# Building the Epistemic Community of AI Safety

Shazeda Ahmed
shazeda@g.ucla.edu
Center on Race and Digital Justice,
University of California - Los Angeles
California, USA

Klaudia Jazwinska
klaudia@princeton.edu
Center for Information Technology
Policy, Princeton University
New Jersey, USA

Archana Ahlawat
archana.ahlawat@princeton.edu
Center for Information Technology
Policy, Princeton University
New Jersey, USA

Amy Winecoff
aw0934@princeton.edu
Center for Information Technology
Policy, Princeton University
New Jersey, USA

Mona Wang
monaw@princeton.edu
Center for Information Technology
Policy, Princeton University
New Jersey, USA

## ABSTRACT

The emerging field of "AI safety" has attracted public attention and large infusions of capital to support its implied promise: the ability to deploy advanced artificial intelligence (AI) while reducing its gravest risks. Ideas from effective altruism, longtermism, and the study of existential risk are foundational to this new field. In this paper, we contend that overlapping communities interested in these ideas have merged into what we refer to as the broader "AI safety epistemic community," which is sustained through its mutually reinforcing community-building and knowledge production practices. We support this assertion through an analysis of four core sites in this community's epistemic culture: 1) online community-building through web forums and career advising; 2) AI forecasting; 3) AI safety research; and 4) prize competitions. The dispersal of this epistemic community's members throughout the tech industry, academia, and policy organizations ensures their continued input into global discourse about AI. Understanding the epistemic culture that fuses their moral convictions and knowledge claims is crucial to evaluating these claims, which are gaining influence in critical, rapidly changing debates about the harms of AI and how to mitigate them.

## KEYWORDS

AI safety, epistemic culture, existential risk, effective altruism

## 1 INTRODUCTION

Imagine you are an undergraduate computer science student at a US research university interested in the ethical consequences of the technology you are learning to build. Seeking a like-minded community, you join a student organization where you read books like *Superintelligence*, and find online forums debating how artificial intelligence (AI) will shape the future of humanity. Motivated by these communities' discussions about how to do the most good in the world, you decide to pursue a career where you work towards addressing risks from AI. You join a tech company where you build large language models (LLMs). In your spare time, you read research papers posted to these communities' web forums on how to make LLMs safer. Suddenly, you realize the community that has informed major decisions in your personal and professional life is increasingly shaping how the technology industry, academia, media, and policymakers think about AI.

This hypothetical scenario approximates a very real personal and professional path for individuals interested in minimizing what they view as the negative long-term consequences of AI—especially those they characterize as existential threats to humanity. Starting in the early 2000s, a robust community has arisen around these issues, attracting individuals interested in applying the interconnected ideas behind effective altruism (EA), longtermism, artificial general intelligence (AGI), and existential risk ("x-risk") to making AI systems safer.

Importantly, these ideas have recently entered the mainstream. In 2022, this shift was propelled in part by the large-scale infusion of capital then-billionaire Sam Bankman-Fried committed to EA and longtermist causes through FTX Foundation's Future Fund, a grant-making body which was associated with his cryptocurrency exchange's philanthropic arm [67]. Many of the organizations, research, media, individuals, and projects selected for FTX Future Fund grants strengthened and expanded the EA and longtermist communities and their influence on how broad swaths of people outside of the community think about AI. In under a year, these ideas have come to take on global significance: discourse about AI posing an existential risk regularly appears in news media coverage and has spurred policymakers on both sides of the Atlantic to turn to this epistemic community for solutions. While the Future Fund dissolved [148] after FTX went bankrupt [87], the community is still going strong and merits closer study.

We contend that the overlapping communities drawn together by these ideas form one coherent "epistemic community": a community with clearly-defined shared values and methods of knowledge production [153]. The impact of this epistemic community, which we hereafter refer to as the "AI safety epistemic community", extends beyond the community's bounds: non-profit and for-profit organizations, as well as academic research groups, have begun attracting sizable donations to fund their work. Furthermore, the AI safety epistemic community has also developed a variety of methods for expanding the reach of their ideas including online forums, career development programs, and policy advocacy. Through an analysis of the landscape of this community, we sought to answer the following research question: **How is the AI safety epistemic community developed and maintained through social, intellectual, and organizational practices?**

In this paper, we illuminate the central ideas and practices that define the emerging epistemic culture of AI safety. We are interested in how this epistemic community has translated their shared moral and normative claims into technical solutions and recommendations for AI policy that may have lasting, global implications. This work contributes to a broader understanding of cultural forces that influence certain types of AI development and deployment. As we note in Section 5.1, the AI safety epistemic community is not the only group concerned with the societal harms AI poses, and is often framed as being in direct opposition to the groups of researchers, advocates, activists, and critics who are collectively referred to as the "AI ethics" community and who emphasize the need to mitigate well-documented, present-day harms of AI systems. This paper will not explore other, parallel communities in depth, as our objective is to provide a rich analytical description of the AI safety epistemic community in particular.

To motivate our analysis, we first explain the theoretical framework of epistemic culture and our methodology. Next, we map the origins of three core ideas (effective altruism, existential risk, and AI safety) that have brought multiple communities under the umbrella of the AI safety epistemic community. Then, we explore four mechanisms for the development and transmission of these concepts in the emerging field of AI safety: online community-building (career advising and web forums), AI forecasting, research papers, and prize competitions. In the discussion, we synthesize the main characteristics of these four mechanisms, and then address the influence they have had outside of the community. We review critiques of the ideas and practices of this emerging field, revealing how the influence of the epistemic culture persists despite these concerns. Finally, we conclude with suggestions for future work that can build on our study's initial landscape of this epistemic culture, as we anticipate that it will only continue to influence how people the world over think about AI.

## 2 METHODS

Whereas Knorr-Cetina's approach to studying epistemic culture involved ethnographic studies of research labs, our methodology instead took advantage of how much of the AI safety community's epistemic culture unfolds online, providing ample documentation of value to our study. We combined approaches from critical technocultural discourse analysis [31] and frame analysis [77] to understand which types of causes, projects, and organizations are prominent in the AI safety epistemic community.

We first compiled a list of individuals and organizations from four separate sources that reflect the spending priorities [1] and field-building potential of this epistemic community. We referenced two major grant-making bodies in this area, Open Philanthropy and the now-defunct FTX Future Fund; the main career services organization for EAs, called 80,000 Hours; and an annual organizational review sourced from a community forum. Specifically, the four data sources included: 1) people and organizations who received funding from Open Philanthropy to work on advanced AI issues in 2022 [140]; 2) people and organizations who were selected for FTX Future Fund grants and regrants to work on AI [67]; 3) organizations referenced on the 80,000 Hours page on preventing an AI-related catastrophe [84]; and 4) organizations from the 2021 Alignment Literature Review and Charity Comparison published on the Alignment Forum by the user "Larks," who published a review of advancements in AI alignment each year between 2018 and 2021 [99]. We sourced from Open Philanthropy and 80,000 Hours to assess the most prominent, established organizations in the field; conversely, we included FTX Future Fund grantees and regrantees to capture new entrants to the field[68]. These lists allowed to us sample for variety and consistency (several people and organizations were cross-listed), with the user-generated review on Alignment Forum balancing out the top-down lists with a view arising from the participatory, community discussion-based approach to field-building.

We then turned our analysis to the output of the funded organizations and individuals, coding each entry from our aggregated list to identify the type of agent (e.g., researcher(s), nonprofit, for-profit company), the topics of work (e.g., AI safety, AI alignment, AI governance), and the method of dissemination (e.g., research, policy, prize competitions, fellowships). To derive these codes, we conducted discourse analysis across websites, research papers, reports, blog posts, forum posts, podcasts, and videos published by the individuals and organizations from our list. Through discourse analysis, we "interrogat[ed] power relations" emerging from these texts, focusing on "the interactions between technology, cultural ideology, and technology practice[31]. We likewise drew from frame analysis to "uncover the grounding assumptions and terms of debate that make some conversations... possible while forestalling alternative visions," [77] a way of gauging where the boundaries of AI safety lie. Both methods enabled us to make clear how discourse translates into practice in the field, and surfaced the four features of AI safety's epistemic culture that we describe in 4.

## 3 BACKGROUND

### 3.1 Epistemic culture

In our approach to understanding the community of individuals interested in AGI, x-risk, EA, AI alignment, and related topics (i.e., the AI safety epistemic community), we make use of the theory of "epistemic culture." Sociologist Karin Knorr-Cetina developed the concept of epistemic culture through an ethnographic study of the knowledge-making practices, communities, and symbolism in high-energy physics and molecular biology research labs [40]. In Knorr-Cetina's words, "Epistemic cultures are cultures that create and warrant knowledge." Investigating the social practices they comprise reveals "how we know what we know" (p.1) [40]. The framework of epistemic culture is useful to apply to the amalgam of ideas, knowledge production and circulation, and community-building that occurs within the AI safety epistemic community for two reasons.

First, epistemic culture is well-suited to illuminating "when domains of social life... curl up upon themselves and become self-referential systems" (p. 364) [39]. In other words, epistemic culture captures when a community cleaves away from the mainstream to develop their own terminology, source texts, and knowledge claims. The AI safety epistemic community comes together to pursue what they see as the marginalized but vital work of protecting humanity from AI's worst potential long-term harms. The technical and intellectual foundations on which the AI safety epistemic community bases their arguments (e.g., the timeline for when AGI will come into being or whether AGI is attainable) are subject to vigorous debate between AI experts (See Section 5.1). Thus, the frame of epistemic culture is useful for understanding how the AI safety epistemic community conceptualizes its work in the face of lack of consensus and speculative subject matter.

Second, a focus on epistemic culture foregrounds the actions that constitute "knowledge as practiced" (p. 8) [40]. As we discuss in Section 4, the AI safety epistemic community exchanges ideas, as well as attracts new participants via a network of web forums, job boards, blogs, conferences, prize competitions, forecasting activities, and other primarily online venues. These are the community's sites of knowledge production, debate, and what many sources in this paper refer to as "field-building" (See Section 4) [1].

We argue that a shared epistemic culture is the connective tissue that keeps the multiple sub-communities within the AI safety epistemic community in conversation with one another. While some individuals support many of the ideas this paper will cover, membership in one group, or belief in one concept, does not amount to blanket endorsement for all of these communities and ideas. For example, not all AI safety researchers identify as effective altruists. Despite these differences, people in this epistemic community collaborate on field-building and knowledge production, undergirded by overlapping moral convictions, strong beliefs about the importance and neglectedness of this work, and shared methodological training.

