Harms from Increasingly Agentic Algorithmic Systems

Alan Chan*†‡
alan.chan@mila.quebec
Mila, Université de Montréal
Montréal, Canada

Rebecca Salganik†
Mila, Université de Montréal
Montréal, Canada

Alva Markelius†
University of Cambridge
Cambridge, UK

Chris Pang†
University of Cambridge
Cambridge, UK

Nitarshan Rajkumar†
University of Cambridge
Cambridge, UK

Dmitrii Krasheninnikov†
University of Cambridge
Cambridge, UK

Lauro Langosco†
University of Cambridge
Cambridge, UK

Zhonghao He†
University of Cambridge
Cambridge, UK

Yawen Duan†
University of Cambridge
Cambridge, UK

Micah Carroll†
University of California, Berkeley
Berkeley, USA

Michelle Lin
McGill University
Montréal, Canada

Alex Mayhew
University of Western Ontario
London, Canada

Katherine Collins
University of Cambridge
Cambridge, UK

Maryam Molamohammadi
Mila
Montréal, Canada

John Burden
Center for the Study of Existential Risk, University of Cambridge
Cambridge, UK

Wanru Zhao
University of Cambridge
Cambridge, UK

Shalaleh Rismani
McGill University, Mila
Montréal, Canada

Konstantinos Voudouris
University of Cambridge
Cambridge, UK

Umang Bhatt
University of Cambridge
Cambridge, UK

Adrian Weller
University of Cambridge
Cambridge, UK

David Krueger*
University of Cambridge
Cambridge, UK

Tegan Maharaj*†
University of Toronto
Toronto, Canada

ABSTRACT

Research in Fairness, Accountability, Transparency, and Ethics (FATE)[1] has established many sources and forms of algorithmic harm, in domains as diverse as health care, finance, policing, and recommendations. Much work remains to be done to mitigate the serious harms of these systems, particularly those disproportionately affecting marginalized communities. Despite these ongoing harms, new systems are being developed and deployed, typically without strong regulatory barriers, threatening the perpetuation of the same harms and the creation of novel ones. In response, the FATE community has emphasized the importance of *anticipating* harms, rather than just responding to them. Anticipation of harms is especially important given the rapid pace of developments in machine learning (ML). Our work focuses on the anticipation of harms from increasingly agentic systems. Rather than providing a definition of agency as a binary property, we identify 4 key characteristics which, particularly in combination, tend to increase the agency of a given algorithmic system: underspecification, directness of impact, goal-directedness, and long-term planning. We also discuss important harms which arise from increasing agency – notably, these include systemic and/or long-range impacts, often on

*Major contributions to the project direction and framing.
†Major contributions to the writing. Authors with only this mark have had their order randomized. Authors without a mark also have their orders randomized.
‡Correspondence to alan.chan@mila.quebec.
[1]We use the term **FATE** as a shorthand, keeping in mind and valuing the ideological diversity of those who work on FATE and related disciplines not captured in this acronym.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594033

marginalized or unconsidered stakeholders. We emphasize that recognizing agency of algorithmic systems does not absolve or shift the human responsibility for algorithmic harms. Rather, we use the term agency to highlight the increasingly evident fact that ML systems are not fully under human control. Our work explores increasingly agentic algorithmic systems in three parts. First, we explain the notion of an increase in agency for algorithmic systems in the context of diverse perspectives on agency across disciplines. Second, we argue for the need to anticipate harms from increasingly agentic systems. Third, we discuss important harms from increasingly agentic systems and ways forward for addressing them. We conclude by reflecting on implications of our work for anticipating algorithmic harms from emerging systems.

