Research Agenda for Sociotechnical Approaches to AI Safety

Samuel Curtis[1], **Ravi Iyer**[2], **Cameron Domenico Kirk-Giannini**[3], **Victoria Krakovna**[4], **David Krueger**[5], **Nathan Lambert**[6], **Bruno Marnette**[7], **Colleen McKenzie**[*7], **Julian Michael**[8], **Evan Miyazono**[9], **Noyuri Mima**[10], **Aviv Ovadya**[12], **Luke Thorburn**[12], and **Deger Turan**[7]

[1]The Future Society
[2]University of Southern California
[3]Rutgers University
[4]Future of Life Institute
[5]University of Cambridge
[6]Allen Institute for AI
[7]AI Objectives Institute
[8]New York University
[9]Atlas Computing
[10]Future University Hakodate
[11]AI & Democracy Foundation
[12]King's College London

Abstract

As the capabilities of AI systems continue to advance, it is increasingly important that we guide the development of these powerful technologies, ensuring they are used for the benefit of society. Existing work analyzing and assessing risks from AI spans a broad and diverse range of perspectives, including some which diverge enough in their motivations and approaches that they disagree on priorities and desired solutions. Yet we find significant overlap among these perspectives' desire for beneficial outcomes from AI deployment, and significant potential for progress towards such outcomes in the examination of that overlap. In this paper we explore one such area of overlap: we discuss areas of AI safety work that could benefit from sociotechnical framings of AI, which view AI systems as embedded in larger sociotechnical systems, and which explore the potential risks and benefits of AI not just as aspects of these new tools, but as possibilities for the complex interactions between humans and our technologies. We present a collection of proposals we believe to be promising directions for including sociotechnical approaches in the pursuit of safe and beneficial AI, demonstrating the potential value of such approaches in addressing the harms, risks, and benefits of current and future AI systems.

[*]Corresponding author: colleen@objective.is

Contents

1 Introduction

As the capabilities of AI systems continue to advance, it is increasingly important that we guide the development of these powerful technologies, ensuring they are used for the benefit of society. High-profile open letters from groups like the Future of Life Institute (Future of Life Institute, 2023), members of the ACM FAccT Community (Castillo et al., 2023), and the Center for AI Safety (Center for AI Safety, 2023), and papers from expert groups (e.g. Bengio et al., 2023; Chan et al., 2023; Weidinger et al., 2023), have called attention to risks from advanced AI and stressed the importance of ensuring that AI systems work in service of human values. Existing work analyzing and assessing risks from AI spans a broad and diverse range of perspectives, including some which diverge enough in their motivations and approaches that they disagree on many questions of priorities and desired solutions—which in the past has led to conflict between these groups (Bristow & Thorburn, 2023; Prunkl & Whittlestone, 2020). Yet we find significant overlap among perspectives' desire for beneficial outcomes from AI deployment, and significant potential for progress towards such outcomes in the exploration of that overlap. To this end, we suggest a collection of research areas that extend AI safety work to incorporate sociotechnical framings of the development of AI: recognizing AI systems as embedded in larger sociotechnical systems, and exploring the potential risks and benefits of AI not just as aspects of these new tools, but as possibilities for the complex interactions between humans and our technologies in systems that include both (Lazar & Nelson, 2023; Selbst et al., 2019). Research in this area includes both work that centers AI's growing influence on these systems, such as risk mitigation for potential superintelligence, but also work that focuses on addressing societal effects of both current and future powerful AI, such as fairness and desirability of impacts, transparency of algorithmic details, and respect for basic human rights (Prabhakaran, Mitchell, et al., 2022).

This paper offers a collection of proposals we believe are promising directions for applying more sociotechnical approaches to questions of developing safe and beneficial AI, drawn from discussions exploring how to integrate different perspectives and priorities for mitigating AI risk. We believe the proposals below exemplify how existing avenues of AI safety research can incorporate considerations of social context and impact, and suggest the potential value of such an approach in addressing the harms, risks, and benefits of current and future AI systems.

1.1 Background

While work on steering AI development towards positive outcomes has recently gained visibility, it builds on substantial prior analysis of how increasingly complex algorithmic systems interface with the populations they affect. These existing lines of inquiry are varied in their foci and approaches, a diversity reflected by the many different contemporary perspectives on AI—and a result of the complex nature of defining what constitutes a safe, fair, or ethical system. In this section, we briefly describe a few areas of existing research that are most relevant to sociotechnical approaches to AI safety, in their consideration of these subjective questions in their approaches to studying new technology.

Early research into AI systems' effects on society has investigated fairness, privacy, transparency, and responsibility (Boyd & Crawford, 2012; Cooper & Vidan, 2022; Hardt, 2014), including issues of bias and socially impactful errors in a variety of algorithmic systems: in the latent representations of language models (Bolukbasi et al., 2016), in image recognition algorithms (Buolamwini & Gebru, 2018), and in predictive systems such as the COMPAS recidivism algorithm (Angwin et al., 2016). This work extends to examinations of the social contexts that create these new technologies, such as new and emerging sources of labor (Irani & Silberman, 2013; Miceli et al., 2022), explorations of what constitutes fairness in the context of algorithmic and other quantitative systems (Dwork et al., 2012; Friedler et al., 2021; Mitchell et al., 2021), and strategies for identifying and mitigating unjust outcomes and impacts of such systems—including evaluating the impacts of well-intentioned interventions (Raji & Buolamwini, 2019; Raji et al., 2020). The ACM conferences on Fairness, Accountability, and Transparency (FAccT) and AI, Ethics, and Society (AIES), founded in 2014 and 2018 respectively, coalesced around these approaches and continue to explore these topics, often described succinctly as the "AI Ethics" perspective.

Other social scientific fields, whose objects of study extend beyond just algorithmic systems, also inform analyses of contemporary AI. Of particular relevance is Science and Technology Studies

(STS), which emerged as a field in the mid 20th century, and has examined the relationship between societies and technologies much less powerful than modern AI. STS scholars highlight the social and historical embeddedness of technology (Dourish, 2001), arguing that no technology is a neutral tool, but rather a product of specific social, political, and economic forces (Winner, 1980), and thus will necessarily serve some values—explicit or implicitly—over others (Nissenbaum, 2001). Social theorists' development of systems-theoretic frameworks like Actor-Network Theory (Law & Hassard, 1999) has laid groundwork for evaluating tools within the sociotechnical systems they participate in, and feminist approaches to STS explore the ways technology can reify contentious social norms (Wajcman, 1991). Researchers in social scientific fields beyond STS examine AI's potential impacts through other lenses: economists analyze and predict the effects of automation on labor markets (Acemoglu & Restrepo, 2020; Frey & Osborne, 2017), including markets for labor to produce training data for machine learning algorithms (Irani & Silberman, 2013); and political scientists investigate the construction of national narratives about AI (Kreps et al., 2022) and the effect of AI-generated media on partisan beliefs (Katzenbach & Bareis, 2022).

Over the last 30 years, parallel investigations have pursued more abstract safety challenges from advanced AI capabilities. Premised on the potential for technology to create AI (or augmentations to human intelligence) that surpasses current human capabilities (Markoff, 2009)—potentially leading to an exponential takeoff in self-improving intelligence (Vinge, 1993)—this work has explored the potential risks of AI systems more intelligent than human beings, including loss of human control over societal decision-making, and runaway optimization towards undesirable targets (Bales et al., 2024). This research includes questions of governance for transformative future technologies, including which policies, regulatory bodies, and best practices might enable successful decision-making about their use (Anderljung et al., 2023; Dafoe, 2017), and how AI might be aligned to human values, given the difficulty of formally specifying those nuanced, changeable, and often uncertain values. Together, these approaches constitute an approach usually referred to as "AI Safety."

1.2 Process

In this paper, we have sought to propose research directions informed by this past work, with an intent to describe how certain areas of research core to the AI Safety approach might incorporate considerations of social context that have been previously explored by other perspectives. The research agenda below was conceived in conversations at events last year[2] between researchers and practitioners interested in increasing collaboration between groups whose approaches to this work are compatible, but usually independent. Building on those conversations, this paper collects proposals that participants identified as relevant to the overlap of multiple perspectives on AI, especially those that extend existing AI Safety work to incorporate the knowledge and concerns of AI Ethics and STS perspectives.

We have excluded some proposals from these original conversations, including those that did not involve novel research (such as creating new public institutions, or changing corporate structures), those that received minimal support among participants, and those for which no contributors were available to discuss the research direction in depth. The result is a research agenda we believe contributes a broad but not exhaustive discussion of promising approaches to shared goals of mitigating the risks of AI while steering its impacts towards societal good.

2 Proposals

The proposals below describe candidates for collaborative research directions that build on the past work and interests of multiple perspectives on AI. We separate the list into three areas of focus that emerged in the discussions that inspired this paper:

1. **Effective AI Governance**: Research to inform and develop institutions tasked with regulation, policymaking, and diplomacy related to the governance of AI
2. **Understanding AI Systems**: Research to improve our ability to understand, assess, and predict the mechanisms and behaviors of AI
3. **Improving AI's Social Impacts**: Research that directly addresses known impacts of AI systems on the societies they affect

[2]Most discussion took place at the 2023 Berkeley CHAI workshop (https://humancompatible.ai/chai2023)

2.1 Effective AI Governance

2.1.1 Develop preference aggregation and social welfare proposals for a sociotechnical world

Deployments of AI impact diverse stakeholders, and in particular, the design choices embedded in widely-used foundation models will plausibly impact a significant majority of the global population. To reflect the interests of such diverse groups, we will sometimes need processes to aggregate preferences and formalize social objectives in ways that are sufficiently concrete to be implementable in code. Such formal social choice methods, where they are used, cannot answer core normative questions about *standing* (who has a say), *measurement* (how preferences are measured), and *aggregation* (how preferences are aggregated) (Baum, 2020), and will need to be combined with procedural and institutional governance systems capable of flexibly supporting complex negotiations between diverse stakeholders (see Proposal 2.1.4). But formal preference aggregation methods can, if implemented well as part of broader governance systems, help facilitate broad consultation and mitigate concentrations of power.