## 3.2 Effective Altruism, Existential Risks, and AI

The intellectual movements of utilitarianism, transhumanism, effective altruism, and longtermism as well as how these movements conceptualize existential risk have had significant bearing on how existential risk and AI are framed in public and scholarly discourse today [18, 53]. In their account of the history of existential risk studies, Beard and Torres [18] identify three phases of the development of ideas about existential risk. The first phase includes conceptualizations of existential risk that draws from the techno-utopian transhumanism movement and utilitarianism. The second phase includes the development of effective altruism and longtermism. The third phase includes an expansion of the philosophical view point of existential risk to include frameworks other than utilitarianism. This phase also draws from a more interdisciplinary perspective on existential risk. Here, we briefly summarize this history to provide context for the cultural and intellectual influences on the AI safety epistemic community.

*Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards* [21], authored by philosopher Nick Bostrom, was a highly influential work in the first wave of existential risk studies [18]. In this paper, Bostrom taxonomizes risk in terms of three properties: scope, intensity, and probability. Scope refers to the size of the population that would be affected. He differentiates "global" risks as those with the potential to affect the entire population, as opposed to "local" or "personal" risks that affect either subpopulations or individuals. For intensity, he differentiates "terminal" from "endurable." Endurable events are those that, even if they are severe, can either be coped with or recovered from. This is distinct from terminal risks, which are those that either completely kill off their targets or irreversibly and negatively alter the targets' ability to live life according to their fullest aspirations. From this perspective, a risk is existential if it is both global in scope and terminal in intensity. In his words, an existential risk is, "one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential." This conceptualization draws from the philosophical framework of utilitarianism, which posits that ethical action should be driven by an estimation of the total well-being that can be achieved through that action. Actions or events that either end all human life or "drastically curtail its potential" can be viewed as a worse-case scenario from a utilitarian perspective [18].

Ideas from within the movements of transhumanism and posthumanism further extrapolate utilitarian ideas about human life [18]. From the transhumanist perspective, technology should also seek to "overcome the human condition" [154] by liberating humans from biological constrains on cognitive capacity and even death through technological advancement [22, 154, 173]. In the words of Bostrom [22], "present technologies, like genetic engineering and information technology, and anticipated future ones, such as molecular nanotechnology and artificial intelligence" can contribute to well-being through a "radical extension of human health-span, eradication of disease, elimination of unnecessary suffering, and augmentation of human intellectual, physical, and emotional capacities." Thus, the creation of "superintelligence" can be seen as a transhumanist endeavor insofar as it achieves any of these goals [22, 154, 173]. On the other hand, pursuit of these goals through technology could also result in an existential catastrophe. Thus, technologies that could realize the most benefit to human well-being also create a significant existential risk.

Building on ideas from the first phase, the second phase of existential risk studies translated existential risk studies into more mainstream perspectives, especially those adopted by the effective altruism community [18]. William MacAskill, an originator of EA, defines effective altruism as "(i) the use of evidence and careful reasoning to work out how to maximize the good with a given unit of resources, tentatively understanding 'the good' in impartial welfarist terms, and (ii) the use of the findings from (i) to try to improve the world" [104]. As we show in Section 4.1, committed EAs often make decisions about their philanthropic donations and career choices in line with the movement's priority cause areas, which include global health and development, animal welfare, and mitigating global catastrophic risks. The importance-tractability-neglectedness (ITN) framework for quantifying how an issue should be prioritized at the margins is a staple within EA [172]. *Importance* or scale defines the amount of good that could be done if a problem is solved or alleviated; *tractability* is about how realistically solvable the issue is; and *neglectedness* illustrates how many resources are directed towards the problem relative to others.

Many EAs additionally identify as "longtermists." Longtermism is "the idea that we should prioritize positively influencing the long-term future of humanity—hundreds, thousands, or even millions of years from now [150]." From this perspective, the consequences of any actions we take now for future people's safety and well-being in the long-term future should enter the calculus about how to prioritize efforts in the present. Thus, longtermist perspectives build on core ideas from effective altruism and utilitarianism by prioritizing actions that can maximize utility. In some cases, EA-related theorists in this phase broadened the notion of what is to be maximized to expected value. Thus, "an existential catastrophe is an event which causes the loss of a large fraction of expected value" [51].

Notably, this perspective further introduces ambiguity into conceptualizations of existential risk. In a framework where value is "whatever it is we care about and want in the world [51]," estimations of existential risk are heavily dependent on whatever the estimator considers valuable. In Bostrom's early conceptualization, he notes, "For there to be a risk, given the knowledge and understanding available, it suffices that there is some subjective probability of an adverse outcome, even if it later turns out that objectively there was no chance of something bad happening" [21]. This notion of probability connects to that employed in a Bayesian tradition in which beliefs can be formulated even with limited knowledge of relevant information and updated over time as more information is acquired. Combining both Cotton's definition of value and and Bostrom's definition of probability results in an expected value calculation that is heavily dependent on the estimator's perspective and knowledge, even if it does result in a precise numerical value.

The third wave of existential risk studies incorporates a wider range of intellectual perspectives, especially those related to ethical frameworks. Instead of thinking about existential risks as arising from a single, causal factor, this perspective acknowledges that existential risks are likely to arise from complex systems of interacting factors [18]. Thus, a comprehensive understanding of existential risks demands an interdisciplinary approach, or even an expansion of the philosophical viewpoint beyond utilitarianism, which is not the dominant framework in other domains of philosophy [52]. In other words, philosophers and AI researchers working in isolation with a single perspective can neither sufficiently capture the interdependent factors that produce existential risks nor develop adequate mitigation strategies. This perspective embraces the idea that the onset of existential catastrophes could be gradual rather than sudden. This justifies a shift towards problems that are occurring in the short and medium term since successful redress of problems arising currently may significantly diminish existential risks that otherwise might arise in the future [18].

## 3.3 AI Safety

A separate but tightly related community that has grown alongside x-risk, longtermist, and EA communities constitutes technical experts who have been concerned with issues they group under the label of "AI safety." The term "AI safety" has been used since the early-to-mid 2010s, and grew in popularity during the period of time when deep reinforcement learning (RL) agents reached impressive landmarks (for instance, the success of AlphaGo). As a result, many of the concepts, failure cases, and vocabulary that AI safety practitioners use borrow heavily from the technical study of RL agents. We address ideas in AI safety in greater depth in Section 4.3. Generally, AI safety practitioners are interested in preventing catastrophic long-term events precipitated by the deployment of machine learning systems. These systems are often modeled as algorithmic agents with capabilities that will inevitably grow far into the future. As their capabilities expand, these agents may in turn become less predictable, understandable, and controllable. Many AI safety advocates are extremely concerned that AI systems could pose an x-risk in the near or far future, and are confident that it is impossible to halt AI developments altogether [27]. This combination of beliefs is disseminated throughout the epistemic culture we detail below.

# 4 HOW THE AI SAFETY EPISTEMIC COMMUNITY IS DEVELOPED AND MAINTAINED

## 4.1 Online Community-Building

*4.1.1 Forums.* Online community-building is a key element of AI safety epistemic culture. Through dedicated web forums, community members develop shared context, language, and reasoning style. We review three popular EA or EA-related online discussion spaces: EA Forum, LessWrong, and the AI Alignment Forum.

EA Forum, run by the nonprofit Centre for Effective Altruism (CEA), is the primary venue for EA-related discussion [59]. LessWrong and the AI Alignment Forum are two closely related forums with overlapping readership and participation [11, 102]. Many posts are cross-posted or referenced on all three forums. Originally founded by Eliezer Yudkowsky, a central figure in AI safety and co-founder of the Machine Intelligence Research Institute (MIRI), LessWrong is dedicated to training people in rationalist reasoning and decision-making [149]. The Alignment Forum was launched in 2018 as a discussion space for AI alignment researchers[20].

In these forums, EA and AI safety communities debate and negotiate the core concepts behind AI safety and alignment. From casually involved EAs to professional AI researchers, participants turn to forums for frequent in-depth discussions, e.g., evaluating and making predictions on when and how AGI might be developed, detailing possible pathways to AI-generated catastrophic and existential events, and brainstorming future directions for AI alignment research. Many leaders from the EA community as well as prominent AI researchers keep tabs on these forums, engage, and inform their opinions and decisions based on content.

The Alignment Forum is unique within AI safety field-building, as it aims to be a central "publication destination for AI Alignment discussion" and to "serve as the archive and library of the field" [20]. Its leaders posit that it fills an essential role that in the past may have been filled by an academic venue such as a conference or journal [20]. Posts cover a variety of topics, from fleshing out key conceptual factors in AGI alignment to exploring new research directions. Some posts are published in the open-access research repository arXiv, a signal of the Alignment Forum's pursuit of scholarly legitimacy, though forum participants disagree on how useful this is [30]. Many alignment-focused organizations cite their Alignment Forum posts as their published research.

In addition to its narrow focus, the Alignment Forum differs from EA Forum and LessWrong in that it gates full participation. Anyone can become a member to post or comment on EA Forum and LessWrong. By contrast, the latest membership update in 2021 shows the Alignment Forum has 50 - 100 members and grows slowly. Members are researchers from major institutions within the epistemic community, including the Future of Humanity Institute (FHI), Berkeley CHAI, DeepMind, OpenAI, MIRI, Open Philanthropy, and

Alignment Research Center (ARC) [20]. Membership is extended to those who existing members believe to have a strong track record of alignment research and to be highly trusted. They can post and comment directly, whereas non-members can only participate by submitting posts and comments to be reviewed before being published or by posting or commenting to LessWrong first. Because the two forums are integrated, members can promote content from LessWrong to the Alignment Forum. The forum leaders believe that these high membership standards improve the quality of discussion and peer review in such a new field [81].

**These three web forums simultaneously attract people newly interested in these topics, sustain an international community of researchers and non-experts, disseminate and enable continued revision of this epistemic community's beliefs, and in some cases promote offline participation in the community by nudging people towards careers in the field.** The Alignment Forum in particular documents a community's attempt to define a nascent field and adapt research production flows to fast-moving technical developments.

*4.1.2 Career advising.* A core impulse of EA communities is to help members think empirically about how they spend their money and time using EA principles, and then follow through on their conclusions. 80,000 Hours is a non-profit organization that serves as EA's primary career planning hub, and is unique in its cohesive bridging of a particular worldview to strategic, step-by-step career planning. 80,000 Hours self-describes as advising

> *a particular audience: namely, people with college degrees who want to make having a positive impact (from an impartial perspective) the main focus of their careers, especially in the problem areas we most recommend; who live in rich, (for the most part) English-speaking countries; and who want to take an analytical approach to their career.* [163]

80,000 Hours provides career planning resources, personal advising through calls with staff members, and a curated job board. Career planning content includes profiles on the "world's biggest and most neglected problems," analyzed through the ITN framework [5]. 80,000 Hours ranks AI risk as the most important issue to work on, because they view this area as both relatively neglected and overwhelmingly important currently and for future generations. In their AI risk problem profile, they justify their assessment, address counterarguments, provide resources for learning more about pathways to existential catastrophes, and recommend both specific issue areas and organizations to work for to alleviate these issues [83]. Of the 54 organizations or sets of roles they recommend, 21 are characterized as "AI policy and governance" or "AI technical safety" [4].