KEYWORDS

algorithmic systems, harms, safety, sociotechnical systems, negative externalities, agency, autonomy, power, delayed impacts, ethics, FATE

ACM Reference Format:

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3593013.3594033

1 INTRODUCTION

The promised benefits of algorithmic systems have not always borne out, and benefits are often tempered by significant negative externalities. Although the deployment of algorithmic systems may result in increased safety or material improvements to human well-being [6, 102, 118], diverse lines of work in Fairness, Accountability, Transparency, and Ethics (FATE) have established the roles that algorithmic systems play in causing harm. Examples include the perpetuation of existing, unjust power relations [19, 38, 60, 104, 174, 197], the generation of toxic language [7, 76], and informational harms [42, 99, 119, 193].

Despite the clear evidence of harms from existing systems, new types of algorithmic systems are continually being developed and deployed, often without strong regulatory barriers [77]. The pace of development has been particularly rapid in the machine learning (ML) community. Just in the last five years, we have witnessed large improvements in the capabilities of systems to perform a variety of real-world tasks, including search [133], drug discovery [102, 175], and dialogue [138].

Researchers in the FATE community have responded to the rapid pace of ML developments by emphasizing the need to *anticipate* harms, rather than just react to them. In particular, many have identified the impact of computational modeling and development in social change [5, 95, 163] and scoped numerous taxonomies of risks, harms, and failures of algorithmic systems [153, 167, 193]. While it is crucial not to idealize or over-hype a model's performance by ignoring model failures [23, 24, 28, 47, 120, 153, 187], it is also important not to understate (and thus fail to anticipate negative consequences of) what these models *can do* and *may be capable of doing* in the near future [31, 90, 103], especially given growing investments in the field [78].

In this paper, we continue the work of anticipating harms by drawing attention to increasingly agentic algorithmic systems. We use agency and agentic in a narrow sense for our work as applied to algorithmic systems, particularly ML systems. While recognizing the many meanings of **agency**, as well as the need not to absolve humans of responsibility pertaining to algorithmic harms [48, 134, 196], we use the term agency consciously to counter the somewhat prevalent view that the developers of an algorithmic system have full control over its behaviour. E.g. Johnson and Verdicchio [100] claim that "the behaviour of computational artefacts is in the control of the humans that design them." And in a systematic review on algorithmic accountability, Wieringa [196] defines algorithms as "basically instructions fed to a computer". While this description is accurate for many purposes, we argue that, particularly for ML-based algorithmic systems, it elides autonomous, responsive, and interactive qualities of these systems which can so easily lead to unforeseen outcomes. Cooper et al. [48], Nissenbaum [134] do identify bugs – including faulty modeling premises and bad model performance – as one way in which humans may not have total control of the operation of an algorithmic system. However, we view agency as distinct from mistakes or bugs and demonstrate the unique and important harms that can result. We note there are significant economic and military incentives to build increasingly agentic systems. Indeed, many in the ML community are explicitly building such systems as a research goal [44, 155, 179].

In summary, our contributions are:

(1) We identify characteristics that tend to increase agency of algorithmic systems, and situate our characterization in the context of diverse perspectives on agency across disciplines. We articulate that even when recognizing agency in algorithmic systems, we can and should emphasize the human responsibility to prevent harms.
(2) We argue for the need to anticipate harms from increasingly agentic systems. Increasingly agentic systems are being developed and there exist strong incentives for this work to continue.
(3) We discuss some harms to be anticipated from increasingly agentic systems. In so doing, we connect to ongoing lines of work in the FATE community, including systemic and delayed effects, an impoverishment of collective decision-making power, and exacerbation of extreme concentrations of power in the hands of a few. We also discuss the role of increasing agency as a source of harms that are yet to be identified.

This paper is **not** about the moral agency or consciousness of algorithms or machines. Instead, we focus on identifying a property of emerging ML systems, argue for the need to anticipate harms from systems that increasingly satisfy this property, and discuss the harms to be anticipated.

2 AGENCY

In colloquial use, agency refers to the ability to take actions or affect outcomes. A difficulty of having concrete discussions on agency is the variety of perspectives through which such a concept can be defined, making confusion and disagreement common. In recognition of this variety of perspectives, we do not attempt to define agency in a binary manner, but instead present a set of characteristics we take to be associated with *increasing* agency, i.e. the more of these characteristics a system has, particularly in combination, the more agency we can consider it to have. We first present our characterization, and follow by contextualizing it in some of the most relevant perspectives and related concepts to our work.