Much of this work is interdisciplinary, drawing on computational social choice, participatory design, machine learning, democratic theory, moral and political philosophy, psychology, and economics. Examples and precedents include the Moral Machine project (Noothigattu et al., 2018), WeBuildAI (M. K. Lee et al., 2019), jury learning (Gordon et al., 2022), democratic fine-tuning via a moral graph (Edelman & Klingefjord, 2023a, 2023b), digital juries (Fan & Zhang, 2020; Pan et al., 2022), and Collective Constitutional AI (Anthropic, 2023b; Collective Intelligence Project, 2023). Most of these draw on the much deeper literature of social choice mechanisms which could be applicable to AI (Brandt et al., 2016).

These and similar mechanisms need to be critically evaluated and strengthened via feedback from diverse perspectives (e.g., Etienne, 2021; Seaver, 2021).

2.1.2 Design effective liability regimes that are widely understood, clarified, and codified in law

Liability, in the context of AI, refers to the legal responsibility of entities for the consequences and risks associated with the development, deployment, and operation of AI systems. This domain of liability intersects with various areas of law, including intellectual property, data protection, consumer rights, and tort law (Kingston, 2018), each of which should be considered in developing a comprehensive legal framework governing AI systems and their societal impacts.

Much research in this domain already exists. However, due to the novel concerns involved in determining liability for the actions of increasingly complex and agentic algorithms (Chan et al., 2023), further research is necessary to characterize effective legal frameworks for this rapidly changing landscape.

One segment of research focuses on understanding the status quo: the applicability of existing laws to outcomes involving AI systems, the extent of current law enforcement, how regulatory agencies exercise existing authority (Henderson, 2023) and communicate their policies (Atleson, 2023), and the impact of regulators' behaviors on those of regulated entities. Another segment of research aims to anticipate how regulators might apply or enforce existing laws, such as those in copyright, data protection, consumer protection, product liability (Leong & Kumarasamy, 2023; Villasenor, 2019), corporate indemnification (Arnold, 2023), and export controls (Brown, 2023), based on historical legal precedents as well as contemporary legal concepts, such as fair use, and the broader societal impacts resulting from the use of AI systems (Madiega, 2023).

A third segment of research would use empirical data to develop technical solutions and implement policies—through advocacy campaigns (AI Now Institute et al., 2023; Moës & Ryan, 2023) or judicial activism (Center for AI and Digital Policy, 2023)—that enable governments to hold entities accountable for their decisions and the risks they impose on society. Technical solutions include watermarks, audit logs, and other methods that help establish causality when harmful or illegal outcomes arise (Madiega, 2023). Policy solutions include regulation of AI development and deployment, and clarification of legal standards. Ex ante requirements imposed on developers and deployers of AI systems would define in advance what standards are required to limit liability when harm occurs (Future Society, 2021), and ex post policies would define methods for monitoring the unpredictable individual and societal impacts of AI systems (Farber, 2022; Transatlantic Reflection Group on

Democracy and the Rule of Law in the Age of "Artificial Intelligence", 2023). Statutory amendments could clarify the burden and standards of proof required to establish causality for harmful or illegal outcomes resulting from AI systems (Madiega, 2023; Moes, 2021).

Altogether, a multifaceted approach to evaluating liability is necessary for understanding the new and complex considerations of responsibility for the harms of AI systems, and for creating a legal environment that safeguards societal welfare by addressing current gaps in attribution and liability, and by informing the development of legal and regulatory mechanisms that are robust to future developments in AI.

2.1.3 Design experiments to empirically study what deliberation processes are most effective for coming to robust agreement on AI governance questions

While AI systems are implemented in code, rich, qualitative, context-specific goals such as fairness are difficult to formalize in ways that apply universally (Selbst et al., 2019), and there is a need for social institutions that can resolve issues arising from AI objective functions and deployments when they are at odds with what people want (Dobbe et al., 2020; Hadfield-Menell & Hadfield, 2019). Many authors have suggested that deliberation (Carson, 2017; Landemore & Page, 2015) can help improve the quality of AI governance and surface common ground (Pinka, 2021), perhaps forming part of what might be called 'society-in-the-loop AI' or an 'algorithmic social contract' (Rahwan, 2018).

Deliberation can support AI governance (Ovadya, 2021, 2023b), but AI can also support deliberation. An emerging set of proposals suggests AI can help find robust agreement among humans with diverse preferences (Bakker et al., 2022; Fish et al., 2023; Ovadya, 2023a; Small et al., 2023), improve the assignment of participants to deliberative groups (Barrett et al., 2023), and might enable new forms of democracy (Grandi, 2018; Kahng et al., 2019; Tantum, 2023). Deliberative processes such as citizen assemblies and deliberative polls are increasingly being conducted or funded by large technology companies (Meta: Ovadya, 2023c; OpenAI: Alignment Assemblies, Recursive Public, Konya et al., 2023; Zaremba et al., 2023; Anthropic: Anthropic, 2023b; Collective Intelligence Project, 2023), and an increasing set of algorithmic tools are available to support different stages of deliberative processes (Polis, Talk to the City).

To translate this work into substantive improvements in AI governance and avoid 'democracy washing', we need to (1) investigate what the normative desiderata for deliberative processes should be, (2) design experiments to empirically evaluate processes against these desiderata, and (3) share the findings in a format that is useful to practitioners and policymakers. Nascent work proposes 'process cards' and 'run reports' for comparing processes (Ovadya, 2023c; AI & Democracy Foundation), but there are many open questions. Most importantly, we need to (4) implement these deliberative processes in a way that gives them real power and control over AI governance decisions.

2.1.4 Research transnational AI governance systems that enable complex democratic decision-making about AI systems

AI systems can have extremely significant transnational impacts, and their governance processes may need to be correspondingly transnational. We have already seen this with the impacts of YouTube (Yesilada & Lewandowsky, 2022) and Meta's recommender systems (Amnesty International, 2022b) and this is likely to be increasingly critical as AI advances continue. To address this challenge will likely require a mix of both existing and new processes to ensure decisions are made with the necessary speed, scale and buy-in for effective responses (Ovadya, 2023d).

We thus need research to design a set of interacting processes and structures for democratic governance, with appropriate checks and balances, that can be used to collectively govern and control AI systems—in coordination with, or as a complement to, existing governance systems. One can think of such a system as roughly analogous to a national constitution in form, with components including aggregation processes and deliberative processes, but also potentially specific delegated roles including executive, judicial, oversight, and administrative bodies that operate either in perpetuity (e.g., a board, a standard legislature), or on demand (e.g., jury duty, citizens' assemblies) (Ovadya, 2024).

To be fully end-to-end, such decision-making systems may also need to be coupled with AI systems themselves, potentially through technical alignment innovations, or related approaches (such as the Open Agency Model; see Drexler, 2023. In contrast to AI alignment as typically envisioned, such

end-to-end systems might go beyond "preference aggregation," e.g. granting specific roles and rights such as veto power, or the ability to engage governance processes regarding decisions around AI development, deployment, or behavior. This might resemble current institutional decision-making processes more than it would aggregating preferences into a single utility function (although it is likely to include aggregation processes as subprocesses, just as elections are sub-processes within the US constitutional order).

We need research to identify approaches for developing and evaluating these systems, not only in simulation, but through real-world pilots that are increasingly end-to-end and with increasing levels of institutionalized deployment and empowerment.

2.1.5 Develop processes for requiring feedback on the fairness and safety of powerful AI systems from a diverse and inclusive set of stakeholders before they are deployed

As powerful AI systems become ever more integrated into our economic, political, and social lives, they exert increasing control over the experiences and life prospects of individuals from a variety of backgrounds. The expanding reach of powerful AI systems has engendered a corresponding increase in the number of people who have a stake in how those systems are deployed. Yet the groups who develop and deploy powerful AI systems are generally not representative of the wide range of individuals whose lives will ultimately be shaped by those systems (Lazar & Nelson, 2023).

This situation is undesirable in at least two respects. First, considerations of justice suggest that those whose experiences and life prospects are shaped by a technology should have a say in how that technology is deployed. An arrangement in which powerful AI systems are developed and deployed without oversight or input from the wide range of groups they will affect is one in which the developers of those systems are likely to lack a complete understanding of the harms they might cause if deployed. Second, members of a group are more likely to know whether and how the deployment of a given technology will affect that group, especially when the group is disadvantaged or marginalized. This may be because they have had personal experiences of the ways in which similar technologies have affected their group, because they have access to testimony from other group members who have had such personal experiences, or because they are more motivated than non-members to acquire knowledge relevant to the question of how the deployment of technologies will affect their group (P. H. Collins, 1990; Dror, 2023). In cases where harms are evident within particular groups before affecting larger populations, members of those groups may have special insight into the causes of those harms; this insight may help mitigate unanticipated risks (Hendrycks & Mazeika, 2022). Members of a group are also less likely than non-members to be subject to certain kinds of barriers to acquiring knowledge about that group (Mills, 2007). Integrating feedback on the fairness and safety of powerful AI systems from a wide variety of different groups, with special attention given to the perspectives of historically disadvantaged and marginalized groups, should accordingly be expected to lead to more accurate beliefs about the possible impacts of those systems. A concern for building safe and fair AI systems leads directly to a concern for consulting a wide range of groups regarding the safety and fairness of those systems prior to deploying them.

Incorporating feedback from a wider variety of perspectives is best achieved through a process that aggregates views without losing their nuance, and regulatory proposals should pay attention to which kinds of diversity are represented. Candidate groups whose feedback would be socially significant, based on known harms of existing AI, include: women (Bolukbasi et al., 2016) racial and ethnic minorities (Buolamwini & Gebru, 2018), people with a variety of sexual orientations and gender identities (Keyes, 2018), the elderly, people with disabilities, people who are socioeconomically marginalized, and people from the Global South. While it is possible that some corporate or governmental actors might voluntarily adopt a policy of soliciting feedback from a diverse set of stakeholders before deploying powerful AI systems, it is more likely that regulatory requirements will be required—and understanding what processes are most effective for this type of oversight will inform such requirements.