In addition to 80,000 Hours, many smaller organizations are dedicated to upskilling and funneling EAs into AI safety roles. These include newer organizations that FTX Future Fund-selected—AI Safety Support [9] and AI Safety Community Field Building Hub [71], for example—as well as more established initiatives such as BlueDot Impact's AGI Safety Fundamentals [6]. Likewise, several AI safety researchers have written explainers and starter guides to initiate people into the field via forum posts [6, 20, 27, 158]. Both 80,000 Hours and these smaller educational, training, and coaching organizations work in tandem towards their goal of field-building.

## 4.2 AI Forecasting

As with forum posts, the practice of forecasting near- and long-term future outcomes for AI and AGI serves at least two functions: community-building and producing knowledge claims. In much of AI-related forecasting, individuals or teams pose a question and predicted guess about when a future event in the development of AI will occur. Questions of when humanity will attain AGI, "transformative AI" [93] "strong AI," [2] and certain AI benchmarks are among the common forecasting topics. In some cases, forecasters proffer the evidence behind their reasoning and document their methodology [10]. Forecasts and their explanations are often posted to the three web forums described in Section 4.1.1, and through forecasting platforms like Metaculus [108] and Elicit [61].

In non-expert forecasting, anyone can post a prediction, and the cross-posting of forecasts on web forums opens others' forecasts up for commentary and revision as new research is published and taken into account. Participants believe they can come to more accurate beliefs on x-risk potentials, especially around AI, by honing their forecasting abilities in these participatory forums [109]. Some organizations have developed explainers and dictionaries [139] to build "forecasting infrastructure." Others have aggregated multiple forecasts to analyze meta trends like how predictions have changed over time [7]. Forecasting also figures into 80,000 Hours' recommendation of working on "epistemics and institutional decision-making" [171]. X-risk focused organizations such as Stanford's Existential Risks Initiative (SERI) [101] and Berkeley Existential Risk Initiative (BERI) have issued grants for events like AI forecasting workshops [100].

Academics focused on nearer-term predictions have also championed forecasting. For example, in 2021 Professor Jacob Steinhardt of UC Berkeley hired professional forecasters to answer questions such as "On June 30, 2022, what will be the state-of-the-art accuracy of a machine-learning model on the MATH dataset?" regarding a dataset of high school-level math problems developed by UC Berkeley researchers, with responses plotted as predicted percentages of the model's accuracy [159, 160]. After the dates in these questions passed, researchers extrapolated broader conclusions based on how closely forecasters' predictions approximated the accuracy of these models. These included assessments that achievement of certain AI benchmarks happened faster than forecasters predicted, and that safety performance is not keeping up with the pace of new AI capabilities. In turn, this epistemic community treats this type of evaluation as justifying more urgency in allocating resources for AI safety work.

Thus **AI forecasting not only produces knowledge claims that these communities base decision-making on and shape research agendas around, but also serves a social function as members of the AI safety epistemic community consume and evaluate one other's forecasts.** Forecasting is further popularized through its citations in research and forecasting prize competitions that can expand the reach of these communities.

## 4.3 AI Safety Research

While web forums enable interaction between AI experts and non-experts, these venues are not the primary arena through which novel technical ideas arise. Technical developments in AI safety are the product of researchers working at academic and non-academic AI safety research centers, and are typically communicated through conference proceedings, journal articles, and technical reports. Prominent AI safety researcher and NYU professor Sam Bowman has described the field of AI safety as "pre-paradigmatic," [29] in that research methods, practices, and norms of communication have yet to solidify. A recent literature review on AI alignment, a subfield of AI safety, echoes this. It notes that alignment "has not yet converged on a single, dominant paradigm or approach ...the current exploratory nature of AI alignment research might be a strength, as exploration helps to avoid ossification" [97]. The researchers currently interested in AI safety have begun to construct such paradigms through the production of empirical research papers, as well as through the iterative scientific discourse of methodological critiques, literature synthesis, and citations, bringing the selection and endorsement of topics and methodologies in the field into sharper relief. While most of the influential scholarship that passes between these groups takes the form of technical computer science papers, philosophy and AI governance-focused papers are also prevalent.

At present, AI safety publications tend to originate from within tech companies and nonprofits, with some collaborators hailing from academia. The non-academic AI safety research organizations we investigated included the companies Anthropic, DeepMind, and OpenAI, as well as research nonprofits Redwood Research and ARC.[3] Non-academic AI safety organizations, which include for-profit, capped-profit, public benefit corporation, and 501(c)(3) non-profit structures, produce research papers that endeavor to serve their stated goals of ensuring that AGI benefits humanity [58, 135] and that future AI systems are aligned with human values. Most of the AI safety research deriving from the non-academic organizations we investigated is posted on the institutional websites of the authors, community forums, and on arXiv. Some of this work is not formally peer-reviewed, although researchers frequently cross-post their papers to the main web forums of this community, where their findings are then debated in an informal alternative to peer review. We note that such informal review methods differ from those of other mainstream scientific disciplines including in computer science, in which revision in response to critiques by subject matter experts is considered a indispensable component of the production of high-quality scientific findings.

Although AI safety has yet to converge on an established set of specific research focuses and methods, several topics have garnered attention from researchers both inside and outside of academia. One influential paper from authors at UC Berkeley, Google, and OpenAI, titled "Unsolved Problems in ML Safety," identified four areas for the field to prioritize: "Alignment", "Monitoring", "Robustness" and "Systemic Safety" [82]. These areas map closely onto a recent Open Philanthropy- and Good Ventures-backed National Science Foundation (NSF) solicitation for grants, "Safe Learning-Enabled Systems," which anticipates issuing up to USD $20 million in funding for academic AI safety research [117]. While these four problem areas are not exhaustive, they provide a useful frame for surveying the landscape of AI safety research to date.

Alignment, delineated by Hendrycks et al. [82] as a focus on how to "build models that represent and safely optimize hard-to-specify human values," is a fast-blooming area of AI safety. AI alignment researchers have grappled with what it would mean for AI to be aligned with human values (so-called "value alignment"). For example, although humans' expressed intentions, explicitly provided instructions, revealed preferences, ideal preferences, and interests are often related, the differences between them have important implications for AI alignment [41, 70]. Furthermore, even if what constitutes alignment could be precisely defined, it may not be possible to access information that would allow for an operationalization of alignment [70]. Thus, "value alignment" work tries to define appropriate paradigms for identifying and representing human preferences and values. While one paper suggests drawing on human rights as an alignment baseline [143], there have also been calls for enlisting social science researchers [90] and relying on upvoting mechanisms to measure human preferences [88]. Yet even if upvoting or some other reaction mechanism can provide information about human preferences that helps align AI, the incorporation of such mechanisms may have negative downstream consequences [138]. As a result, even at the level of subtopics, AI alignment research is inextricable from philosophical questions about morality [70] and social scientific questions about human behavior.

Hendrycks et al. [82] describe "monitoring" as "identifying hazards" such as malicious uses and anomalies. To them, this issue area also encompasses work on making models "honest": many AI safety researchers posit that as some AI systems are incentivized to produce results they expect will be rewarded, they may in some instances produce outputs that are false and can cause harm. For instance, some researchers are concerned with the possibility of AI systems whose capabilities and reward-driven development lead to them "gaining and maintaining power over humans and the real-world environment" (i.e., "power-seeking AI") [34], and note that deception could ensue as a result. Among proposed methodological solutions are eliciting latent knowledge (ELK), a framework for identifying instances in which ML systems will misrepresent reality to human observers, and mitigating this misrepresentation by detecting and then preventing it [43]; and constitutional AI, in which a pre-specified set of principles (a "constitution") serves as the sole human oversight guiding an "AI assistant" that reduces harmful behaviors of other AI systems [16]. Relatedly, AI safety researchers have observed that as certain models become exponentially larger, or "scale," they may develop capabilities their designers did not intend or expect them to have, such as large language model GPT-3 being able to do arithmetic. "Monitoring" includes investigating methods for identifying when and where these new capabilities will emerge and how uncertainty in these outcomes might be reduced. Examples of this include work on scaling laws, which involves determining how the properties of a model change as the model scales, and work on mechanistic interpretability, which involves reverse-engineering and interpreting ML models to pinpoint the "algorithmic patterns, motifs, or frameworks" [60] that may indicate the "mechanism" through which new capabilities develop [130].

"Robustness" involves "withstanding hazards" such as "black swan" or "long-tail" (i.e., very unlikely but very consequential) events, as well as being capable of effectively responding to attacks by adversarial agents. As motivating examples of long-tail events, Hendrycks et al. [82] cite computer crashes of automated trading systems and the COVID-19 pandemic. To avoid similarly rare, but highly consequential, outcomes from AI systems, these authors suggest generating more benchmarks and stress-testing of environments using simulated data. With respect to adversarial attacks, they encourage researchers to expand their research beyond conventional topics, such as how to ensure robustness against adversarial attempts to undermine ML classifiers via very small perturbations to model inputs. Instead, they suggest researching wider conceptualizations of what constitutes an adversarial attack, including threats that are obviously different from non-adversarial inputs but that can nevertheless evade detection.

Finally, "systemic safety" addresses the contexts in which AI systems are deployed. Here, Hendrycks et al. [82] raise examples of cyberattacks, as well as citing more critical literature, such as research on LLMs as "stochastic parrots" reproducing the social harms embedded in the online sources on which they have been trained [19]. Systemic safety also includes the AI governance literature that comes out of the same epistemic culture and institutions this paper covers [56]. The authors acknowledge that much of the systemic safety research occurs in other fields, such as privacy, fairness, and ethics, while also proposing solutions that tie back to the epistemic community this paper researches, including improving ML forecasting capabilities to predict future events.