2.1 Characteristics that are Associated with Increasing Agency in Algorithmic Systems

When we say that an algorithmic system has a degree of agency, we mean that it is to some extent an agent or agentic. **Agency** is the property, **agent** is the role, and **agentic** is the adjective. Our characterization of agency is specific to algorithmic systems and is not meant to define agency for humans or arbitrary entities. We will sometimes use "agentic system" in place of "agentic algorithmic system" for brevity.

We identify 4 key characteristics associated with increasing agency in algorithmic systems, especially in combination: underspecification, directness of impact, goal-directedness, and long-term planning.

(1) **Underspecification:** the degree to which the algorithmic system can accomplish a goal provided by operators or designers, without a concrete specification of how the goal is to be accomplished [53].
(2) **Directness of impact:** the degree to which the algorithmic system's actions affect the world without mediation or intervention by a human, i.e. without a human in the loop.
(3) **Goal-directedness:** the degree to which the system acts as if it is designed/trained to achieve a particular quantifiable objective.
(4) **Long-term planning:** the degree to which the algorithmic system is designed/trained to make decisions that are temporally dependent upon one another to achieve a goal and/or make predictions over a long time horizon.

To illustrate the notion of increasing agency, consider the task of compiling a literature review on a certain subject. With a search engine, the human user must type in queries, click on related works, read papers, look through bibliographies, record relevant information in a document, and edit the text. A system that was more agentic than the search engine, still for the same task, could simply be queried with the topic of the desired literature review, and would automatically look through related works on the internet without user intervention, like WebGPT can do to some extent [131]. The user would not need (or be able to) to specify which papers were relevant nor have to compile papers manually into a document.

2.2 Prior Work on Agency

Agency is a central concept in many fields of academia [159]. Dennett [56] provides one of the most popular analyses of when and how to attribute agency, focusing on the notion that agents behave intentionally. Orseau et al. [140] and more recently Kenton et al. [108] have attempted to formalize this notion of agency in the context of artificial intelligence. In cognitive science and psychology, agency is conceptualized relatively similarly, as having intentions, plans, goals, communication, and reasoning [113, 171] – entities with agency can plan, act, memorize, exert self-control, and communicate with others. While these notions of agency focus on individuals making rational choices in pursuit of some goal, in sociology, agency is often thought of as contextualized within, constrained by, and/or contrasted with structure [64].

Principal-agent theory [61, 96] provides more intuition for how we characterize agency. Principal-agent theory concerns itself with a *principal* who delegates tasks to an *agent* in order to achieve their goals. The agent acts (directness of impact) on behalf of the principal to achieve the principal's goals, which may be long-horizon (long-term planning). Crucially, the agent and principal have different incentives[2] and information: the principal does not tell the agent how to complete the tasks (underspecification). In our context, we view the principal as humans and the agent as algorithmic systems, as done in prior work [86]. It is in this sense that we consider algorithmic systems to have agency.

Our notion of increasing agency also takes inspiration from how the term agent is used in AI research. In the most popular introductory text on artificial intelligence, Russell and Norvig [158, p. 58] define a **rational agent** as follows: "For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has." Russell and Norvig [158, p. 60] further states that "To the extent that an agent relies on the prior knowledge of its designer rather than on its own percepts and learning processes, we say that the agent lacks autonomy. A rational agent should be autonomous—it should learn what it can to compensate for partial or incorrect prior knowledge." While our characterization does not consider (ir)rationality, goal-directedness and underspecification are captured in this definition.