2.1.6 Identify key support for whistleblowers at AI companies and research labs

Internal reports of institutional wrongdoing have been essential to public awareness of unsafe and unjust actions, practices, and decisions, and may be especially important in the context of AI due to the high stakes and dynamics of this sector. Development of such systems involves high potential impact, complex technical detail, and private-sector progress covered by non-disclosure agreements—and

because technological progress has outpaced legislation to govern newly capable AI, most companies in the sector are largely self-regulated by internal standards (Hadfield & Clark, 2023). As a result, protecting AI whistleblowers will likely require updated protections—possibly with more detail specific to these dynamics—beyond the general whistleblower acts and directives many countries have already adopted[3]. Further, the need for fair and representative AI systems that encode broadly agreeable values means that issues of workplace discrimination are of particular importance within companies developing these systems (Birhane, Kalluri, et al., 2022). Together, these factors suggest whistleblowers may be both more important for timely identification of AI-related risks, harms, and injustices, and also less able to point to specific illegal activity when making such disclosures.

Exploring the different constraints, concerns, and incentives of potential whistleblowers in this domain could help inform the creation of institutions to support them coming forward. Specific protections for whistleblowers in AI could include explicit expansion of protections for revealing a wider variety of problems that experts agree constitute significant harms—such as questions of algorithmic bias (Angwin et al., 2016), or deceptive, highly capable AI (Ngo et al., 2023)—even before ratification of laws against such harms. Support outside of new legislation may also help whistleblowers face potential consequences. For example, a legal defense fund specific to AI-related disclosures would increase incentives for disclosing potential AI risk even when legal questions of wrongdoing are not clearly defined. Operational support in the form of dedicated advisors on information security best practices, or connections to regulators and media contacts, could also help maintain individual anonymity, and could assist with priority treatment of high-stakes cases.

Understanding which of these forms of assistance would be most impactful for potential whistleblowers could help clarify the priorities of whistleblower protection organizations aiming to support AI whistleblowers specifically, making it more likely we learn of harmful choices by institutions developing or deploying AI in time for public and private strategies to mitigate them.

2.1.7 Design architectures that prioritize safety, transparency, and human review

The difficulty of understanding and predicting the internal dynamics of current-generation AI systems impedes our ability to ensure their safety, but developing novel architectures that prioritize interpretability by design may help future, highly capable systems remain safe, corrigible, and reliable. One possible such design would be a modular architecture that assumes every constituent AI component is untrustworthy, and enables human review of each component. One approach to facilitating this separation of functions would be to design a system with AI-based tools for each of the following steps:

1. Help users generate a specification that encapsulates desired properties of a solution,
2. Generate proposed solutions from the specification, and
3. Generate proof that the solutions generated satisfy a set of relevant criteria.

A beneficial property of these systems is that the architecture is compatible with existing human governance structures. Most alignment processes implicitly aggregate opinions of multiple people through data used for training or fine-tuning. This approach scales to large volumes of data, but does not allow for close oversight or control of what that implicit process encodes, which can result in unfair, biased, offensive, or otherwise harmful models (Barocas & Selbst, 2016; Paullada et al., 2021). By comparison, safe-by-design systems could synthesize stakeholder opinions explicitly in the specification of constraints, which could be reviewed and iterated on by humans before a solution is generated. Additionally, specifications can be combined, with a solution generated by such an architecture for a user needing to satisfy national, local, corporate, and individual constraints. This explicit encoding of specifications could increase transparency into which opinions and preferences are best satisfied by a particular system, offering paths towards understanding which cultures and values a system implicitly supports (Prabhakaran, Qadri, & Hutchinson, 2022). Further, while such systems are not sufficient to ensure democratic decision-making, they offer a platform for transparent, participatory approaches to governing AI systems, which can make such processes more inclusive, safer, and less risky for groups whose input is often overlooked (Birhane, Isaac, et al., 2022).

[3] See https://www.oig.dhs.gov/whistleblower-protection, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L1937, https://www.whistleblowers.org/whistleblower-laws-around-the-world/.

There are multiple explorations of this general direction. David "davidad" A. Dalrymple and Eric Drexler co-created one such design, posting separately about the Open Agency Architecture (Dalrymple, 2022) and Open Agency Model (Drexler, 2023). Davidad is now pursuing this direction with his research programme at ARIA (Dalrymple, 2023a, 2023b; Kolly & Segerie, 2023). Open Agency systems can be characterized by the elicitation and aggregation of safety criteria and other desiderata from stakeholders to form a specification, generation of proposals meeting that specification, and proofs or evidence that those proposals satisfy the specification. In a related vein, Tegmark and Omohundro use the term "provably safe systems"(Tegmark & Omohundro, 2023) to emphasize the importance of AI systems mathematically proving specified properties. Yoshua Bengio has also presented (Bengio, 2023) an agenda that fits within this general direction.

In addition to the above researchers advancing feasibility of universal Open Agency systems, multiple groups are translating this work from theory to application. A useful prototype system must at least include a specification language, a specification generator, a solution generator, and a world-model to validate the solution against the specification. Atlas Computing, the AI Objectives Institute, and Digital Gaia are all implementing useful, minimalist prototypes of the Open Agency Architecture in different subject domains, namely cybersecurity and biochemistry (Atlas), supply chain economics (AOI), and environmental policy (DG).

2.2 Understanding AI Systems

2.2.1 Improve interpretability of model behavior and outputs

Producing human-interpretable models of AI systems allows developers and users to better understand their capabilities, limitations, and failure modes so they can more safely and productively deploy or use them.

For developers, model interpretability research can help identify the underlying features and algorithms a model relies on to make its predictions. Circuits-style mechanistic interpretability methods (Elhage et al., 2021; Lepori et al., 2023; Olah et al., 2020) can help characterize high-level concepts or algorithms represented by small groups of neurons, while concept mapping techniques using probing classifiers (Tenney et al., 2018, *inter alia*) or model gradients (Kim et al., 2018; Sundararajan et al., 2017, *inter alia*) may characterize how such concepts appear as activation patterns in the network. Combining these with perturbation methods like model editing (Meng et al., 2022) , causal analysis (Geiger et al., 2023), or other interventions on a model's intermediate representations (Yu et al., 2023) can establish that a model depends on certain high-level features, concepts, or knowledge to make decisions. These features may include:

- Concepts which we don't want the model to use, e.g., reliance on protected characteristics which we may want to suppress or remove from the model's representation or make unavailable to the model's final decision procedure (Belrose et al., 2023; Ravfogel et al., 2020).
- Representations of knowledge that may drive the model's decisions in undesirable ways, for example by virtue of being false (Burns et al., 2023), violating individual privacy (Carlini et al., 2021), (Lukas et al., 2023), or enabling dangerous activity such as the production of weapons or biological agents (Sandbrink, 2023). In such cases, models may be edited to remove the influence of these features, if possible (Eldan & Russinovich, 2023), or not deployed if the risk of harm is too high.
- Other aspects of a model's internal representation which could lead to undesirable deployment behavior or generalization patterns, e.g., implicit representation of goals (Hubinger et al., 2019) or optimization algorithms (von Oswald et al., 2023) which could result in dangerous behavior for models of very high capability (Omohundro, 2008).

For users, methods for communicating model internal state are important for empowering users to use AI outputs to make decisions in line with their objectives. This can mean:

- Citing sources and providing extractive quotes (Menick et al., 2022) or directly executing code or using tools with well-defined semantics (Schick et al., 2023);

- Architecting the model to expose interpretable, learned intermediate variables corresponding to concepts which explain the model's decision, and which a user can intervene or provide feedback on to steer model behavior (Koh et al., 2020; Zarlenga et al., 2022).
- Explicitly estimating model uncertainty (Gal, 2016; Hendrycks & Gimpel, 2016; Kadavath et al., 2022) or refusing to answer questions where the model is likely to be incorrect; or,
- Explicitly tracking certain kinds of state, and employing new UX paradigms to communicate such state, for example, (Viégas & Wattenberg, 2023)'s *system model and user model*.

Making AI systems more transparent to users with explicit models of their behavior can help prevent mistakes due to over-reliance on model outputs (see Hill, 2020; *People v. Zachariah C. Crabill*, 2023). More generally, designing user interaction protocols by which AI improves human decision making can feed back into the supervision process, enabling scalable oversight (Amodei et al., 2016) of increasingly-capable models to reliably perform tasks that human overseers cannot reliably perform or verify on their own given realistic time and resource constraints (see Bowman et al., 2022; Michael et al., 2023).

An important property of interpretability methods is how accurately they describe the behavior of the underlying model, often referred to as *faithfulness* (Jacovi & Goldberg, 2020). Some approaches to interpretability, such as extractive quoting or tool use in a language model, are faithful by construction, whereas machine learning based methods like probing, model editing, or instance-based explanation methods—which provide a reason that the model made an individual prediction—must be empirically evaluated for faithfulness (Dasgupta et al., 2022; DeYoung et al., 2020; Turpin et al., 2023).