Some experts in the field have questioned whether academia is well-suited to AI safety research since studying the effects of large models often demands having access to the massive computational resources necessary to train such models—resources that are concentrated within a few tech firms [29]. Even still, a growing handful of university professors and graduate students work on and disseminate their AI safety research in more traditional publication venues and conferences. We initially identified seven researchers who were listed as grantees or regrantees of the FTX Future Fund—Lionel Levine (Cornell University), Anca Dragan (University of California, Berkeley), Daniel S. Brown (University of Utah—or who were mentioned in an 80,000 Hours post describing key organizations within academic labs—Sam Bowman (NYU), Dylan Hadfield-Menell (MIT), David Krueger (University of Cambridge), and Vincent Conitzer (Carnegie Mellon University). After investigating the research profiles of these seven researchers, we also included He He and Mengye Ren as both are listed as co-PIs of Sam Bowman's lab, and Adrian Weller since he is listed as a member of the Computational and Biological Learning Laboratory at University of Cambridge alongside David Krueger. We then surveyed the professional profiles and publication records of these researchers. The majority were trained in computer science or related disciplines such as robotics or electrical engineering and hold appointments in computer science or computer science-adjacent fields (e.g., data science). Most are early-career, with eight out of 10 having completed their PhDs in the last 10 years. Over the past three years, they mainly published papers in disciplinary computer science conferences (e.g., Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML)). Where taxonomic meta-data was available, publications were most often classified in subareas of computer science such as artificial intelligence, machine learning, computer vision (and related topics), language, and robotics, though several publications were classified under topics associated with social science, humanities, or law.

On the whole, our analysis suggests that academic AI safety researchers tend to disseminate their work in disciplinary computer science venues and address topics that are broadly of interest to the computer science researcher community. In other words, **the reach of these researchers is not constrained to research communities primarily interested in AI safety or closely related topics, but rather, is likely to have influence in traditional computer science domains**. Furthermore, because most researchers we investigated had accepted faculty appointments relatively recently, their impact both within and beyond the subfield of AI safety will likely be magnified over time as their trainees also secure positions in academia or in industry research centers. As a result, **scholarly research represents one of the core avenues through which ideas about AI safety have begun to take hold outside the AI safety epistemic community**. As with web forums and forecasting, research papers allow for collaborations across boundaries, with many co-authored papers bringing together representatives of industry and academia, and in some cases reflecting how some individuals move between both companies and universities. The fourth and final site of the epistemic culture we investigated, prize competitions, also traverses the boundary between industry and the academy.

## 4.4 Prize Competitions

Prize competitions are another venue in which ideas from effective altruism, longtermism, and conceptions of x-risk interact with AI safety. Often technical in nature, each contest frames a particular problem in machine learning and solicits solutions. Competition websites include submission guidelines, along with research publications that entrants are encouraged to consult before submitting papers, code, and other documentation of their work. Winning individuals and teams receive cash prizes from prize pools that are orders of magnitude larger than those academic computer science conferences tend to grant.

We assessed 10 prize competitions organized by institutions with an explicit connection to x-risk and AI safety to better understand the communities, practices, and problem formulations they socially reproduce and incentivize. Two of these were one-time events hosted by AI safety nonprofits. In 2022, ARC ran a competition which aimed to build on their work on eliciting latent knowledge (ELK). ARC solicited strategies, code, and pseudocode that took into account a bank of counterexamples raised by the organizers and commentators on the competition announcement post [42]. The nonprofit Fund for Alignment Research (FAR) also ran their Inverse Scaling Prize that year to find "new examples of tasks where pretrained language models exhibit inverse scaling: that is, models get worse at the task as they are scaled up"[107].

In addition to standalone prize competitions, institutions in the AI safety epistemic community have hosted contests within established academic conferences and workshops. The Center for AI Safety (CAIS), a "research and field-building nonprofit," supported a competition on Trojan attack detection in deep neural networks at the 2022 NeurIPS conference [36] and organized the 2022 NeurIPS ML Safety workshop. The workshop's Best Paper Award encouraged participants to work within the four thematic areas specified in the paper "Unsolved Problems in ML Safety" discussed in Section 4.3. One of the paper's authors served as a judge for CAIS' non-academic SafeBench benchmarking prize competition, which referenced these same categories (replacing "Systemic Safety" with "Safety Applications") [35]. The SafeBench guidelines note, however, that "submissions will be judged according to their relevance to risks from advanced AI, not to these categories."

Two of the prize competitions that were originally to be funded by FTX Future Fund appear to have erased their x-risk and longtermist premises with no explanation offered. What was first announced as the NeurIPS ML Safety workshop's "Existential Risk paper prize" [119] to find the "best accepted ML Safety papers that have the highest quality x-risk analyses" was rebranded as the Paper Analysis Awards "for accepted papers that provide analysis of how their work relates to the sociotechnical issues posed by AI" [120]. Similarly, the Future Fund's original description of the European Conference on Computer Vision (ECCV) 2022 Workshop on Adversarial Robustness in the Real World noted that "The best papers are selected to have higher relevance to long-term threat models than usual adversarial robustness papers" [67]. The eventual virtual workshop that was held, however, did not mention longtermism [131]. Given the collapse of FTX, these revisions might signify trepidation around bringing x-risk and longtermist ideas into academic spaces or the waning influence of a major funder's desired frame.

Two additional nonacademic competitions that bridge to academia were sponsored by CAIS and ML Safety (a social community CAIS sponsors). The first was their Moral Uncertainty competition, which solicited submissions that use "machine learning models to estimate when human values are in conflict and estimate moral uncertainty" and required prize-qualifying entrants to "have a corresponding accepted paper at reputable machine learning conferences and/or journals" [114]. The second was their Autocast Forecasting competition to "build a machine learning model that makes accurate and calibrated forecasts" responding to forecasting questions sourced from platforms such as Metaculus. The organizers "reserve[d] the right to… integrate this round into a NeurIPS competition" [112].

**Prize competitions present an opportunity for ideas in this epistemic community to be socially reproduced within larger audiences that have little to no prior exposure to x-risk and themes within AI safety.** The community has also appended social events to these workshops, such as the ML Safety Social [113] and the AI Safety Unconference [8] at NeurIPS. The latter event, which has run three times to date, provides reading recommendations from the alignment and AI safety literature as well as Bostrom's *Superintelligence*, a reminder of how intertwined these communities continue to be. These competitions provide community-building efforts, incentivize the production of research and solutions to long-term technical problems in AI safety, and in some cases seek for this work to acquire academic legitimacy.

**Table 1: AI Safety Prize Competitions**

| Prize Competition | Funder(s) | Prize Pool Total |
|---|---|---|
| Best Paper Award, ECCV 2022 Workshop on Adversarial Robustness in the Real World (AROW) [131] | Open Philanthropy | $30,000 |
| Trojan attack detection on deep neural networks, NeurIPS 2022 [36] | Open Philanthropy | $50,000 |
| Best Paper Award at ML Safety workshop, NeurIPS 2022 [120] | Open Philanthropy | $50,000 |
| Paper Analysis Awards[1] at ML safety workshop, NeurIPS 2022 [120] | Open Philanthropy | $50,000 |
| Machine Learning Moral Uncertainty Competition [114] | Center for AI Safety | $100,000 |
| Autocast Forecasting Competition [112] | Center for AI Safety | $125,000 |
| MineRL BASALT competition, NeurIPS 2022 [110] | FTX Future Fund[2], Encultured.ai, Microsoft | $155,000 |
| Inverse Scaling Prize [107] | FTX Future Fund[3] | $250,000[4] |
| Eliciting Latent Knowledge prize competition [42] | Alignment Research Center | $274,000 |
| SafeBench Competition [35] | Center for AI Safety | $500,000 |

[1] Originally "Existential Risk Prize"; [2] Original description was that "grant will be administered by the Berkeley Existential Risk Initiative; [3] Organizers did not disclose where the replacement funding they were able to raise originated from; [4] Original amount promised by FTX, but organizers note that they were only able to raise $50,000 after FTX bankruptcy

## 5 DISCUSSION

Throughout this paper, we have cited how many people in the AI safety epistemic community describe their work as important, yet neglected or marginalized by others. However, this community has stepped out of the margins and into a more mainstream position to affect public thinking about AI. From OpenAI's launch of ChatGPT to Anthropic's USD$4 billion valuation [44], we have seen how a small group of actors can set off a wave of attention, investment, and debate about whether and how to deploy certain AI systems. In part, this shift into the mainstream derives from the large sums of money involved—from the ultimately withdrawn financial injections of FTX Future Fund to the more concrete recent investments in AI chatbots. Yet, our paper argues that the ongoing influence of this epistemic culture over broader AI discourse is not solely driven by the community's financial resources. **It is instead the cohesive, interwoven social structures of knowledge-production and community-building that will ensure the continued reproduction of their ideas about AI outside of the boundaries of this community.**

In our review of online community-building through career advising and web forums, forecasting, research, and prize competitions, we show how each of these individual sites serve multiple functions. Career advising hubs and web forums enable both new initiates who have dabbled with ideas from EA and AI safety, as well as more deeply involved members of these communities, to socially and professionally structure their time around contributing to their goal of preventing AI x-risks. Forecasting produces both professionally sourced and crowd-sourced predictions about future AI outcomes, which in turn are cited in research and used as evidence to justify decisions about which work to prioritize and the timelines along which to do so. Forecasting is also social; posting predictions and the reasoning behind them leads to debate and interaction with others about their forecasts. Technical research papers and reports are the main knowledge production output of the AI safety epistemic community. They also enable ideas from this community to travel between industry, academia, and the less formal social spaces of web forums to spark further debate. Research papers and many prize competitions attempt to legitimize this epistemic community's views on AI within academia. Prize competitions in particular present an opportunity to recruit new people into exploring ideas about x-risk and approaches to AI safety that have been shaped by this epistemic culture. All four of these mechanisms interlock to socially reproduce a specific set of ideas about AI safety.

What happens when the products of this epistemic culture take root outside of this tight-knit, globally distributed epistemic community? To answer this question, we start by summarizing critiques from within and outside of the community. We then discuss how, in spite of these criticisms, this community's ideas have begun to exert influence in media narratives about AI as well as in AI policy forums. We close with suggestions for future work that can build on the foundation of this paper.

## 5.1 Public critique

Along with the rise in public awareness about EA and its associated views on AI, there has been an uptick in criticism from both within and outside of the movement. Within EA circles, some have expressed concern that AI safety receives a disproportionate amount of attention and resources from the community relative to the level of certainty over risks and timelines to advanced AI[49]. There is also internal debate over the distinction between working on advancing AI capabilities versus safety [168][33]. Another line of internal critique points out the lack of diversity and pluralism within the community, and unaccountable trust in leaders[52, 54].

In 2023, the EA community grappled with its demographic homogeneity, allegations of sexism and harassment within the movement [12], as well as the resurfacing of an email containing racist rhetoric from Bostrom [72]. In his public response, Bostrom did not reject the possibility of genetic differences contributing to different levels of intelligence between people of different races [24]. This prompted the Global Catastrophic Risk Institute to issue a statement distancing themselves from Bostrom's perspective [17]. This inter-community grappling with how one central figure in the epistemic culture has treated eugenicist ideas occurs in parallel with the growth of a body of research identifying how the influence of eugenics has been present throughout the historical project of building AI and AGI [74, 94].