Reinforcement learning is a field that concentrates on the construction of agents. In the field's premier introductory text, Sutton and Barto [178, p. 47-8] states that the "learner and decision maker is called the agent. The thing it interacts with, comprising everything outside the agent, is called the environment. These interact continually, the agent selecting actions and the environment responding to these actions and presenting new situations to the agent. The environment also gives rise to rewards, special numerical values that the agent seeks to maximize over time through its choice of actions." Note that reinforcement learning is not the only way of constructing agents, however. For instance, recent work has shown that foundation models can perform planning tasks [93]. Even simple predictive algorithms, depending on their training procedure, can follow incentives to affect the world in unexpected ways – for example by shifting user interests rather than improving at their predictive task [111], thus increasing their agency.

One of the difficult discussions surrounding agency is interaction of agency and responsibility, for humans and for machines. Goetze [80] identifies a **responsibility gap** between engineers

[2] It is coherent to talk about the incentives of algorithmic systems. See Everitt et al. [68].

and the outcomes of their designed systems – people designing autonomous systems are far removed from the consequences of their deployment. The authors contend that regardless of the system's autonomy, human designers must be the ones held accountable. [176] investigate attitudes toward agency of fictional AI robots, and find survey respondents do not typically consider AI systems to be moral agents – they tend to place moral responsibility on developers, not on AI systems as agents. Similarly, Robinette et al. [157] examines (over)trust of autonomous systems and find in emergency situations, people will follow robots into further danger, because they attribute the agency of the robot to the (assumed capable and responsible) designers. As people in these situations appear to, we distinguish agency from responsibility, and emphasize that attribution of agency to an autonomous system in no way shifts moral responsibility from humans onto that system. Leufer [116], Myths [130] examine another aspect of this problem, describing AI **agency as a myth** which masks human agency (and therefore responsibility). The authors contend that anthropomorphization of AI systems contributes to mystification of the underlying technology and sociotechnical blindness [100], wherein people "believe AI systems got to be the way they are without human intervention", and obscuring of the (often exploitative) human labour which enables AI systems to exist [82, 144, 147]. While we strongly agree with all these points, we reach the opposite conclusion – AI agency (in the sense of our work) is not a myth, it is a reality of increasing sociotechnical importance. It is precisely because of the importance of problems like these (responsibility gap, mystification, sociotechnical blindness, masking human agency and labour, etc.) and their far-ranging implications that we need to carefully examine the agency of AI systems, not dismiss it out of hand. If we think it is categorically impossible for AI systems to have agency, we will never be able to accurately recognize when we are giving up our agency to them.

A related concept we wish to distinguish from agency is autonomy. In our framework, autonomy corresponds most closely to directness of impact, with some overlap in the three other categories. While it is often an intuitive or useful description of a system, we find it combines distinct phenomena we wish to distinguish with our characteristics. Bekey [22] defines **autonomy** as "the ability to operate without a human operator for a protracted period of time." Many factory robots are highly autonomous, but they operate strictly within the confines of a factory, and the actions they take affect only the intended outcome (e.g. the product they're making) – they are autonomous but do not have agency. Welsh [194] presents a series of protocols which can be used to govern the use of lethal autonomous weapons, emphasizing the need for human-in-the-loop decision making – i.e. to ensure all agency rests with human controllers.

In this vein, our focus on agency also shares many commonalities with work from the FATE community on establishing the harms of **automated decision-making (ADM)**. ADM involves the use of algorithms to make decisions or enact policies without human intervention. Given its applications in recommendations [119, 127], health-care systems [70, 135, 164], the judicial sector [18, 83, 205], and public services [27, 122, 173], ADM can often exhibit similar kinds of diffuse and long-term harms to those we discuss coming from increased agency. Given the commonalities, many of the harms of ADM also apply to increasingly agentic systems, as we discuss in Section 4. But there are two key differences between the body of work on ADM and our work. First, with the term agency we emphasize lack of explicit or low-level instructions for behaviour - we might specify a task, but not *how* to solve that task. Second, our work explicitly targets systems that are *increasingly* agentic, such as reinforcement-learning systems that are capable of making decisions in an open-ended environment over long time horizons without human intervention. Such systems have not been the focus of work in ADM simply because they have not yet seen widespread public deployment. We thus consider our focus on agency to be a continuation of current work on ADM.