While in some cases faithfulness may be automatically evaluated, in general it depends on the accuracy of a *human* (e.g., user)'s model of the system when using an interpretability method. More broadly, the sociotechnical relevance of interpretability methods extends to the decisions that humans make on the basis of these methods. As such, empirically assessing progress on interpretability requires measuring the usefulness of interpretability methods for understanding and steering model behavior in human–AI interactive settings (Bansal et al., 2019; Chaleshtori et al., 2023; Doshi-Velez & Kim, 2017; M. Lee et al., 2023). Such evaluation methods may measure, for example:

- Human–AI collaborative performance on tasks (Chaleshtori et al., 2023). Good models of AI behavior will help users draw inferences on the basis of their outputs, and decide when and how to trust them. The accuracy and breadth of these inferences may be measured through the ability of a human–AI team to exceed the performance of either party alone (Bansal et al., 2021), improving the ability of humans to supervise models as measured in the *sandwiching* experimental paradigm (Bowman et al., 2022; Cotra, 2021). Measuring human–AI team performance in the presence of interpretability or explainability methods can also highlight problems with misleading, confusing, or overconfident explanations which harm overall decision making (Ghassemi et al., 2021).
- Comprehensive metrics of human–AI interaction quality, capturing aspects such as helpfulness, ease of use, or partial correctness (K. M. Collins, Jiang, et al., 2023; M. Lee et al., 2023). Such evaluations may serve to guide the development of interpretability methods or interaction paradigms which serve their intended use case in ways which are difficult to measure in static evaluations or automated measurements of accuracy.
- Simulatability (Chen et al., 2023; Doshi-Velez & Kim, 2017; Hase & Bansal, 2020), the ability of humans to predict how the model will behave (e.g., under counterfactual perturbations of an input) when leveraging some interpretability method, which directly tests the accuracy of the human's mental model of the system.
- Steerability of model behavior under interventions, whether conversational (i.e., direct correction or questioning, (K. M. Collins, Jiang, et al., 2023) or concept-driven, guided by interpretability methods (Koh et al., 2020; Zarlenga et al., 2022). Models should be responsive to such interventions, changing their predictions in the expected way (Lanham et al., 2023) as well as, ideally, robust to human mistakes and uncertainty about the associated concepts (K. M. Collins, Barker, et al., 2023).
- How the interpretability method can help humans anticipate or prevent undesirable model behaviors in interactive settings, including deception or manipulation of users (Park et al., 2023) and harmful actions taken as part of plans to meet user-specified goals (Scheurer et al., 2023).

Ideally, empirical evaluations should aim to assess interpretability methods in a wide range of domains, especially those which approximate a model's intended (or expected) usage scenario as closely as possible. Another important aspect to consider is that a human's mental model of a system derives from more than the interpretability method itself, but also the human's prior knowledge, culture, goals in using the system, and background assumptions about how the system works or how to interpret the output of an interpretability method. Successfully communicating a mental model of a system to a wide variety of users requires understanding the varieties of user and how they interact with and interpret the outputs of interpretability methods. As examples, this may involve studying methods by which humans can learn to effectively use interpretable AI systems (Mozannar et al., 2021, 2023), developing an understanding of the possible failure modes—e.g., inferring the existence of structure where there is none (Adebayo et al., 2018)—and developing and testing end-to-end frameworks for interfacing with interpretability methods (Slack et al., 2023). In this way, progress in interpretability depends on a combination of technical, social, and design questions.

2.2.2 Study AI accidents in their sociotechnical context to learn from their failures

Understanding how and why AI systems fail in real-world contexts can provide insight for how such systems can be made safer and more reliable. In parallel with efforts to avert unexpected harms, preparing for detailed and holistic analysis of these systems' negative impacts when they happen will help us gain the knowledge necessary to prevent similar failures in the future. Because failure cases can be difficult to investigate in retrospect without sufficient monitoring in place, it is essential to build the infrastructure necessary for these *post facto* analyses in advance, based on predictions of what failures may occur.

Previous work on risk analysis and safety engineering in other fields offers models for rigorous *in situ* investigation of system weaknesses precipitating adverse events in complex sociotechnical environments. Protocols for detailed forensic analysis of systemic harms arising post-deployment, such as the Systems Theoretic Accident Model and Processes (Leveson, 2004) could be expanded and refined for AI-specific use. Such protocols would facilitate an understanding of problems arising not only from faulty elements of a particular system—where such elements can be technical tools and systems, or human factors such as individual decision processes and institutional culture—but also from detrimental relationships between elements that do not have obvious flaws in isolation. Past issues of user manipulation in recommender systems (Kasirzadeh & Evans, 2023), for example, are clearest when viewed as a harmful feedback relationship between users and algorithms (Krueger et al., 2020)—and many other current and future societal harms from AI will likely follow this pattern (Siddarth et al., 2021). Detailing the interlinked social and technical drivers of actual AI failures can surface preventative approaches (Avin et al., 2021).

Incorporating such risk-analysis frameworks into existing documentation and monitoring efforts could significantly improve the understanding of potential AI risks we gain from such projects. Developing sociotechnical analysis templates for use with the AI Incident Database, for example, or expanding the NIST AI Risk Management Framework to include guidance on systems-theoretic risk assessment, would help guide responses to AI failures in the direction of detailed and holistic understanding. The results of past retrospective failure analyses in other domains, such as the 1986 Challenger explosion, and the safety measures taken in response, suggest there is high potential impact of the insights gained from such processes—and there is reason for optimism about our ability to avoid repeated failures of AI systems.

2.2.3 Improve public understanding of the capabilities and societal impacts of AI

As artificial intelligence becomes more complex and ubiquitous, improving public understanding of these systems is increasingly important, both for individuals' own informed use and for effective, legitimate democratic governance of these new technologies (Cave et al., 2019; Lazar, 2022).

Work in this direction would aim to provide the general public with the knowledge and tools necessary to navigate the rapidly evolving landscape of artificial intelligence. Central to this project is the use of interactive tools that transform the abstract complexities of AI into tangible, interactive learning experiences (Vogel et al., 2006). These tools would be not just visual aids, but interactive platforms where users can simulate scenarios, visualize the decision-making pathways of AI, and experience the results of these systems in real time (Mima, 2018). This hands-on approach demystifies AI

technologies, making them more accessible and understandable to individuals regardless of their technical background.

In addition to interactive learning, effective public education should emphasize the ethical and societal implications of AI systems (Mima, 2021). Curated educational content, workshops, and seminars will help participants explore topics such as algorithmic bias, privacy, and the ethical use of AI (Barocas & Selbst, 2016; Zhang et al., 2023). This curriculum aims to stimulate critical thinking and ethical reasoning, and to equip the public with the cognitive tools to question and understand the moral dimensions of AI applications. It's not enough to know how AI works; understanding when and why its use may be problematic is crucial to promoting a socially responsible approach to AI (Hoffman et al., 2023).

Finally, educational resources must evolve as AI advances. To ensure sustainability, such projects should incorporate dynamic feedback systems that invite input from a wide range of users to keep the content relevant and up-to-date. Collaborations with academia, industry leaders, and policymakers will help integrate AI literacy campaigns into broader educational and regulatory frameworks. By establishing an ongoing dialogue between AI developers and the public, these projects can contribute to a culture of informed consent and ethical responsibility in the age of AI, helping the public remain influential in the trajectory of AI development, deployment, and use.

2.2.4 Develop evaluation frameworks for key aspects of AI systems

As frontier models become more capable, they should be evaluated for a variety of safety criteria and dangerous capabilities. Safety criteria would measure to what extent the model poses sociotechnical and alignment risks, including bias, misinformation, deception, and power-seeking (Perez et al., 2023; Weidinger et al., 2023). Safety evaluations would include robustness assessments (to what extent does the system consistently behave acceptably) and impact assessments (of anticipated or observed effects of specific deployments and use cases on impacted groups) (AI Policy and Governance Working Group, 2023). Evaluations of dangerous capabilities (including persuasion / manipulation, self-proliferation and cyber-offense) would measure misuse and accident risks posed if the model is given undesirable objectives by bad actors or develops them on its own (Shevlane et al., 2023).

Some of these types of evaluations are of interest both from sociotechnical and alignment perspectives. For example, a model's persuasion abilities would contribute to misinformation and misuse risks as well as enable the model to escape human control. Broad collaboration between different research communities on evaluations of common interest would produce more holistic and robust measures of a model's risks and impacts.

For all of these types of evaluations, it would be beneficial to create effective infrastructure for frequently evaluating frontier models, institutional policies for incorporating these evaluations in deployment decisions (Anthropic, 2023a; Brundage et al., 2020; OpenAI, 2023), external auditing for evaluation outcomes, and sharing the best evaluation designs among frontier AI labs.

2.2.5 Evaluate the reward models used in RLHF

Reinforcement learning from human feedback (RLHF) has gained popularity due to its success in fine-tuning large language models like ChatGPT and Claude (Bai et al., 2022; Christiano et al., 2017; OpenAI, 2022). However, the lack of transparency and rigorous evaluation of the reward models learned by RLHF raises concerns about the long-term impacts of this technique, and the limitations of working with aggregate human data in RLHF highlight the need to consider whose values are being encoded and prioritized in these models. Addressing these concerns requires better understanding of the reward models used in RLHF and the values they encode (to the extent that they encode values at all). It also requires a better understanding of how pretraining and fine-tuning interact to influence LLMs' behavior. This paradigm also extends to evaluating models trained with new methods such as Direct Preference Optimization (DPO), which induce a policy directly from a policy – it should be confirmed that the underlying optimization of preferences does what we expect via the theory and data. Recent work on reward model interpretability (Marks et al., 2024) has laid groundwork for more precise inspection of implicit reward models, but further research on reward model evaluation is needed in the following key research directions:

1. Evaluation of reward model capabilities: While the most common evaluation technique for learned reward models is to measure a model's agreement with data set aside from the

training set, this method does not control for the complex nature of such datasets, including potential conflicts in aggregate inputs from multiple individuals or contexts. Evaluation methods that test reward models in a manner analogous to their use would provide more insight into how RHLF-trained models will perform—for example, testing whether a reward model designed to prioritize factuality and honesty meets benchmarks for exhibiting those qualities. One specific example exists in recent work, a benchmark to evaluate reward model consistency as its ability to maintain consistent scores over changes to text that do not alter meaning (Shen et al., 2023), but more tools are needed to cover the many use cases of RLHF models.