Outside of the community, critiques similarly contest ideological and technical underpinnings of x-risk and notions of intelligence. For instance, scholars have problematized Bostrom's prioritization of human intelligence above all other traits, especially given the historical harms that have been premised on hierarchies of individual or group-level intelligence [47, 155]. Although Bostrom's *Superintelligence* [23] was not the subject of these critiques, in it Bostrom describes genetic engineering as one of the mechanisms through which a superintelligent agent might be achieved. Though he goes into some detail (p. 47-52) as to how social dynamics might accelerate superintelligence by this method, he makes only a passing mention of the history of violence associated with eugenicist ideas (p. 52). Thus, concerns about the links between eugenics and EA-inflected ideas about intelligence apply not only to AI, but to such thinking more broadly.

Finally, critics challenge the very notion that AGI is attainable and that current advances in AI warrant the coordinated, resource-intensive response this epistemic communities has marshaled. The influential 2016 OpenAI paper "Concrete Problems in AI Safety," [13] and a 2020 rebuttal paper, "Concrete Problems in AI Safety, Revisited," [146] pre-figured contemporary arguments between the AI safety epistemic community's proponents and detractors. More recent work continues to challenge the prioritization of AI safety despite current fundamental failures of AI functionality [147].

Warnings against AI doomsday hype [32, 118] and AI "snake oil" [116] also undercut foundational concerns of the AI safety field. In 2023, both the Future of Life Institute [125] and the Center for AI Safety [65] issued open letters calling on the public to recognize the existential risk that AI poses, with both letters attracting thousands of signatures from a wide range of figures including prominent technology CEOs and respected computer science professors. In response, organizers of the Fairness, Accountability, and Transparency in Machine Learning (FAccT) conference posted a letter raising concerns about real-world, empirically proven AI harms and gathered over 250 signatures from academics, technologists, and advocates [50], and . Direct critiques of the AI safety epistemic community itself condemn its EA connection and argue that the community's role in general AI development will lead it to harm already marginalized groups rather than benefit humanity [73], as well as point out that this epistemic community's prize competitions may rely on unpaid labor of non-prize-winning entrants [79]. In spite of these criticisms, the AI safety epistemic community still wields influence in media and policy venues that are interested in AI.

## 5.2 Media and Policy Influence

*5.2.1 Media.* The wider EA community has steered resources towards promoting EA ideas in non-EA media outlets, which creates an opening for shaping public narratives about AI. The community has influenced media in several ways. First, they seek to attract positive attention from the press. Organizations like the Future of Life Institute (FLI) and the Center for Long-Term Resilience have boasted in their annual reports about the coverage they received in the mainstream media [66, 124]. The publication of MacAskill's book and the swift rise of FTX also led to widespread, generally favorable, coverage of EA ideas in 2021 and 2022 [103, 137, 152, 161], including views on AI- and AGI-related x-risks.

Second, recognizing the field as having the potential for high impact, EA organizations have encouraged members to cultivate careers in journalism [64]. For example, in 2022, the Effective Altruism Fund sponsored Training for Good's Tarbell Fellowship for early-career journalists interested in covering emerging technologies, particularly AI.[162]. Similarly, Open Philanthropy sponsored a fellowship where one of the project goals was to "demystify longtermist and EA concepts, especially for skeptics" [133]. Beyond training programs, some EA advocates have invested in journalistic outlets as a way to promote EA-related journalism. For example, Bankman-Fried was the largest investor in Semafor, a media startup launched in 2022 [55]. Through his family foundation, he also donated six-and seven-figure grants to popular publications like ProPublica [144], The Intercept [78], and Vox's "Future Perfect" vertical for reporting on topics such as "technological and innovation bottlenecks that hamper human progress" [167].

Third, members of the community have created new media projects with EA-aligned goals and values. Asterisk, a magazine launched in 2022 with backing from the Effective Ventures Foundation and Open Philanthropy, proclaims that its editorial perspective is "shaped by the philosophy of Effective Altruism, but not limited to it" [14]. One of the contributors [142] is a writer for Vox's Future Perfect vertical, which is transparent about its roots in the effective altruism movement [105] and which has covered EA steadily [106, 141, 151]. Finally, alternative media to journalistic coverage include tabletop [98] and video games [62] that teach players about AI safety and AI x-risk, and a museum exhibit about their potential catastrophic outcomes [2][91].

How the downfall of FTX will influence EA-focused media endeavors is not yet clear [63, 115, 145, 167]. However, given the multiple avenues of influence, EA ideas are likely to continue to gain ground within the media. Despite mainstream English-language media's coverage having shifted from curiosity about EA toward a more suspicious stance, tech policy is more open to suggestions from EA, x-risk, and AI safety communities.

5.2.2 *Policy.* The AI safety epistemic community's modes of influencing AI policy shifted after the popularization of ChatGPT. From the mid-2010s up through late 2022, a mix of think tanks, nonprofits, academic research centers, and companies in this epistemic community produced reports on AI governance; made policy recommendations to mostly Global North-based governments [25, 89, 136, 169, 170] and multilateral bodies such as the United Nations, G20, and OECD [38, 46, 69, 96, 122]; provided Congressional testimonies in broader hearings about AI; and weighed in on specific initiatives such as the National Institute of Standards and Technology's (NIST) AI Risk Management Framework (RMF)[129].

Three themes emerged across their initial recommendations: 1) internationally cooperative versus rivalrous [4] AI development [26], 2) self-regulation, and 3) designated roles for government and academia. One example of support for self-regulation can be seen in an FLI representative's remarks at a NIST workshop praising the RMF's "soft law," "sector-agnostic" approach in which organizations can volunteer to publicly disclose risks they have discovered in their AI systems [128]. Finally, regarding actions academia and government should take, Anthropic co-founder Jack Clark posited that under the creation of the US government's National AI Research Resource (NAIRR), the government could provide academia the funding for "computational power closer to that found in industry" to conduct both experimental research and research that would provide "accountability for the private sector" [166]. This equates accountability not with with changing behavior within companies, but instead with giving universities the compute resources required to run large models so that they can produce similar AI safety research to that being done in industry. Critics argue that NAIRR would instead entrench power within the largest companies who have these resources and limiting what counts as "AI research"—i.e., sidelining the humanities and social sciences [57].

From late 2022 to the present, the AI safety epistemic community's influence over policymakers, particularly in the United States and United Kingdom, has markedly grown. UK Prime Minister Rishi Sunak announced that his government would commit to spending GBP £100 million on AI safety and host an AI Safety Summit in fall 2023 [76]. Anthropic, DeepMind, and OpenAI claim they will grant UK researchers 'priority access' to their models for safety evaluations[75]. Recent critics have also pointed out how members of the AI safety epistemic community are occupying government and civil service roles in the UK related to AI, raising concerns about conflicts of interest [45].

Congressional hearings featuring influential figures in this epistemic culture—ranging from OpenAI CEO Sam Altman [132] to both the founder and executive director of the think tank Center on Security in Emerging Technology (CSET) testifying in the same hearing [86]—are now regular occurrences. In 2023, the White House hosted several high-profile meetings with companies including Anthropic and OpenAI, resulting in outcomes such as a voluntary pledge by these and other AI companies promising transparency and safety on vague terms [95]. Ideas from the earlier period of the epistemic community's attempts to influence policy re-appear in recent high-profile events. For instance, at a NIST workshop Clark had noted that Anthropic hired crowd workers to red-team LLMs, or to pose malicious requests that researchers could take into account when updating systems to steer clear of these types of abuses, and proposed that the US government could create markets for AI red team services [127]. In 2023, the Biden-Harris administration announced support for an AI red-teaming event to be held at that year's DEF CON conference [85].

In sum, **many of this epistemic community's policy recommendations amount to maintaining the status quo of non- or self-regulation**, and support the approach companies like OpenAI and Microsoft are already taking—i.e., releasing unregulated LLMs to the public as a means of gathering feedback about the kinds of misuses and harms users will encounter in them. These recommendations blend in with a long-held tendency in the United States to avert over-regulation of technology that is feared to stymie innovation, yet they come at a time when many policymakers lack a comprehensive understanding of AI's harms [92]. At the same time, the AI safety epistemic community is likely to direct more members towards careers in policy; on 80,000 Hours page of highest-impact career paths, government and policy roles appear first, with emphasis on AI policy [3, 164]. Open Philanthropy sponsors a US policy fellowship [134] with placements in executive branch offices, Congressional offices, and think tanks, while Training for Good has a policy fellowship in the EU [165]. In addition, the Centre for the Governance of AI (GovAI) runs summer and winter fellowships for "candidates who are strongly considering using their careers to study or shape the long-term implications of AI" [37].

## 5.3 Future work

Future research could focus on at least two areas. The first is development of AI safety as an academic research field. Researchers could analyze papers coming out of university AI safety labs [15, 121, 123] and supported by the National Science Foundation's allocation of USD $20 million in grants for AI safety research [117]. Related work could focus on student-run AI safety and alignment groups [80, 111] and the expansion of prize competitions to include college and high school students[156] [157].

Another track for future research could be to study the institutional elements of this epistemic culture. From justifications for university research labs [28] to the Anthropic founders' decision to leave OpenAI and commit to the "focused research bet" of their company's work on AI safety [126], what are the affordances of creating dedicated spaces for this work? Several of these organizations operate as public benefit corporations, capped-profit companies, and 501(c)(3) nonprofits, which implies their work has a clear benefit to society. What counts as benefit to the public's experience of AI within this metrics-driven, longtermist-minded epistemic culture?

## 5.4 Limitations

One of the main limitations of this work is its reliance on publicly available data. One significant aspect of the epistemic culture that we did not include here is local offline meetups, ranging from different EA chapters' meetings to bigger conferences such as the annual EA Global conferences. Interviews or surveys may have uncovered different findings or gone into more depth on how the people participating in this epistemic culture experience it. Likewise, we were not able to capture the magnitude, e.g., of how many people ultimately compete in prize competitions.

Furthermore, because our analysis intended to provide a high-level map of the AI safety epistemic community, it was not also possible for us to conduct an exhaustive analysis of topics within each of the four areas of interest. Subsequent work could build on our analysis by expanding the breadth or depth of analysis within each of the areas. For example, additional insight into the AI safety epistemic community could be gleaned through a qualitative content analysis or quantitative topic analysis of the themes embedded within the work of AI safety research papers.

# 6 CONCLUSION

In this paper, we have traced how a set of ideas and practices are crystallizing into an epistemic culture that connects people whose interests lie at the intersection of effective altruism, existential risk, longtermism, and AI safety. Research papers, career advising, web forums, forecasting, and prize competitions all double as sites of knowledge production and community-building within this epistemic culture. The resulting AI safety epistemic community is not only successful at attracting young people who find its mission compelling, but also at mobilizing resources that keep people in these careers and communities.