Some philosophical work on agency also focuses on mental states such as consciousness, emotions, and subjective experience [159]. Entities with these mental states have personalities, and feel things like pleasure, curiosity, pain, embarrassment, fear and joy. Our work does not address experience or consciousness, only agency.

2.3 Potential Objections to our Use of Agency

One objection against framing algorithmic systems as agents is that it distracts from the responsibility of humans. As noted above, we characterize agency as separate from responsiblity. As many authors suggest, we strongly agree that attention should be directed towards holding corporations, regulators, developers, etc. (*actors* for short in this section) accountable [48, 100, 134, 196]. This claim is not in contention with the idea that algorithmic systems can be agentic in our narrow sense. Principals (actors) can be held responsible on behalf of their agents, such as when employers are held liable for negligent hiring when employees cause harm [88].

We should also require more than just individual accountability. In addition to focusing on individual actors, we should also attend to structural factors that shape their behaviours. A developer is likely blameworthy at least to some extent when a system causes harm, but structural factors like economic incentives or company culture to push forward likely also play significant roles [186, 206]. As we will discuss in Section 4, viewing algorithmic systems as agents can in fact highlight harms and the collective responsibility we have to prevent them.

3 THE NEED TO ANTICIPATE HARMS FROM INCREASINGLY AGENTIC SYSTEMS

We argue for the need to anticipate harms from increasingly agentic systems. Anticipation is about two things: (1) the development of systems with increasing agency and (2) the deployment of systems with more agency than those already deployed. We touch upon trends in ML development and deployment as well as some reasons to expect these trends to continue. In Section 3.3 we respond to some potential objections.

3.1 Trends in Development and Deployment

We aim to show two things in this section. First: development of increasingly agentic systems has proceeded by consistently overcoming technical challenges. Second: deployment of increasingly agentic systems has occurred because these systems have increasingly practical skills that are useful for real-world applications.

3.1.1 Overcoming Technical Challenges to Build Increasingly Agentic Systems. Reinforcement learning (RL), as one of the major paradigms of machine learning, has a major focus on the construction of agents [178, 179]. In particular, RL is about designing systems to learn, without human intervention, to achieve a goal encoded in a reward function. Prior to 2013, RL systems were developed mainly for a restricted set of simple domains [49]. The introduction of deep learning to RL systems produced superhuman performance on a wider variety of narrow tasks with limited to no human supervision, including but not limited to increasingly complex board games [35, 146, 168, 169] and video games [129, 161, 199]. Subsequent work has greatly improved the performance of RL systems on more complex, open-ended environments. For instance, DreamerV3 [87] collected diamonds from scratch without human data or curricula in MineCraft, which has been a longstanding challenge because the task is extremely complex and open-ended. Another striking example comes from Diplomacy, a complex, multiplayer board game involving tactical coordination and natural language negotiation. The recent Cicero [17], integrating a language model with planning and RL algorithms, is the first AI to achieve human-level performance in Diplomacy. Such systems have demonstrated strong capabilities to interact with complex environments and humans to accomplish their goals that require long-horizon planning.

We emphasize that for all the systems we have mentioned in this section so far, designers do not specify how the tasks were to be completed. In the current scientific paradigm of large-scale deep-learning, one instead provides high-level learning algorithms that tend to be task-agnostic or adapt to new tasks efficiently [21, 57]. One particular example to highlight is AdA [181], which adapts to open-ended, novel, embodied 3D problems as quickly as humans, without human specification of how to solve problems.

3.1.2 The Increasing Deployment of Increasingly Agentic Systems. The practicality of systems has increased along with their agency. Increasing practicality means that increasingly agentic systems are more likely to be found making decisions in the real world. Major companies have been deploying increasingly agentic systems to control parts of their operation. For example, DeepMind and Google use RL for controlling commercial cooling systems and data centers [67, 105]. Amazon has applied RL to supply chain optimization problems [160]. Additionally, there has been an increasing amount of research in recommender systems to optimize long-term metrics such as engagement via reinforcement learning [10]. Major recommendation companies such as Meta [74], YouTube [16], and Spotify [65] have already deployed RL-based recommender systems on their live products.