2. Evaluation of reward model safety: Red-teaming is the primary means of evaluating the safety, toxicity, and potential harms of LLMs (Ganguli et al., 2022). While red-teaming the text output of LLMs evaluates on multiple dimensions, red-teaming reward models is simpler and less interactive, as the output of the reward model is a score, not a text to read and comprehend. Scoring reward models by comparing LLM text outputs to established "neutral" texts would provide a richer basis for comparison. Reward models should also be evaluated for sensitivity to adversarial input strings, which could indicate downstream exploitation during the RL step. Today, a third-party researcher can play with a closed model and find an adversarial attack even if no red-teaming protocols are shared about the system, but the same is not true for a private reward model, which may be re-used in the future to train other models. As a starting point, practitioners could consider the efforts to identify adversarial prompts that may escalate toxicity in the evaluation of LLM generations, for instance as is done in the RealToxicityPrompts dataset (Gehman et al., 2020).
3. Sociotechnical specification of preference: Given the complex intellectual origins of quantifying and optimizing preferences, specific recommendations should be made around what information a model of human preferences should and should not encode. For example, accounting for individual preferences about other individuals' actions can produce preference aggregations that hamper those latter individuals' agency. Doing so may reduce the performance of models in the short term, but increases the potential for multi-stakeholder engagement and reduction of harms with the development of LLMs.

2.3 Improving AI's Social Impacts

2.3.1 Track and measure the impact of AI in social media and LLMs

Recent senate testimony has highlighted how the experiences users report to be harmful often do not conform to the metrics that companies report (Hendrix, 2023), which focus on content that violates company policies. Mirroring evidence internal to platforms (*Harmful Non-Violating Narratives*, 2021), incidents of hate speech are often less common than violence-inciting fear speech, though the latter often do not violate policy (Saha et al., 2023). If we want to understand and hold companies accountable for their current use of AI, which underlies social media recommender systems (Clegg, 2023), we need to understand how user experience changes, and whether advances in AI are increasing or decreasing users' experiences of encountering harmful recommended content (e.g. Almost half of US teens report being harassed online (Vogels, 2022), and algorithmic objective functions have been shown to increase those experiences (Horwitz et al., 2023). In a US study by Common Sense Media (Robb & Mann, 2023), the average age of pornography exposure online is age 12, and 58% of youth say they came across pornography online by accident).

If increasingly powerful AI were to make social media systems better, by enabling more value-aligned recommendations, or worse, with AI generated misinformation and non-consensual sexual imagery being rampant, how would we know? To answer this question, USC's Neely Center has been replicating internal platform research (Pahwa, 2021) on user experience, by surveying a representative sample of users as to their positive and negative experiences across platforms (Fast et al., 2023). The results have already been used by the press (Counts & Nakano, 2023) to hold technology companies accountable for negative experiences and been consumed internally by companies seeking to improve the user experience of their products.

We'd like to extend this approach to a larger global sample, allowing us to understand how harmful experiences may be affecting sub-samples and people not just in the United States, but in places around the world where we have signal that the impact of AI driven social media systems can be grave (Amnesty International, 2022a), but where we often lack measurement. Further, we are also

extending this paradigm to understanding people's direct experiences with generative AI systems. In our first analyses during spring of 2023 (Fast et al., 2023), roughly 10% of Americans reported using Large Language Models. Our latest polling from Fall 2023 (Motyl et al., 2024) indicates that these numbers have grown, with 18% having reported using LLMs. Most reported using LLMs out of curiosity, and few reported using them for higher-risk purposes, such as companionship and mental health. Most people found these tools useful and few reported harm. But as such use cases grow, we hope to understand both positive and negative experiences across use cases for these new technologies, to mitigate risks before they become widespread—arguably something we did not do before AI powered social media algorithms became widely used.

2.3.2 Develop LLM tools to mitigate polarization

Several tools exist for analyzing opinion clusters with AI – including Polis, Remesh, Cortico and Talk to the City—but their primary focus has been on presenting the stances of individuals involved in deliberation, rather than mediating the process of deliberation itself in a way that would be beneficial to participants. As shown in a recent comprehensive multi-disciplinary review (Caluwaerts et al., 2023), deliberation processes can have a positive influence both on idea-based polarization (extreme ideologies) and on affective polarization (strong emotions). However, most of the studies captured in this review suggest that a good facilitation process is essential for deliberation processes to have a positive impact on depolarisation. Some prototypes have been built to explore how LLMs could partially automate the work of facilitators (see for instance Depolarized), but we believe more work needs to be done in this direction. Further work could also explore tools for directing individuals towards more grounded discussions (Konya et al., 2023), guided by existing research on bridging (Ovadya & Thorburn, 2023) and depolarization (Stray, 2022) and the work of practitioners such as Braver Angels. The bridging algorithms powering Community Notes on X could also be integrated to deliberation platforms to try and take polarization explicitly into account to mitigate its effects.

2.3.3 Ensure influence from AI systems is beneficial, or at least non-manipulative

AI systems—including recommender systems, chat interfaces, and other applications of generative AI—frequently play a role in determining people's information environment, and so constitute a vector by which people might be manipulated or otherwise deleteriously influenced (Ammann, 2023).

Preventing wrongful manipulation (and other forms of undesirable influence) is important for respecting human autonomy, but it is challenging to (1) characterize what manipulation is both philosophically (Noggle, 2022) and technically (Carroll et al., 2023), (2) articulate when it is wrongful (Benn & Lazar, 2022; Bezou-Vrakatseli et al., 2023), and (3) formalize non-manipulative objectives or baselines against which AI systems can be evaluated. There is some early work on estimating and penalizing preference changes (Carroll et al., 2022), and highlighting that solutions to manipulation require knowledge of meta-preferences (Ashton & Franklin, 2022), but much more work is needed.

In the case of recommenders, the possibility of feedback loops—between what is shown, what is thought, and what is engaged with (Thorburn, 2023)—has led to concerns that a recommender system may learn to increase its reward not by making better recommendations, but by manipulating users so that its objective is more easily satisfied (Thorburn, 2022). This phenomena has been variably called 'preference manipulation', 'user tampering' (Kasirzadeh & Evans, 2023), and 'auto-induced distributional shift' (Krueger et al., 2020).

Progress towards non-manipulative recommender systems would also inform the design of generative AI applications, which work has already shown can be manipulative. For example, the sentiment of suggested responses in chat interfaces influences the trajectory of human conversations (Hohenstein et al., 2023).

3 Limitations & Future Work

This collection of proposals is intended as an exploration of potential sociotechnical approaches to the development of safe and beneficial AI. Far from an exhaustive list, it represents the interests and backgrounds of its authors and other contributors to our exploratory discussions. Our backgrounds and the context of these conversations represent perspectives on AI that center safety concerns, risk

mitigation, and harms from existing ML systems, but are not representative of the full breadth of social concerns relevant to guiding the development of AI.

Further work in this direction would benefit from expanding to incorporate more knowledge and methods from other approaches to these questions, beyond the overlap we have found with our areas of concern. Specific areas overlooked in this agenda include research exploring the beliefs, values, and ontologies that drive particular approaches to AI development and deployment, and research exploring the complexities of individual preferences. Future explorations would also benefit from the knowledge of domain experts with backgrounds in social justice, legal theory, and international governance, for a deeper exploration of the existing dynamics of our social structures that AI may reinforce or disrupt.

4 Conclusion

This paper proposes a set of research directions that describe how certain areas of research core to AI safety can incorporate sociotechnical approaches to mitigating the risks and improving the outcomes of AI systems. The collection builds on conversations discussing areas of common concern among multiple approaches to beneficial AI deployment, and presents research areas we believe are promising directions for continued exploration. Proposals are varied in their targets and emphases, but they share grounding in a sociotechnical approach to governing, understanding, and guiding the use of these new technologies.

While research has already begun in many of the directions described above, significant work remains to explore them in full. In some cases, this work will depend on cooperation across sectors to explore policy solutions and other government interventions. We believe that the sociotechnical approach to AI safety offers a strong foundation for pursuing this work from a holistic, systemic understanding of how technology and society interact. This agenda, while not exhaustive, suggests potential concrete avenues for incorporating such approaches, and for exploring how more collaboration could lead to progress on the shared challenge of guiding AI progress towards broad societal benefit.

Acknowledgements. We thank David Krueger, Seth Lazar, Fazl Barez, Carroll Wainwright, Micah Carroll, Anna Leshinskaya, Usman Anwar, and Katherine Collins for their feedback on earlier versions of this paper. Any errors or limitations of this work remain those of the authors. We would additionally like to thank participants in discussions that inspired this paper, held at at Berkeley Center for Human-Compatible AI (CHAI), the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), and the Stanford Institute for Human-Centered AI (HAI).

References

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. *Journal of Political Economy*, *128*(6), 2188–2244. https://doi.org/10.1086/705716

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 9525–9536.

AI Now Institute, Amba Kak, & Sarah Myers West. (2023). *General purpose AI poses serious risks, should not be excluded from the EU's AI act – policy brief* [AI Now Institute]. https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act

AI Policy and Governance Working Group. (2023). Comment of the AI policy and governance working group on the NTIA AI accountability policy request for comment docket NTIA-230407-0093. https://www.ias.edu/sites/default/files/AI%20Policy%20and%20Governance%20Working%20Group%20NTIA%20Comment.pdf

Ammann, N. (2023). *The value change problem*. https://www.lesswrong.com/s/3QXNgNKXoLrdXJwWE

Amnesty International. (2022a). *Myanmar: Facebook's systems promoted violence against rohingya; meta owes reparations – new report*. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

Amnesty International. (2022b). *Myanmar: The social atrocity: Meta and the right to remedy for the rohingya*. https://www.amnesty.org/en/documents/asa16/5933/2022/en/

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

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). Frontier AI regulation: Managing emerging risks to public safety. http://arxiv.org/abs/2307.03718

Angwin, J., Mattu, J., Larson, L., & Kirchner, S. (2016). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Anthropic. (2023a). *Anthropic's responsible scaling policy*. https://www.anthropic.com/news/anthropics-responsible-scaling-policy

Anthropic. (2023b). *Collective constitutional AI: Aligning a language model with public input*. https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input

Arnold, M. (2023). *Statement to the U.S. Senate: AI insight forum - privacy and liability* [Legal priorities project]. https://www.legalpriorities.org/research/ai-insight-forum.html