To our knowledge, there is no comparable effort—both financially and socially—from any other community to influence AI's trajectory. By virtue of their dispersal throughout industry, academia and, increasingly, policy, this highly coordinated epistemic community will continue to influence global discourse about AI. Understanding the epistemic culture that fuses their moral convictions and knowledge claims is crucial to evaluating these claims as they circulate within critical conversations about AI's harms and how to mitigate them. While it remains to be seen whether this community's views will become an epistemic monopoly, the template they have provided for how to build a field will undoubtedly continue to shape academic research priorities, industry practices, regulatory responses, and government resource allocation for AI.

# ACKNOWLEDGMENTS

[blinded]

# REFERENCES

[1] 2023. AI Alignment Fieldbuilding. https://www.alignmentforum.org/tag/ai-alignment-fieldbuilding.

[2] 2023. Misalignment Museum. https://www.misalignmentmuseum.com/.

[3] 80,000 Hours. 2022. Key categories of impactful careers. https://80000hours.org/career-reviews/.

[4] 80,000 Hours. 2023. Jobs. https://jobs.80000hours.org/.

[5] 80,000 Hours. 2023. Start here: why we're here and how we can help. https://80000hours.org/start-here/.

[6] AGI Safety Fundamentals. 2023. AI Alignment Course - AGI Safety Fundamentals. https://www.agisafetyfundamentals.com/ai-alignment-curriculum.

[7] AI Impacts. 2015. MIRI AI Predictions Dataset. https://aiimpacts.org/miri-ai-predictions-dataset/.

[8] AI Safety Events. 2022. AI Safety Unconference at NeurIPS 2022. https://aisafetyevents.org/events/aisuneurips2022/.

[9] AI Safety Support. 2023. AI Safety Support. https://www.aisafetysupport.org/.

[10] Alignment Forum. 2022. Technological Forecasting. https://www.agisafetyfundamentals.com/ai-alignment-curriculum.

[11] Alignment Forum. 2023. Alignment Forum. https://www.alignmentforum.org/.

[12] Charlotte Alter. 2023. Effective Altruism Promises to Do Good Better. These Women Say It Has a Toxic Culture Of Sexual Harassment and Abuse. https://time.com/6252617/effective-altruism-sexual-harassment/

[13] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. https://doi.org/10.48550/arXiv.1606.06565

[14] Asterisk. 2023. About. https://asteriskmag.com/about.

[15] Embodied Intelligence Group at MIT CSAIL. 2023. Algorithmic Alignment Group. https://algorithmicalignment.csail.mit.edu/.

[16] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. https://web.archive.org/web/20230308041823/https://www.anthropic.com/constitutional.pdf.

[17] Seth Baum. 2023. GCRI Statement on Race and Intelligence. https://gcrinstitute.org/gcri-statement-on-race-and-intelligence/.

[18] SJ Beard and Phil Torres. 2020. Ripples on the Great Sea of Life: A Brief History of Existential Risk Studies. *Available at SSRN 3730000* (2020).

[19] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[20] Ruben Bloom and Oliver Habyka. 2021. Welcome & FAQ! https://www.alignmentforum.org/posts/Yp2vYb4zHXEeoTkJc/welcome-and-faq.

[21] Nick Bostrom. 2002. Existential risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology* 9 (2002). https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c

[22] Nick Bostrom. 2005. Transhumanist values. *Journal of philosophical research* 30, Supplement (2005), 3–14.

[23] Nick Bostrom. 2014. *Superintelligence: paths, dangers, strategies* (first edition ed.). Oxford University Press, Oxford. OCLC: ocn881706835.

[24] Nick Bostrom. 2023. Apology for an Old Email. https://nickbostrom.com/oldemail.pdf.

[25] Nick Bostrom, Haydn Belfield, and Sam Hilton. 2020. Evidence to the UK Parliament Science & Technology Committee's Inquiry on a new UK research funding agency. https://committees.parliament.uk/writtenevidence/9504/html/.

[26] Niel Bowerman. 2019. The case for building expertise to work on US AI policy, and how to do it. https://80000hours.org/articles/us-ai-policy/.

[27] Sam Bowman. 2022. AI Safety and Neighboring Communities: A Quick-Start Guide, as of Summer 2022. https://www.alignmentforum.org/posts/EFpQcBmfm2bFfM4zM/ai-safety-and-neighboring-communities-a-quick-start-guide-ass.

[28] Sam Bowman. 2022. Why I Think More NLP Researchers Should Engage with AI Safety Concerns. https://wp.nyu.edu/arg/why-ai-safety/.

[29] Sam Bowman. 2023. What's the Deal with AI Safety? Motivations and Ways to Get Involved. Talk given at Princeton University, Princeton, NJ.

[30] Jan Brauner. 2022. Your posts should be on arXiv. https://www.alignmentforum.org/posts/TYTEJxzeK3jBMq2TZ/your-posts-should-be-on-arxiv.

[31] André Brock. 2018. Critical Technocultural Discourse Analysis. *New Media & Society* (2018).

[32] Jenna Burrell. 2023. Artificial Intelligence and the Ever-Receding Horizon of the Future. https://techpolicy.press/artificial-intelligence-and-the-ever-receding-horizon-of-the-future/.

[33] Steve Byrnes. 2021. Safety-capabilities tradeoff dials are inevitable in AGI. https://www.alignmentforum.org/posts/tmyTb4bQQi7C47sde/safety-capabilities-tradeoff-dials-are-inevitable-in-agi.

[34] Joseph Carlsmith. 2022. Is Power-Seeking AI an Existential Risk? http://arxiv.org/abs/2206.13353 arXiv:2206.13353 [cs].

[35] Center for AI Safety. 2022. SafeBench. https://benchmarking.mlsafety.org/index.html.

[36] Center for AI Safety. 2022. Trojan Detection Challenge. https://trojandetection.ai/.

[37] Centre for the Governance of AI. 2022. Summer & Winter Fellowships. https://www.governance.ai/post/summer-fellowship-2023.

[38] Centre for the Study of Existential Risk. 2021. Science 20 Report: 'Foresight: Science for Navigating Critical Transitions'. https://www.cser.ac.uk/news/science-20-report-foresight-science-navigating-cri/.

[39] Karin Knorr Cetina. 2007. Culture in global knowledge societies: knowledge cultures and epistemic cultures. *Interdisciplinary Science Reviews* 32, 4 (Dec 2007), 361–375. https://doi.org/10.1179/030801807X163571

[40] Karin Knorr Cetina. 2009. *Epistemic Cultures: How the Sciences Make Knowledge.* Harvard University Press, Cambridge, MA, USA. https://doi.org/10.2307/j.ctvxw3q7f

[41] Brian Christian. 2020. *The alignment problem: Machine learning and human values* (first edition ed.). W.W. Norton & Company, New York, NY.

[42] Paul Christiano. 2022. Prizes for ELK proposals. https://www.alignmentforum.org/posts/QEYWkRoCn4fZxXQAY/prizes-for-elk-proposals.

[43] Paul Christiano, Ajeya Cotra, and Mark Xu. 2021. Eliciting latent knowledge: How to tell if your eyes deceive you. https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit.

[44] Kate Clark. 2023. https://www.theinformation.com/articles/openai-rival-anthropic-raises-funding-at-4-1-billion-valuation

[45] Laurie Clark. 2023. How Silicon Valley doomers are shaping Rishi Sunak's AI plans. https://www.politico.eu/article/rishi-sunak-artificial-intelligence-pivot-safety-summit-united-kingdom-silicon-valley-effective-altruism/.

[46] Sam Clarke, Jess Whittlestone, Matthijs M. Maas, Haydn Belfield, José Hernández-Orallo, and Seán Ó hÉigeartaigh. 2020. Submission of Feedback to the European Commission's Proposal for a Regulation laying down harmonised rules on artificial intelligence. https://www.cser.ac.uk/resources/feedback-european-regulation/.

[47] Claire Colebrook. 2018. *Lives worth living: Extinction, persons, disability.* University of Minnesota Press, United States, 151–171.

[48] Committee on Commerce, Science, and Transportation, United States Senate. 2016. Statement of Greg Brockman, Hearing on "The Dawn of Artificial Intelligence". https://www.commerce.senate.gov/services/files/ae7e9ee3-df1b-4d94-96d1-267ebd206c48.

[49] ConcernedEAs. 2023. Doing EA Better. https://forum.effectivealtruism.org/posts/54vAiSFkYszTWWWv4/doing-ea-better-1.

[50] ACM FAccT Conference. 2023. Statement on AI Harms and Policy. https://facctconference.org/2023/harm-policy.html.

[51] Owen Cotton-Barratt and Toby Ord. 2015. Existential risk and existential hope: definitions. (2015).

[52] Carla Zoe Cremer and Luke Kemp. 2021. Democratising Risk - or how EA deals with critics. https://forum.effectivealtruism.org/posts/gx7BEkoRbctjkyTme/democratising-risk-or-how-ea-deals-with-critics-1.

[53] Carla Zoe Cremer and Luke Kemp. 2021. Democratising risk: In search of a methodology to study existential risk. *arXiv preprint arXiv:2201.11214* (2021).

[54] Carla Zoe Cremer and Luke Kemp. 2022. Democratising Risk: In Search of a Methodology to Study Existential Risk. https://doi.org/10.48550/ARXIV.2201.11214

[55] Crunchbase. 2023. Semafor - Funding, Financials, Valuation & Investors. https://www.crunchbase.com/organization/semafor/company_financials.

[56] Allan Dafoe. 2018. AI Governance: A Research Agenda. Future of Humanity Institute. https://www.governance.ai/research-paper/agenda

[57] AI Now Institute Data & Society. 2021. Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource. https://ainowinstitute.org/AINow-DS-NAIRR-comment.pdf.

[58] DeepMind. 2023. DeepMind About Page. https://www.deepmind.com/about.

[59] Effective Altruism Forum. 2023. Effective Altruism Forum. https://forum.effectivealtruism.org/.

[60] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A Mathematical Framework for Transformer Circuits. urlhttps://transformer-circuits.pub/2021/framework/index.html.

[61] Elicit Forecast. 2023. Elicit Forecast. https://forecast.elicit.org/.

[62] Encultured.AI. 2023. Encultured.AI. https://www.encultured.ai/.

[63] Lee Fan, Ken Klippenstein, and Daniel Boguslaw. 2023. New FTX Filing Pulls Back Curtain on SBF's Massive Influence-Peddling Operation. https://theintercept.com/2023/01/30/ftx-sam-bankman-fried-lobbying-pr/.

[64] Cody Fenwick. 2022. Journalism. https://80000hours.org/career-reviews/journalism/.

[65] Center for AI Safety. 2023. Statement on AI Risk. https://www.safe.ai/statement-on-ai-risk.