Systems that can competently operate across different data modalities and tasks are plausibly more useful than more narrow systems, regardless of how agentic they are. Before the current era of large language models (LLMs) [30, 36, 150, 172], few systems competently performed out-of-the-box on a range of natural language tasks [36]. Recent models [12, 155, 202] can even handle multiple data modalities simultaneously. GATO [155] can complete tasks using the same model and weights in vastly different domains, such as Atari, image captioning, dialogue, and robotics. As systems become increasingly agentic, the systems that are increasingly domain general seem likely to see more practical application.

Increasingly agentic systems are also becoming more available to the general public. Although language models are only trained on next-token prediction, they can be leveraged to interact with APIs and accomplish a wide variety of multi-step digital tasks with increasingly less explicit human intervention [43, 125, 131]. Adept's ACT-1 [9] is a system in development which purportedly can perform an arbitrary task on your computer, such as searching for and buying an item online, through a single text command. OpenAI's ChatGPT [139] has plug-ins that can interface with a wideo variety of web-apps, including Gmail and a web browser. AutoGPT [3] chains together an arbitrary number of GPT-4 calls to accomplish a high-level task on one's computer without one's intervention.

Despite the progress so far, systems still have limitations and there are still barriers to the deployment of more agentic systems. For example, the raw task performance of generalist systems [155, 202] is still limited to tasks where expert data is available and has not achieved human level on all tasks. In the realm of language models, recent studies [98, 185] have shown that large language models can perform poorly on planning and reasoning tasks, and such systems are prone to hallucinate unintended text, which fails to meet users' intents on many real-world scenarios. However, we note the pace of development and deployment are still rapidly increasing, not decreasing. We anticipate that current limitations and barriers will be surpassed or ignored in the pressure to deploy.

3.2 Factors in the Continued Development and Deployment of Increasingly Agentic Systems

For current AI models, there are strong incentives for continued investment and development despite uncertainty around how their future capabilities will emerge [72]. Similarly, a number of reasons suggest the potential for development and deployment of increasingly agentic algorithmic systems. These factors are the economic and military advantages afforded by increasingly agentic systems, scientific curiosity and prestige, a lack of regulatory barriers, and emergent agency. The first three reasons are sociopolitical, while the last reason concerns potentially surprising technical properties of ML systems. Taken together, these increase our subjective likelihood that systems will become increasingly agentic.

3.2.1 Economic Incentives. Actors who deploy more agentic systems than their competitors would likely generate more profit because of increased automation. First, more agentic systems might be able to perform tasks much more cheaply than a less agentic systems. A less agentic system by definition would require more human intervention, whether to make decisions or specify explicit procedures for task completion. Second, more agentic systems will often be more effective at performing tasks than less agentic systems. Part of an increase of agency is the degree to which a system achieves a goal without operators or designers to specify how. The upshot is that the search space of solutions to a problem is larger for a more agentic system, which could result in solutions that would be much more efficient than those a human could have found. That AlphaGo [168] beat Lee Sedol, the world Go champion, with the apparently confusing move 37 is evidence of this possibility.

3.2.2 Military Incentives. Militaries may perceive that increasingly agentic systems could provide capability advantages over adversaries that are constrained by human decision-making. The introduction by any one military of a more agentic system could upset a balance of power and force others to pursue similar developments in an unsafe race to the bottom, mirroring other races for technologies such as nuclear weapons, and ballistic and hypersonic missiles [51]. The UK's defence AI strategy [1] frames advances in AI as being an area of "geostrategic competition" and "a battleground for competing ideologies", but also imposes no governance or oversight mechanisms on increasingly agentic systems, focusing such efforts "on effects rather than the nature of any particular technology." Total bans on developments for such highly autonomous systems may be difficult to introduce and maintain, and it may be easier to pursue nonproliferation of such advances beyond a small set of technologically advanced users [183].