Ashton, H., & Franklin, M. (2022). Solutions to preference manipulation in recommender systems require knowledge of meta-preferences. http://arxiv.org/abs/2209.11801

Atleson, M. (2023). *Keep your AI claims in check*. https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check

Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., Anderljung, M., Krawczuk, I., Krueger, D., Lebensold, J., Maharaj, T., & Zilberman, N. (2021). Filling gaps in trustworthy development of AI. *Science*, *374*(6573), 1327–1329. https://doi.org/10.1126/science.abi7176

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. http://arxiv.org/abs/2204.05862

Bakker, M. A., Chadwick, M. J., Sheahan, H. R., Tessler, M. H., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M. M., & Summerfield, C. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. http://arxiv.org/abs/2211.15006

Bales, A., D'Alessandro, W., & Kirk-Giannini, C. D. (2024). Artificial intelligence: Arguments for catastrophic risk. *Philosophy Compass*, *19*(2). https://doi.org/https://doi.org/10.1111/phc3.12964

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445717

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732. https://www.jstor.org/stable/24758720

Barrett, J., Gal, K., Gölz, P., Hong, R. M., & Procaccia, A. D. (2023). Now we're talking: Better deliberation groups through submodular optimization. *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, *37*, 5490–5498. https://doi.org/10.1609/aaai.v37i5.25682

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI & SOCIETY*, *35*(1), 165–176. https://doi.org/10.1007/s00146-017-0760-1

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2023). LEACE: Perfect linear concept erasure in closed form. https://openreview.net/forum?id=awIpKpwTwF

Bengio, Y. (2023). *Towards AI safety that improves with more compute*. https://www.youtube.com/watch?v=SlNvufmdBHU

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S.,

Gao, Q., Acharya, A., Krueger, D., Dragan, A., … Mindermann, S. (2023). Managing AI risks in an era of rapid progress. http://arxiv.org/abs/2310.17688

Benn, C., & Lazar, S. (2022). What's wrong with automated influence. *Canadian Journal of Philosophy*, *52*(1), 125–148. https://doi.org/10.1017/can.2021.23

Bertolini, A. (2020). *Artificial intelligence and civil liability* (PE 621.926). Report Commissioned by Policy Department for Citizens' Rights and Constitutional Affairs. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/621926/IPOL_STU(2020)621926_EN.pdf

Bezou-Vrakatseli, E., Brückner, B., & Thorburn, L. (2023). SHAPE: A framework for evaluating the ethicality of influence. http://arxiv.org/abs/2309.04352

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? opportunities and challenges for participatory AI. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8. https://doi.org/10.1145/3551624.3555290

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184. https://doi.org/10.1145/3531146.3533083

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356–4364.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., … Kaplan, J. (2022). Measuring progress on scalable oversight for large language models. http://arxiv.org/abs/2211.03540

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (Eds.). (2016). *Handbook of computational social choice*. Cambridge University Press. https://doi.org/10.1017/CBO9781107446984

Bristow, T., & Thorburn, L. (2023). Don't slip into binary thinking about AI. http://arxiv.org/abs/2312.14230

Brown, I. (2023). *Expert explainer: Allocating accountability in AI supply chains*. https://www.adalovelaceinstitute.org/resource/ai-supply-chains/

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

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2023). Discovering latent knowledge in language models without supervision. https://openreview.net/forum?id=ETKGuby0hcs

Cabrera, Á. A., Perer, A., & Hong, J. I. (2023). Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction*, *7*, 136:1–136:21. https://doi.org/10.1145/3579612

Caluwaerts, D., Bernaerts, K., Kesberg, R., Smets, L., & Spruyt, B. (2023). Deliberation and polarization: A multi-disciplinary review. *Frontiers in Political Science*, *5*. https://www.frontiersin.org/articles/10.3389/fpos.2023.1127372

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. *Proceedings of the 30th USENIX Security Symposium*. https://www.usenix.org/system/files/sec21-carlini-extracting.pdf

Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). Characterizing manipulation from AI systems. http://arxiv.org/abs/2303.09387

Carroll, M., Dragan, A., Russell, S., & Hadfield-Menell, D. (2022). Estimating and penalizing induced preference shifts in recommender systems. http://arxiv.org/abs/2204.11966

Carson, L. (2017). *Deliberation* [newDemocracy Foundation]. https://www.newdemocracy.com.au/2017/03/22/deliberation/

Castillo, C., Chouldechova, A., De-Arteaga, M., Ekstrand, M., Lazar, S., & Members of the FAccT Community. (2023). *ACM FAccT – Statement on AI harms and policy*. https://facctconference.org/2023/harm-policy

Cave, S., Coughlan, K., & Dihal, K. (2019). "Scary robots": Examining public responses to AI. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 331–337. https://doi.org/10.1145/3306618.3314232

Center for AI and Digital Policy. (2023). *In the matter of Open AI (Federal Trade Commission 2023)*. https://www.caidp.org/cases/openai/

Center for AI Safety. (2023). *Statement on AI risk*. https://www.safe.ai/statement-on-ai-risk

Chaleshtori, F. H., Ghosal, A., & Marasovic, A. (2023). On evaluating explanation utility for human-AI decision-making in NLP. https://openreview.net/forum?id=8BR8EaWNTZ

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., … Maharaj, T. (2023). Harms from increasingly agentic algorithmic systems. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666. https://doi.org/10.1145/3593013.3594033

Chen, Y., Zhong, R., Ri, N., Zhao, C., He, H., Steinhardt, J., Yu, Z., & McKeown, K. (2023). Do models explain themselves? Counterfactual simulatability of natural language explanations. http://arxiv.org/abs/2307.08678

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

Clegg, N. (2023). *How AI influences what you see on facebook and instagram* [Meta]. https://about.fb.com/news/2023/06/how-ai-ranks-content-on-facebook-and-instagram/

Collective Intelligence Project. (2023). *CIP and Anthropic launch Collective Constitutional AI*. https://cip.org/blog/ccai

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2023). Evaluating language models for mathematics through interactions. http://arxiv.org/abs/2306.01694

Collins, K. M., Barker, M., Espinosa Zarlenga, M., Raman, N., Bhatt, U., Jamnik, M., Sucholutsky, I., Weller, A., & Dvijotham, K. (2023). Human uncertainty in concept-based AI systems. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 869–889. https://doi.org/10.1145/3600211.3604692

Collins, P. H. (1990). *Black feminist thought, 30th anniversary edition: Knowledge, consciousness, and the politics of empowerment*. Routledge.

The constitution of the United States: A transcription [National archives]. (2015). https://www.archives.gov/founding-docs/constitution-transcript

Cooper, A. F., & Vidan, G. (2022). Making the unaccountable internet: The changing meaning of accounting in the early ARPANET. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 726–742. https://doi.org/10.1145/3531146.3533137

Cotra, A. (2021). The case for aligning narrowly superhuman models. https://www.alignmentforum.org/posts/PZtsoaoSLpKjjbMqM/the-case-for-aligning-narrowly-superhuman-models

Counts, A., & Nakano, E. (2023). Twitter's surge in harmful content a barrier to advertiser return. *Bloomberg.com*. https://www.bloomberg.com/news/articles/2023-07-19/twitter-s-surge-in-harmful-content-a-barrier-to-advertiser-return

Dafoe, A. (2017). *AI governance: A research agenda*. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

Dalrymple, D. A. (2022). An open agency architecture for safe transformative AI. https://www.alignmentforum.org/posts/pKSmEkSQJsCSTK6nH/an-open-agency-architecture-for-safe-transformative-ai

Dalrymple, D. A. (2023a). A list of core AI safety problems and how I hope to solve them. https://www.alignmentforum.org/posts/mnoc3cKY3gXMrTybs/a-list-of-core-ai-safety-problems-and-how-i-hope-to-solve

Dalrymple, D. A. (2023b). Mathematics and modelling are the keys we need to safely unlock transformative AI. https://www.aria.org.uk/wp-content/uploads/2023/10/ARIA-Mathematics-and-modelling-are-the-keys-we-need-to-safely-unlock-transformative-AI-v01.pdf

Dasgupta, S., Frost, N., & Moshkovitz, M. (2022). Framework for evaluating faithfulness of local explanations. *Proceedings of the 39th International Conference on Machine Learning*. https://proceedings.mlr.press/v162/dasgupta22a/dasgupta22a.pdf

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020). ERASER: A benchmark to evaluate rationalized NLP models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4443–4458). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.408

Directive (EU) 2019/1937 of the European Parliament and of the Council of 23 October 2019 on the protection of persons who report breaches of Union law (2019). http://data.europa.eu/eli/dir/2019/1937/oj/eng

Dobbe, R. I., Gilbert, T. K., & Mintz, Y. (2020). Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 242. https://doi.org/10.1145/3375627.3375861

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. http://arxiv.org/abs/1702.08608

Dourish, P. (2001). *Where the action is: The foundations of embodied interaction*. The MIT Press. https://doi.org/10.7551/mitpress/7221.001.0001

Drexler, E. (2023). *The open agency model* [AI alignment forum]. https://www.alignmentforum.org/posts/5hApNw5f7uG8RXxGS/the-open-agency-model

Dror, L. (2023). Is there an epistemic advantage to being oppressed? *Noûs*, *57*(3), 618–640. https://doi.org/10.1111/nous.12424

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Edelman, J., & Klingefjord, O. (2023a). *Introducing democratic fine-tuning* [Meaning Alignment Institute]. https://meaningalignment.substack.com/p/introducing-democratic-fine-tuning

Edelman, J., & Klingefjord, O. (2023b). *OpenAI x DFT: The first moral graph* [Meaning Alignment Institute]. https://meaningalignment.substack.com/p/the-first-moral-graph

Eldan, R., & Russinovich, M. (2023). Who's Harry Potter? Approximate unlearning in LLMs. http://arxiv.org/abs/2310.02238

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html

Etienne, H. (2021). The dark side of the 'moral machine' and the fallacy of computational ethical decision-making for autonomous vehicles. *Law, Innovation and Technology*, *13*(1), 85–107. https://doi.org/10.1080/17579961.2021.1898310