[66] Center for Long-Term Resilience. 2022. Annual Report 2021. https://www.longtermresilience.org/post/annual-report-2021.

[67] FTX Future Fund. 2022. Areas of Interest. https://web.archive.org/web/20221104165229/https://ftxfuturefund.org/area-of-interest/artificial-intelligence/

[68] FTX Future Fund. 2022. Apply for Funding. https://web.archive.org/web/20220625173044/https://ftxfuturefund.org/apply/.

[69] Future of Life Institute. 2021. FLI Position Paper on the EU AI Act. https://futureoflife.org/wp-content/uploads/2021/08/FLI-Position-Paper-on-the-EU-AI-Act.pdf?x72900.

[70] Iason Gabriel. 2020. Artificial Intelligence, Values and Alignment. *Minds and Machines* 30, 3 (Sept. 2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[71] Vael Gates. 2022. Announcing the AI Safety Field Building Hub, a new effort to provide AISFB projects, mentorship, and funding. https://forum.effectivealtruism.org/posts/ozm4SpiChfAAAGnw5/announcing-the-ai-safety-field-building-hub-a-new-effort-to.

[72] Matthew Gault and Jordan Pearson. 2023. Prominent AI Philosopher and 'Father' of Longtermism Sent Very Racist Email to a 90s Philosophy Listserv. https://www.vice.com/en/article/z34dm3/prominent-ai-philosopher-and-father-of-longtermism-sent-very-racist-email-to-a-90s-philosophy-listserv.

[73] Timnit Gebru. 2023. Effective Altruism Is Pushing a Dangerous Brand of 'AI Safety'. https://www.wired.com/story/effective-altruism-artificial-intelligence-sam-bankman-fried.

[74] Timnit Gebru. 2023. Eugenics and the Promise of Utopia through AGI. https://www.youtube.com/watch?v=P7XT4TWLzJw

[75] Gov.UK. 2023. PM London Tech Week speech: 12 June 2023. https://www.gov.uk/government/speeches/pm-london-tech-week-speech-12-june-2023.

[76] Gov.UK. 2023. UK to host first global summit on Artificial Intelligence. https://www.gov.uk/government/news/uk-to-host-first-global-summit-on-artificial-intelligence.

[77] Daniel Greene, Anna Lauren Hoffman, and Luke Stark. 2019. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial

Intelligence and Machine Learning. *HICSS 2019* (2019).

[78] Ryan Grim. 2022. How AIPAC and DMFI Outspent the Democratic Insurgency. https://theintercept.com/2022/10/16/democratic-party-progressive-israel-aipac-dmfi/.

[79] Alex Hanna and Emily Bender. 2023. Mystery AI Hype Theater 3000, Episode 5 - Sam Bankman-Fried's Future Fund and Fresh AI Hell. https://videos.trom.tf/w/p/4gykGcMrmHHs7bG2Y6qK9W.

[80] Harvard AI Safety Team. 2023. Harvard AI Safety Team. https://haist.ai/.

[81] Peter Hase. 2021. The Alignment Forum should have more transparent membership standards. https://www.lesswrong.com/posts/dsr9vvcLjyLZGuBnN/the-alignment-forum-should-have-more-transparent-membership.

[82] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2022. Unsolved Problems in ML Safety. https://doi.org/10.48550/arXiv.2109.13916 arXiv:2109.13916 [cs].

[83] Benjamin Hilton. 2022. Preventing an AI-related catastrophe - Problem profile. https://80000hours.org/problem-profiles/artificial-intelligence/

[84] Hilton, Benjamin. 2022. Preventing an AI-related catastrophe. https://80000hours.org/problem-profiles/artificial-intelligence/

[85] The White House. 2023. Red-Teaming Large Language Models to Identify Novel AI Risks. https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/.

[86] House Committee on Science, Space, and Technology. 2023. Full Committee Hearing - Artificial Intelligence: Advancing Innovation Towards the National Interest. https://science.house.gov/2023/6/artificial-intelligence-advancing-innovation-towards-the-national-interest.

[87] Kalley Huang. 2022. Why Did FTX Collapse? Here's What to Know. https://www.nytimes.com/2022/11/10/technology/ftx-binance-crypto-explained.html.

[88] Hugging Face. 2023. stanfordnlp/SHP · Datasets at Hugging Face. https://huggingface.co/datasets/stanfordnlp/SHP

[89] Global Catastrophic Risk Institute. 2020. RFI Response: National Artificial Intelligence Research and Development Strategic Plan— White House Office of Science and Technology Policy. https://gcrinstitute.org/files/2022-03-OSTP-GCRI.pdf.

[90] Geoffrey Irving and Amanda Askell. 2019. AI Safety Needs Social Scientists. *Distill* 4, 2 (Feb. 2019), e14. https://doi.org/10.23915/distill.00014

[91] Khari Johnson. 2023. Welcome to the Museum of the Future AI Apocalypse. https://www.wired.com/story/welcome-to-the-museum-of-the-future-ai-apocalypse/. *WIRED* (March 2023).

[92] Cecilia Kang and Adam Satariano. 2023. As A.I. Booms, Lawmakers Struggle to Understand the Technology. https://www.nytimes.com/2023/03/03/technology/artificial-intelligence-regulation-congress.html

[93] Holden Karnosky. 2021. Forecasting Transformative AI, Part 1: What Kind of AI? https://www.lesswrong.com/posts/mMDNeNfEKCKPjJTNC/forecasting-transformative-ai-part-1-what-kind-of-ai

[94] Yarden Katz. 2020. *Artificial whiteness: politics and ideology in artificial intelligence.* Columbia University Press, New York.

[95] Makena Kelly. 2023. Meta, Google, and OpenAI reveal their safety plans following White House summit. https://www.theverge.com/2023/7/21/23803244/meta-google-openai-microsoft-artificial-intelligence-ai-white-house-commitments.

[96] Luke Kemp, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Seán Ó hÉigeartaigh, Jane Leung, and Zoe Creme. 2019. Advice to UN High-level Panel on Digital Cooperation. https://www.cser.ac.uk/news/advice-un-high-level-panel-digital-cooperation/.

[97] Jan H. Kirchner, Logan Smith, Jacques Thibodeau, Kyle McDonell, and Laria Reynolds. 2022. Researching Alignment Research: Unsupervised Analysis. https://doi.org/10.48550/arXiv.2206.02841 arXiv:2206.02841 [cs].

[98] Daniel Kokotajlo. 2022. AI takeover tabletop RPG: "The Treacherous Turn". https://www.lesswrong.com/posts/b5EqwQZw7ww2K28Ki/ai-takeover-tabletop-rpg-the-treacherous-turn. *LessWrong* (November 2022).

[99] Larks. 2021. 2021 AI Alignment Literature Review and Charity Comparison. https://www.alignmentforum.org/posts/C4tR3BEpuWviT7Sje/2021-ai-alignment-literature-review-and-charity-comparison

[100] LessWrong. 2019. AI Forecasting online workshop. https://www.lesswrong.com/events/wXBC4PMePfjKFFHYJ/ai-forecasting-online-workshop.

[101] LessWrong. 2021. Forecasting Compute - Transformative AI and Compute [2/4]. https://www.lesswrong.com/posts/sHAaMpdk9FT9XsLvB/forecasting-compute-transformative-ai-and-compute-2-4.

[102] LessWrong. 2023. LessWrong. https://www.lesswrong.com/.

[103] Gideon Lewis-Kraus. 2022. The Reluctant Prophet of Effective Altruism. https://www.newyorker.com/magazine/2022/08/15/the-reluctant-prophet-of-effective-altruism.

[104] William MacAskill. 2019. *The Definition of Effective Altruism.* Oxford University Press, Oxford, UK, 10–28. https://doi.org/10.1093/oso/9780198841364.003.0001

[105] Dylan Matthews. 2018. Future Perfect, explained. https://www.vox.com/future-perfect/2018/10/15/17924288/future-perfect-explained.

[106] Dylan Matthews. 2022. How effective altruism went from a niche movement to a billion-dollar force. https://www.vox.com/future-perfect/2022/8/8/23150496/effective-altruism-sam-bankman-fried-dustin-moskovitz-billionaire-philanthropy-crytocurrency.

[107] Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022. The Inverse Scaling Prize. https://github.com/inverse-scaling/prize

[108] Metaculus. 2023. Metaculus. https://www.metaculus.com.

[109] Michael. 2020. Database of existential risk estimates. https://forum.effectivealtruism.org/s/MGRxKfQsZXiw9JHwD/p/JQQAQrunyGGhzE23a.

[110] MineRL. 2022. MineRL BASALT competition. https://minerl.io/basalt/.

[111] MIT AI Alignment. 2023. MIT AI Alignment. https://www.mitalignment.org/.

[112] ML Safety. 2022. The Autocast Competition. https://forecasting.mlsafety.org/.

[113] ML Safety. 2022. ML Safety Social at NeurIPS 2022. https://www.mlsafety.org/social.

[114] ML Safety. 2022. The Moral Uncertainty Research Competition. https://moraluncertainty.mlsafety.org/.

[115] Benjamin Mullin and David Yaffe-Bellany. 2023. Media Start-Up Semafor Plans to Buy Out Sam Bankman-Fried's Investment. https://www.nytimes.com/2023/01/18/business/semafor-sam-bankman-fried-investment.html.

[116] Arvind Narayanan and Sayash Kapoor. 2023. AI Snake Oil newsletter. https://aisnakeoil.substack.com/.

[117] National Science Foundation. 2023. Safe Learning-Enabled Systems Program Solicitation, NSF 23-562. https://www.nsf.gov/pubs/2023/nsf23562/nsf23562.pdf.

[118] Nature. 2023. Stop talking about tomorrow's AI doomsday when AI poses risks today. https://www.nature.com/articles/d41586-023-02094-7.

[119] NeurIPS. 2022. 2022 NeurIPS ML Safety Workshop. https://web.archive.org/web/20220713190527/https://neurips2022.mlsafety.org/.

[120] NeurIPS. 2022. 2022 NeurIPS ML Safety Workshop. https://neurips2022.mlsafety.org/.

[121] NYU Alignment Research Group. 2023. The NYU Alignment Research Group. https://wp.nyu.edu/arg/.

[122] OECD Education and Skills Today. 2019. What does artificial intelligence mean for values and ethics? https://oecdedutoday.com/artificial-intelligence-education-values-ethics-oecd-forum-ai/.

[123] Foundations of Cooperative AI Lab. 2023. FOCAL@CMU. https://www.cs.cmu.edu/~focal/.

[124] Future of Life Institute. 2019. Annual Report. https://futureoflife.org/wp-content/uploads/2019/02/2018-Annual-Report.pdf?x51579.

