/10.1177/1358902002008002512

Perry, B., & Uuk, R. (2019). AI governance and the policymaking process: Key considerations for reducing AI risk. *Big Data and Cognitive Computing*, *3*(2), Article 2. https://doi.org/10.3390/bdcc3020026

Piper, K. (2023). Can society adjust at the speed of artificial intelligence? *Vox*. https://www.vox.com/future-perfect/2023/3/18/23645013/openai-gpt4-holden-karnofsky-artificial-intelligence-ai-safety-existential-risk

PricewaterhouseCoopers (2017). What doctor? Why AI and robotics will define New Health. *PricewaterhouseCoopers*. https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/ai-robotics-new-health.pdf

Prunkl, C., & Whittlestone, J. (2020). Beyond near- and long-term: Towards a clearer account of research priorities in AI ethics and society. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 138–143. https://doi.org/10.1145/3375627.3375803

Ramirez, A., & Saraoglu, H. (2009). *An Analytic Approach to Selecting a Nonprofit* (SSRN Scholarly Paper 1488870). https://doi.org/10.2139/ssrn.1488870

Roose, K. (2023a). AI poses 'risk of extinction,' industry leaders warn. *The New York Times*. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

Roose, K. (2023b). Why a conversation with Bing's Chatbot left me deeply unsettled. *The New York Times*. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

Sætra, H. S., & Danaher, J. (2023). Resolving the battle of short- vs. long-term AI risks. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00336-y

Schiff, D., Ayesh, A., Musikanski, L., & Havens, J. C. (2020). IEEE 7010: A new standard for assessing the well-being implications of artificial intelligence. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2746–2753. https://doi.org/10.1109/SMC42975.2020.9283454

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2021). Explaining the principles to practices gap in AI. *IEEE Technology and Society Magazine*, *40*(2), 81–94. https://doi.org/10.1109/MTS.2021.3056286

Schiff, D. S. (2023). Looking through a policy window with tinted glasses: Setting the agenda for U.S. AI policy. *Review of Policy Research*, *40*(5), 729–756. https://doi.org/10.1111/ropr.12535

Schiff, D. S., Laas, K., Biddle, J. B., & Borenstein, J. (2022). Global AI ethics documents: What they reveal about motivations, practices, and policies. In K. Laas, M. Davis, & E. Hildt (Eds.), *Codes of Ethics and Ethical Guidelines: Emerging Technologies, Changing Fields* (pp. 121–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-86201-5_7

Schwartz, D., Selman, J. M. G., Wrege, P., & Paepcke, A. (2021). Deployment of embedded edge-AI for wildlife monitoring in remote regions. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1035–1042. https://doi.org/10.1109/ICMLA52953.2021.00170

Shapiro, D., & Cody, S. (2015). Data quality to further philanthropy's mission. *Mathematica Policy Research Reports*. https://www.mathematica.org/publications/data-quality-to-further-philanthropys-mission

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm (arXiv: 1712.01815). *arXiv*. https://doi.org/10.48550/arXiv.1712.01815

Simpson, B., Willer, R., & Feinberg, M. (2022). Radical flanks of social movements can increase support for moderate factions. *PNAS Nexus*, *1*(3), 110. https://doi.org/10.1093/pnasnexus/pgac110

Singer, P. (1975). *Animal Liberation: A New Ethics for Our Treatment of Animals*. Eweb: 10461. https://repository.library.georgetown.edu/handle/10822/769929

Singer, P. (1998). *Ethics into Action: Henry Spira and the Animal Rights Movement*. Rowman & Littlefield.

Singer, P. (2019). *The Life You Can Save: How to Do Your Part to End World Poverty* (10th Anniversary ed. edition). www.thelifeyoucansave.org.

Stahl, B. C., Schroeder, D., & Rodrigues, R. (2023). AI for good and the SDGs. In B. C. Stahl, D. Schroeder, & R. Rodrigues (Eds.), *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges* (pp. 95–106). Springer International Publishing. https://doi.org/10.1007/978-3-031-17040-9_8

Stern, K. (2013). *With Charity For All: Why Charities Are Failing and a Better Way to Give*. Knopf Doubleday Publishing Group.

Stix, C., & Maas, M. M. (2021). *Bridging the Gap: The Case for an 'Incompletely Theorized Agreement' on AI Policy by Charlotte Stix, Matthijs M. Maas: SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3756437

Strickland, E. (2022). Timnit Gebru is building a slow AI movement. *IEEE Spectrum*. https://spectrum.ieee.org/timnit-gebru-dair-ai-ethics

Sulaeman, D. (2018). Smart charities? Analyses of IT-enabled charitable fundraising. *PACIS 2018 Proceedings*, 340.

Taillandier, A. (2021). "Staring into the singularity" and other posthuman tales: Transhumanist stories of future change. *History and Theory*, *60*(2), 215–233. https://doi.org/10.1111/hith.12203

Van Dyke, N., & Amos, B. (2017). Social movement coalitions: Formation, longevity, and success. *Sociology Compass*, *11*(7), e12489. https://doi.org/10.1111/soc4.12489

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762

Vequist, D. (2014). *Nonprofit Fundraising Transformation Through Analytics* (pp. 116–125). https://doi.org/10.4018/978-1-4666-7272-7.ch009

Voida, A. (2014). A case for philanthropic informatics. In S. Saeed (Ed.), *User-Centric Technology Design for Nonprofit and Civic Engagements* (pp. 3–13). Springer International Publishing. https://doi.org/10.1007/978-3-319-05963-1_1

Wakunuma, K., Ogoh, G., Eke, D., & Akintoye, S. (2022, January 1). Responsible AI, SDGs, and AI Governance in Africa. *2022 IST-Africa Conference (IST-Africa)*, 1–13. https://doi.org/10.23919/IST-Africa56635.2022.9845598.

White, H. (2019). The twenty-first century experimenting society: The four waves of the evidence revolution. *Palgrave Communications*, *5*(1), Article 1. https://doi.org/10.1057/s41599-019-0253-6

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration*, *42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Yudkowsky, E. (2013). *Intelligence Explosion Microeconomics*. Machine Intelligence Research Institute.

Yudkowsky, E. (2016). *The AI Alignment Problem: Why It's Hard, and Where to Start*. Machine Intelligence Research Institute.

Yudkowsky, E. (2023). The only way to deal with the threat from AI? Shut it down. *Time*. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3), 101577. https://doi.org/10.1016/j.giq.2021.101577