Fan, J., & Zhang, A. X. (2020). Digital juries: A civics-oriented approach to platform governance. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376293

Farber, D. A. (2022). *Catastrophic uncertainty and regulatory impact analysis*. http://dx.doi.org/10.2139/ssrn.4170257

Fast, N., Schroeder, J., Iyer, R., & Motyl, M. (2023). *Unveiling the neely ethics & technology indices* [Designing tomorrow]. https://psychoftech.substack.com/p/unveiling-the-neely-ethics-and-technology

Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., & Wüthrich, M. (2023). Generative social choice. http://arxiv.org/abs/2309.01291

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, *64*(4), 136–143. https://doi.org/10.1145/3433949

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

Future Society. (2021). Trust in excellence & excellence in trust. https://thefuturesociety.org/wp-content/uploads/2021/12/Summary-Trust-in-Excellence-Excellence-in-Trust-The-Future-Society.pdf

Gal, Y. (2016). *Uncertainty in deep learning* [Unpublished doctoral dissertation]. University of Cambridge. https://www.cs.ox.ac.uk/people/yarin.gal/website/thesis/thesis.pdf

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., … Clark, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. http://arxiv.org/abs/2209.07858

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. http://arxiv.org/abs/2009.11462

Geiger, A., Potts, C., & Icard, T. (2023). Causal abstraction for faithful model interpretation. http://arxiv.org/abs/2301.04709

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T., & Bernstein, M. S. (2022). Jury learning: Integrating dissenting voices into machine learning models. *CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3502004

Grandi, U. (2018). Agent-mediated social choice. http://arxiv.org/abs/1806.07199

Hadfield, G. K., & Clark, J. (2023). Regulatory markets: The future of AI governance. https://arxiv.org/abs/2304.04914

Hadfield-Menell, D., & Hadfield, G. K. (2019). Incomplete contracting and AI alignment. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 417–422. https://doi.org/10.1145/3306618.3314250

Hardt, M. (2014). *How big data is unfair* [Medium]. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

"Harmful Non-Violating Narratives" is a problem archetype in need of novel solutions [Gizmodo Facebook Papers Directory]. (2021). https://www.documentcloud.org/documents/21950059-tier1_c19_pr_0321

Hase, P., & Bansal, M. (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5540–5552). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.491

Henderson, J. G. (2023). *FTC authorizes compulsory process for AI-related products and services* [Federal Trade Commission]. https://www.ftc.gov/news-events/news/press-releases/2023/11/ftc-authorizes-compulsory-process-ai-related-products-services

Hendrix, J. (2023). *Transcript: Senate hearing on social media and teen mental health with former facebook engineer arturo bejar – TechPolicy.press*. https://techpolicy.press/transcript-senate-hearing-on-social-media-and-teen-mental-health-with-former-facebook-engineer-arturo-bejar

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. https://openreview.net/forum?id=Hkg4TI9xl&noteId=ryGs_6r4g

Hendrycks, D., & Mazeika, M. (2022). X-risk analysis for AI research. http://arxiv.org/abs/2206.05862

Hill, K. (2020). Wrongfully accused by an algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Hoffman, R. R., Mueller, S. T., Klein, G., Jalaeian, M., & Tate, C. (2023). Explainable AI: Roles and stakeholders, desirements and challenges. *Frontiers in Computer Science*, *5*. https://www.frontiersin.org/articles/10.3389/fcomp.2023.1117848

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, *13*(1), 5487. https://doi.org/10.1038/s41598-023-30938-9

Horwitz, J., Hagey, K., & Glazer, E. (2023). Facebook wanted out of politics. it was messier than anyone expected. *Wall Street Journal*. https://www.wsj.com/articles/facebook-politics-controls-zuckerberg-meta-11672929976

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. http://arxiv.org/abs/1906.01820

Irani, L. C., & Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in amazon mechanical turk. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 611–620. https://doi.org/10.1145/2470654.2470742

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4198–4205). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.386

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., … Kaplan, J. (2022). Language models (mostly) know what they know. http://arxiv.org/abs/2207.05221

Kahng, A., Lee, M. K., Noothigattu, R., Procaccia, A., & Psomas, C.-A. (2019). Statistical foundations of virtual democracy. *Proceedings of the 36th International Conference on Machine Learning*, 3173–3182. https://proceedings.mlr.press/v97/kahng19a.html

Kasirzadeh, A., & Evans, C. (2023). User tampering in reinforcement learning recommender systems. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 58–69. https://doi.org/10.1145/3600211.3604669

Katzenbach, C., & Bareis, J. (2022). Talking AI into being: The narratives and imaginaries of national ai strategies and their performative politics. *Science, Technology, and Human Values*, *47*(5), 855–881. https://doi.org/10.1177/01622439211030007

Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW). https://doi.org/10.1145/3274357

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning*, 2668–2677. https://proceedings.mlr.press/v80/kim18d.html

Kingston, J. K. C. (2018). Artificial intelligence and legal liability. https://arxiv.org/abs/1802.07782

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. *Proceedings of the 37th International Conference on Machine Learning*, 5338–5348. https://proceedings.mlr.press/v119/koh20a.html

Kolly, G., & Segerie, C.-R. (2023). Davidad's bold plan for alignment: An in-depth explanation. https://www.alignmentforum.org/posts/jRf4WENQnhssCb6mJ/davidad-s-bold-plan-for-alignment-an-in-depth-explanation

Konya, A., Schirch, L., Irwin, C., & Ovadya, A. (2023). Democratic policy development using collective dialogues and AI. http://arxiv.org/abs/2311.02242

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117. https://doi.org/10.1017/XPS.2020.37

Krueger, D., Maharaj, T., & Leike, J. (2020). Hidden incentives for auto-induced distributional shift. http://arxiv.org/abs/2009.09153

Landemore, H., & Page, S. E. (2015). Deliberation and disagreement: Problem solving, prediction, and positive dissensus. *Politics, Philosophy & Economics*, *14*(3), 229–254. https://doi.org/10.1177/1470594X14544284

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., … Perez, E. (2023). Measuring faithfulness in chain-of-thought reasoning. http://arxiv.org/abs/2307.13702

Law, J., & Hassard, J. (1999). *Actor network theory and after*. Blackwell, Boston, MA.

Lazar, S. (2022). Legitimacy, authority, and democratic duties of explanation. *Oxford Studies in Political Philosophy*.

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

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., & Procaccia, A. D. (2019). WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 181:1–181:35. https://doi.org/10.1145/3359283

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., Gerard-Ursin, I., Li, X. L., Ladhak, F., Rong, F., Wang, R. E., Kwon, M., Park, J. S., Cao, H., Lee, T., Bommasani, R., Bernstein, M. S., & Liang, P. (2023). Evaluating human-language model interaction. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=hjDYJUn9l1

Leong, B., & Kumarasamy, J. (2023). *Third-party liability and product liability for AI systems*. https://iapp.org/news/a/third-party-liability-and-product-liability-for-ai-systems/

Lepori, M. A., Serre, T., & Pavlick, E. (2023). Uncovering intermediate variables in transformers using circuit probing. http://arxiv.org/abs/2311.04354

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety Science*, *42*(4), 237–270. https://doi.org/10.1016/S0925-7535(03)00047-X

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing leakage of personally identifiable information in language models, 346–363. https://doi.org/10.1109/SP46215.2023.10179300

Madiega, T. (2023). *Artificial intelligence liability directive* (PE 739.342). European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf

Markoff, J. (2009). Scientists worry machines may outsmart man. *The New York Times*. https://www.nytimes.com/2009/07/26/science/26robot.html

Marks, L., Abdullah, A., Neo, C., Arike, R., Torr, P., & Barez, F. (2024). Beyond training objectives: Interpreting reward model divergence in large language models. http://arxiv.org/abs/2310.08164

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, *35*, 17359–17372. https://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html

Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., & McAleese, N. (2022). Teaching language models to support answers with verified quotes. http://arxiv.org/abs/2203.11147

Miceli, M., Yang, T., Alvarado Garcia, A., Posada, J., Wang, S. M., Pohl, M., & Hanna, A. (2022). Documenting data production processes: A participatory approach for data work. *Proceedings of the ACM on Human-Computer Interaction*, *6*, 510:1–510:34. https://doi.org/10.1145/3555623

Michael, J., Mahdi, S., Rein, D., Petty, J., Dirani, J., Padmakumar, V., & Bowman, S. R. (2023). Debate helps supervise unreliable experts. http://arxiv.org/abs/2311.08702

Mills, C. W. (2007). White ignorance. In S. Sullivan & N. Tuana (Eds.), *Race and epistemologies of ignorance* (pp. 11–38). State Univ of New York Pr.

Mima, N. (2018). The potential of local science festivals for a sustainable society. *2018 Science and You, International Conference on Science Communication*.