[125] Future of Life Institute. 2023. Pause Giant AI Experiments: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[126] Future of Life Institute podcast. 2022. Daniela and Dario Amodei on Anthropic. https://futureoflife.org/podcast/daniela-and-dario-amodei-on-anthropic/.

[127] National Institute of Standards and Technology. 2022. Building the NIST AI Risk Management Framework: Workshop #3, AIRM2.3 How to Measure AI Risk across the AI Lifecycle. https://www.nist.gov/news-events/events/2022/10/building-nist-ai-risk-management-framework-workshop-3.

[128] National Institute of Standards and Technology. 2022. Building the NIST AI Risk Management Framework: Workshop 3, Exactly What Is an AI RMF Profile? https://www.nist.gov/news-events/events/2022/10/building-nist-ai-risk-management-framework-workshop-3.

[129] National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[130] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context Learning and Induction Heads. urlhttps://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[131] European Conference on Computer Vision. 2022. ECCV 2022 Workshop on Adversarial Robustness in the Real World. https://eccv22-arow.github.io/.

[132] US Senate Committee on the Judiciary. 2023. Oversight of A.I.: Rules for Artificial Intelligence - Subcommittee Hearing. https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence.

[133] Open Philanthropy. 2022. Effective Altruist Communications Fellowship — Summer 2022 Fellowship. https://www.openphilanthropy.org/grants/effective-altruist-communications-fellowship-summer-2022-fellowship/.

[134] Open Philanthropy. 2022. Open Philanthropy Technology Policy Fellowship. https://www.openphilanthropy.org/open-philanthropy-technology-policy-fellowship/.

[135] OpenAI. 2023. Planning for AGI and beyond. https://openai.com/blog/planning-for-agi-and-beyond.

[136] Toby Ord, Angus Mercer, and Sophie Dannreuther. 2021. Future Proof: The Opportunity to Transform the UK's Resilience to Extreme Risks. https://www.longtermresilience.org/futureproof.

[137] Alexander Osipovich. 2021. This Vegan Billionaire Disrupted the Crypto Markets. Stocks May Be Next. https://www.wsj.com/articles/this-vegan-billionaire-disrupted-the-crypto-markets-stocks-may-be-next-11618565408.

[138] Orestis Papakyriakopoulos, Severin Engelmann, and Amy Winecoff. 2023. Upvotes? Downvotes? No Votes? Understanding the relationship between reaction mechanisms and political discourse on Reddit. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Hamburg, Germany. https://doi.org/10.1145/3544548.3580644

[139] Parallel, LLC. 2019. AI Forecasting Dictionary. https://www.agisafetyfundamentals.com/ai-alignment-curriculum.

[140] Open Philanthropy. 2022. Potential Risks from Advanced Artificial Intelligence. https://www.openphilanthropy.org/focus/potential-risks-advanced-ai/

[141] Kelsey Piper. 2022. Sam Bankman-Fried tries to explain himself. https://www.vox.com/future-perfect/23462333/sam-bankman-fried-ftx-cryptocurrency-effective-altruism-crypto-bahamas-philanthropy.

[142] Kelsey Piper. 2022. What We Owe The Future. https://asteriskmag.com/issues/1/review-what-we-owe-the-future.

[143] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. 2022. A Human Rights-Based Approach to Responsible AI. In *2022 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery, Arlington, VA, USA. https://doi.org/10.48550/arXiv.2210.02667

[144] ProPublica. 2022. Bankman-Fried Family Donates $5 Million to ProPublica. https://web.archive.org/web/20221101142355/https://www.propublica.org/atpropublica/bankman-fried-family-donates-5-million-to-propublica.

[145] ProPublica. 2022. ProPublica to Return Grant Funded by Bankman-Fried Family. https://www.propublica.org/atpropublica/bankman-fried-family-donates-5-million-to-propublica.

[146] Inioluwa Deborah Raji and Roel Dobbe. 2020. Concrete Problems in AI Safety, Revisited. https://drive.google.com/file/d/1Re_yQDNFuejoqjZloTgQpILosDGtt5ei/view

[147] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 959–972. https://doi.org/10.1145/3531146.3533158

[148] Sam Reynolds. 2022. FTX's 'Effective Altruism' Future Fund Team Resigns. https://www.coindesk.com/business/2022/11/10/ftxs-effective-altruism-future-fund-team-resigns/.

[149] Ruby. 2019. A Brief History of LessWrong. https://www.lesswrong.com/posts/S69ogAGXcc9EQjpcZ/a-brief-history-of-lesswrong.

[150] Sigal Samuel. 2022. Effective altruism's most controversial idea. https://www.vox.com/future-perfect/23298870/effective-altruism-longtermism-will-macaskill-future

[151] Sigal Samuel. 2023. How to reform effective altruism after Sam Bankman-Fried. https://www.vox.com/future-perfect/23564571/effective-altruism-sam-bankman-fried-holden-karnofsky-ai.

[152] Theodore Schleifer. 2021. How a crypto billionaire decided to become one of Biden's biggest donors. https://www.vox.com/recode/2021/3/20/22335209/sam-bankman-fried-joe-biden-ftx-cryptocurrency-effective-altruism.

[153] Hendrik R. Schopmans. 2022. From Coded Bias to Existential Threat: Expert Frames and the Epistemic Politics of AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Oxford United Kingdom, 627–640. https://doi.org/10.1145/3514094.3534161

[154] Aura-Elena Schussler. 2019. Transhumanism as a new techno-religion and personal development: In the framework of a future technological spirituality. *Journal for the Study of Religions and Ideologies* 18, 53 (2019), 92–106.

[155] Joshua Schuster and Derek Woods. 2021. *Calamity theory: three critiques of existential risk*. University of Minnesota Press, Minneapolis London.

[156] Prometheus Science. 2022. Results. https://prometheus.science/elk-results.

[157] Prometheus Science. 2023. What is Elk? https://prometheus.science/what-is-elk.

[158] Ze Shen Shen. 2022. A newcomer's guide to the technical AI safety field. https://www.alignmentforum.org/posts/5rsa37pBjo4Cf9fkE/a-newcomer-s-guide-to-the-technical-ai-safety-field.

[159] Jacob Steinhardt. 2017. Forecasting ML Benchmarks in 2023. https://www.alignmentforum.org/posts/arveXgFbJwascKtQC/forecasting-ml-benchmarks-in-2023.

[160] Jacob Steinhardt. 2022. AI Forecasting: One Year In. https://bounded-regret.ghost.io/ai-forecasting-one-year-in/.

[161] The Ezra Klein Show. 2022. The Sentences That Could Change the World - and Your Life. https://www.nytimes.com/2022/08/09/opinion/ezra-klein-podcast-will-macaskill.html.

[162] The Tarbell Fellowship. 2022. The Tarbell Fellowship. https://www.tarbellfellowship.org/.

[163] Benjamin Todd. 2019. Advice on how to read our advice. https://80000hours.org/key-ideas/advice-on-how-to-read-our-advice.

[164] Benjamin Todd. 2021. Government and policy in an area relevant to a top problem. https://80000hours.org/articles/government-policy/.

[165] Training for Good. 2022. EU Tech Policy Fellowship 2023. https://www.trainingforgood.com/europe-tech-policy.

[166] U.S. Senate Committee on Commerce, Science, and Transportation. 2022. Written Testimony of Jack Clark. https://www.commerce.senate.gov/2022/9/securing-u-s-leadership-in-emerging-compute-technologies.

[167] Vox Staff. 2022. Support Future Perfect. https://www.vox.com/2020/1/7/21020439/support-future-perfect.

[168] Thomas W and Dan H. 2022. Perform Tractable Research While Avoiding Capabilities Externalities [Pragmatic AI Safety #4]. https://forum.effectivealtruism.org/posts/WmrCQTkTgDuk5RhCP/perform-tractable-research-while-avoiding-capabilities.

[169] Jess Whittlestone, Shahar Avin, Katherine Collins, Jack Clark, and Jared Mueller. 2022. Future of compute review - submission of evidence. https://www.longtermresilience.org/post/future-of-compute-review-submission-of-evidence.

[170] Jess Whittlestone, Diane Cooke, Shahar Avin, Kayla Matteucci, Seán Ó hÉigeartaigh, Haydn Belfield, and Markus Anderljung. 2022. The UK Defence AI Strategy: ensuring safe and responsible use of AI. https://www.longtermresilience.org/post/the-uk-defence-ai-strategy-ensuring-safe-and-responsible-use-of-ai.

[171] Jess Whittlestone and the 80,000 Hours team. 2017. Epistemics and institutional decision-making. https://80000hours.org/problem-profiles/improving-institutional-decision-making/.

[172] Robert Wiblin. 2016. A framework for comparing global problems in terms of expected impact. https://80000hours.org/articles/problem-framework/.

[173] Eliezer Yudkowski. 2001. The Singularitarian Principles Version 1.0.2. https://web.archive.org/web/20081229202843/http://yudkowsky.net:80/obsolete/principles.html

# NOTES

[1] While our methodology involved tracing flows of capital from funders to institutions or individuals, we emphasize that funding itself is not the basis of our analysis in this paper. Rather, we relied on funding streams as a useful tool for identifying important actors.

[2] In a variation on the prediction approach to forecasting, MIRI's The Uncertain Future tool uses their own collated data to model a user's beliefs about when "human-level AI" will happen. See http://theuncertainfuture.com/ and accompanying report, http://intelligence.org/files/ChangingTheFrame.pdf

[3] These institutions are unique insofar as all three steadily produce research publications, with DeepMind being the least product-oriented of the three firms. Thus, the topical focus and methods of knowledge production within these organizations may not generalize to other research organizations with different company structures or business priorities.

[4] Despite this epistemic community's claims to want to support a non-rivalrous geopolitical landscape for AI development, we found that tech firm representatives often raise talking points that work against this stated objective. For example, we reviewed three Congressional testimonies from 2016 and 2022, one from Greg Brockman of OpenAI, and two from Jack Clark (in his roles at OpenAI and Anthropic). Both men spoke to the need for the United States to retain its position as a world leader in AI in order to "define [AI's] culture and values" [48] as well as to reap what they characterize as AI's inevitable economic benefits. This position formed the backdrop to consistent references to how other countries are heavily investing in AI development, and how, for instance, China is "far ahead of" the US when it comes to AI capabilities used for re-identifying individuals from surveillance footage [166]. Elsewhere, other Anthropic employees have spoken about not wanting to fuel an "AI race" dynamic [126]. Yet it can be difficult to separate narratives of geopolitical competition from these testimonies' comparisons of US investment and advances in AI to those of other countries; their emphases on public measurements of AI systems' progress; and their references to international contests on different types of AI tasks as a way of gauging other countries' AI advancements.

Received 30 September 2023; revised TBD; accepted TBD