Mima, N. (2021). *Living in the age of AI: Creativity and empathy for designing the future*. Iwanami Shoten. https://www.iwanami.co.jp/book/b591618.html

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Moes, N. (2021). *Product liability directive - adapting liability rules to the digital age, circular economy and global value chains*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Product-Liability-Directive-Adapting-liability-rules-to-the-digital-age-circular-economy-and-global-value-chains/F2663292_en

Moës, N. (2021). *European commission - have your say* [The Future Society]. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12979-Product-Liability-Directive-Adapting-liability-rules-to-the-digital-age-circular-economy-and-global-value-chains/F2663292_en

Moës, N., & Ryan, F. (2023). *Heavy is the head that wears the crown: A risk-based tiered approach to governing general-purpose AI*. https://thefuturesociety.org/heavy-is-the-head-that-wears-the-crown/

Motyl, M., Narang, J., & Fast, N. (2024). *Tracking chat-based ai tool adoption, uses, and experiences*. https://psychoftech.substack.com/p/tracking-chat-based-ai-tool-adoption
Mozannar, H., Lee, J. J., Wei, D., Sattigeri, P., Das, S., & Sontag, D. (2023). Effective human-AI teams via learned natural language rules and onboarding. https://openreview.net/forum?id=V2yFumwo5B&noteId=VrOvAcMx4o
Mozannar, H., Satyanarayan, A., & Sontag, D. (2021). Teaching humans when to defer to a classifier via exemplars. http://arxiv.org/abs/2111.11297
Ngo, R., Chan, L., & Mindermann, S. (2023). The alignment problem from a deep learning perspective. http://arxiv.org/abs/2209.00626
Nissenbaum, H. (2001). How computer systems embody values. *Computer*, *34*(3), 118–120. https://doi.org/10.1109/2.910905
NIST. (2022). AI risk management framework: Second draft. https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf
Noggle, R. (2022). The ethics of manipulation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/
Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. (2018). A voting-based system for ethical decision making. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11512
Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*, *5*(3), e00024.001. https://doi.org/10.23915/distill.00024.001
Omohundro, S. M. (2008). The basic AI drives. *Proceedings of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 483–492.
OpenAI. (2022). *Introducing ChatGPT: Optimizing language models for dialogue.* https://openai.com/blog/chatgpt
OpenAI. (2023). *Preparedness*. https://openai.com/safety/preparedness
Ovadya, A. (2021). *Towards platform democracy: Policymaking beyond corporate CEOs and partisan pressure*. https://www.belfercenter.org/publication/towards-platform-democracy-policymaking-beyond-corporate-ceos-and-partisan-pressure
Ovadya, A. (2023a). 'Generative CI' through collective response systems. http://arxiv.org/abs/2302.00672
Ovadya, A. (2023b). *'Platform Democracy'—a very different way to govern powerful tech*. https://reimagine.aviv.me/p/platform-democracy-a-different-way-to-govern
Ovadya, A. (2023c). *Deliberative polls, citizen assemblies, and an online deliberation platform*. https://reimagine.aviv.me/p/deliberative-poll-vs-citizen-assembly-meta-pilot
Ovadya, A. (2023d). Reimagining democracy for AI. *Journal of Democracy*, *34*(4), 162–170. https://www.journalofdemocracy.org/articles/reimagining-democracy-for-ai/
Ovadya, A. (2024). *Build wise systems: Combining competence, alignment, and robustness* [Medium]. https://aviv.medium.com/building-wise-systems-combining-competence-alignment-and-robustness-a9ed872468d3
Ovadya, A., & Thorburn, L. (2023). Bridging systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance. http://arxiv.org/abs/2301.09976
Pahwa, N. (2021). Facebook asked users what content was "good" or "bad for the world." some of the results were shocking. *Slate*. https://slate.com/technology/2021/11/facebook-good-bad-for-the-world-gftw-bftw.html
Pan, C. A., Yakhmi, S., Iyer, T. P., Strasnick, E., Zhang, A. X., & Bernstein, M. S. (2022). Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction*, *6*, 82:1–82:31. https://doi.org/10.1145/3512929
Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). AI deception: A survey of examples, risks, and potential solutions. http://arxiv.org/abs/2308.14752
Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11). https://doi.org/10.1016/j.patter.2021.100336
People v. Zachariah C. Crabill, 23PDJ067 (2023). https://coloradosupremecourt.com/PDJ/Decisions/Crabill,%20Stipulation%20to%20Discipline,%2023PDJ067,%2011-22-23.pdf
Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C.,

Yan, D., Amodei, D., … Kaplan, J. (2023). Discovering language model behaviors with model-written evaluations. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 13387–13434). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.847

Pinka, R. (2021). Synthetic deliberation: Can emulated imagination enhance machine ethics? *Minds and Machines*, *31*(1), 121–136. https://doi.org/10.1007/s11023-020-09531-w

Prabhakaran, V., Mitchell, M., Gebru, T., & Gabriel, I. (2022). A human rights-based approach to responsible AI. https://arxiv.org/abs/2210.02667

Prabhakaran, V., Qadri, R., & Hutchinson, B. (2022). Cultural incongruencies in artificial intelligence. http://arxiv.org/abs/2211.13069

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

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435. https://doi.org/10.1145/3306618.3314244

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151. https://doi.org/10.1145/3375627.3375820

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7237–7256). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.647

Robb, M. B., & Mann, S. (2023). 2022 teens and pornography. *Common Sense*.

Saha, P., Garimella, K., Kalyan, N. K., Pandey, S. K., Meher, P. M., Mathew, B., & Mukherjee, A. (2023). On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, *120*(11), e2212270120. https://doi.org/10.1073/pnas.2212270120

Sandbrink, J. B. (2023). Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. https://arxiv.org/abs/2306.13952

Scheurer, J., Balesni, M., & Hobbhahn, M. (2023). Technical report: Large language models can strategically deceive their users when put under pressure [Apollo Research]. https://www.apolloresearch.ai/research/our-research-on-strategic-deception-presented-at-the-uks-ai-safety-summit

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. https://openreview.net/forum?id=Yacmpz84TH)

Seaver, N. (2021). Everything lies in a space: Cultural data and spatial reality. *Journal of the Royal Anthropological Institute*, *27*, 43–61. https://doi.org/10.1111/1467-9655.13479

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Shen, L., Chen, S., Song, L., Jin, L., Peng, B., Mi, H., Khashabi, D., & Yu, D. (2023). The trickle-down impact of reward (in-)consistency on RLHF. http://arxiv.org/abs/2309.16155

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023). Model evaluation for extreme risks. http://arxiv.org/abs/2305.15324

Siddarth, D., Acemoglu, D., Allen, D., Crawford, K., Evans, J., Jordan, M., & Weyl, E. G. (2021). *How AI fails us*. https://ethics.harvard.edu/how-ai-fails-us

Slack, D., Krishna, S., Lakkaraju, H., & Singh, S. (2023). Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, *5*(8), 873–883. https://doi.org/10.1038/s42256-023-00692-8

Small, C. T., Vendrov, I., Durmus, E., Homaei, H., Barry, E., Cornebise, J., Suzman, T., Ganguli, D., & Megill, C. (2023). Opportunities and risks of LLMs for scalable deliberation with polis. http://arxiv.org/abs/2306.11932

Stray, J. (2022). Designing recommender systems to depolarize. *First Monday*, *27*(5). https://doi.org/10.5210/fm.v27i5.12604

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html

Tantum. (2023). *Democracy on mars: Red sky thinking* [Democracy on Mars]. https://tantum.substack.com/p/democracy-on-mars-red-sky-thinking

Tegmark, M., & Omohundro, S. (2023). Provably safe systems: The only path to controllable AGI. http://arxiv.org/abs/2309.01933

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S. R., Das, D., & Pavlick, E. (2018). What do you learn from context? probing for sentence structure in contextualized word representations. https://openreview.net/forum?id=SJzSgnRcKX

Thorburn, L. (2022). *Is optimizing for engagement changing us?* [Understanding recommenders]. https://medium.com/understanding-recommenders/is-optimizing-for-engagement-changing-us-9d0ddfb0c65e

Thorburn, L. (2023). *When you hear "filter bubble", "echo chamber", or "rabbit hole" — think "feedback loop"* [Understanding recommenders]. https://medium.com/understanding-recommenders/when-you-hear-filter-bubble-echo-chamber-or-rabbit-hole-think-feedback-loop-7d1c8733d5c

Transatlantic Reflection Group on Democracy and the Rule of Law in the Age of "Artificial Intelligence". (2023). A manifesto on enforcing law in the age of 'artificial intelligence'. *European Law Journal*, *29*(1), 249–255. https://doi.org/10.1111/eulj.12474

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. https://openreview.net/forum?id=bzs4uPLXvi)

Viégas, F., & Wattenberg, M. (2023). The system model and the user model: Exploring AI dashboard design. http://arxiv.org/abs/2305.02469

Villasenor, J. (2019). *Products liability law as a way to address AI harms* [Brookings]. https://www.brookings.edu/articles/products-liability-law-as-a-way-to-address-ai-harms/

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. https://ntrs.nasa.gov/citations/19940022856

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, *34*(3), 229–243. https://doi.org/10.2190/FLHV-K4WA-WPVQ-H0YM

Vogels, E. A. (2022). *Teens and cyberbullying* [Pew research center: Internet, science & tech]. https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/

von Oswald, J., Niklasson, E., Schlegel, M., Kobayashi, S., Zucchet, N., Scherrer, N., Miller, N., Sandler, M., Arcas, B. A. y., Vladymyrov, M., Pascanu, R., & Sacramento, J. (2023). Uncovering mesa-optimization algorithms in transformers. http://arxiv.org/abs/2309.05858

Wajcman, J. (1991). *Feminism confronts technology*. Penn State Press.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). Sociotechnical safety evaluation of generative AI systems. http://arxiv.org/abs/2310.11986

Whistleblower laws around the world [National whistleblower center]. (N.d.). https://www.whistleblowers.org/whistleblower-laws-around-the-world/

Whistleblower protection [Office of inspector general]. (N.d.). https://www.oig.dhs.gov/whistleblower-protection

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136. https://www.jstor.org/stable/20024652

Yesilada, M., & Lewandowsky, S. (2022). Systematic review: YouTube recommendations and problematic content. *Internet policy review*, *11*(1), 1652. https://doi.org/10.14763/2022.1.1652

Yu, Q., Merullo, J., & Pavlick, E. (2023). Characterizing mechanisms for factual recall in language models. http://arxiv.org/abs/2310.15910

Zaremba, W., Dhar, A., Ahmad, L., Eloundou, T., Santurkar, S., Agarwal, S., & Leung, J. (2023). *Democratic inputs to AI*. https://openai.com/blog/democratic-inputs-to-ai

Zarlenga, M. E., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lio, P., & Jamnik, M. (2022). Concept embedding models: Beyond the accuracy-explainability trade-off. https://openreview.net/forum?id=HXCPA2GXf_

Zhang, H., Lee, I., Ali, S., DiPaola, D., Cheng, Y., & Breazeal, C. (2023). Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study. *International Journal of Artificial Intelligence in Education*, *33*(2), 290–324. https://doi.org/10.1007/s40593-022-00293-3