3.2.3 Scientific Curiosity and Prestige. Developing increasingly agentic systems is an object of scientific curiosity and also confers status, a motivation that contributes to a prestige race at varying levels between actors in the AI research system. For individual researchers this emerges through standard metrics such as paper publications and grant awards that support climbing the academic career ladder, but for many leading figures the ambitions transcend these: Geoffrey Hinton – a pioneer deep learning research – has stated that "the prospect of discovery is too sweet" in spite of his beliefs that "political systems will use [AI] to terrorize people" [109], and Rich Sutton – a pioneer in reinforcement learning – has stated that creating "beings of far greater intelligence than current humans" (that would necessarily be agentic) will be "the greatest intellectual achievement of all time" and "a great and glorious goal" [177]. For companies, developing increasingly agentic systems could drive the most impactful research and development outputs, increasing attractiveness to the best scientific talent in a competitive hiring pool. For nations, highly visible scientific demonstrations may act as demonstrations of broader state capacity, and prestige may be as motivating a force as security for competitive races with peers and adversaries [20]

3.2.4 Lack of Regulatory Barriers. Regulatory efforts for AI have focused largely on salient risks, rather than on anticipatory governance mechanisms that are proactive to future advances in AI capabilities [77]. For example, the EU AI act currently proposes to target regulation according to tiers of risk determined by type of data use and deployment setting, and efforts in the UK take a sectoral focus on regulating only the applications of AI, but neither covers development of agentic AI systems that could both be intrinsically high-risk and could underlie progress and use across a variety of sectors and domains [2, 59]. Accordingly, development and deployment in this space is effectively unregulated and without any clear possibility of regulation in the near future.

3.2.5 Emergent Agency. Even if designers do not explicitly build more agency into their systems, it may emerge from general capability improvements. Recent works discuss emergent behaviors of large language models. Bommasani et al. [30] introduce emergence as a "behavior of a system [that] is implicitly induced rather than explicitly constructed; it is both the source of scientific excitement and anxiety about unintended consequences." For example, LLMs are trained to model a distribution of internet text; this training leads to *emergent behavior* such as learning from very few examples [136], or arithmetic [36], or even the ability itself to perform sequential reasoning [192]. Many of these abilities only emerge at a certain scale, or after a certain point in the training process [191].

When emergent behavior increases the agency of a system we can speak of *emergent agency*. One particularly striking example is the ability of LLMs to simulate the **human agents** who are the sources of the training data. For example, maraoz [124] uses GPT-3 to write a transcript of a conversation between themselves and Albert Einstein, and others have used LLMs to retroactively simulate user studies from psychology and economics [11]. The seeming fidelity of such texts has motivated some to argue that LLMs have a general ability to simulate human agents [14].

Some emergent capabilities relate directly to our characterization of agency. Wei et al. [192] show that adding "let's think step-by-step" vastly improves sequential reasoning capabilities in LLMs, a capability which is useful for performing tasks over long time horizons. Team et al. [181] show that scaling up a particular approach leads to RL systems that capably adapt to open-ended, novel 3D problems as well as humans can, without human intervention on how to solve the problem.

3.3 Potential Objections to our Characterization of ML Progress

3.3.1 The Need for Anticipation of Increasingly Agentic Systems is Small. Earlier, we distinguished between two things to anticipate: **(1)** the increasing agency of developed systems and **(2)** the deployment of systems with more agency than those already deployed. We respond to objections against both points.

One could accept the need for attention to **(1)**, but maintain that the need is small given that technical improvements to increase agency occur much more slowly than we have characterized. Indeed, past beliefs in rapid pace of artificial intelligence research have been overoptimistic [58]. Barriers to increasing agency include acting capably over long time horizons [185] and with an accurate understanding of the world [24]. These challenges are real and there is by no means any certainty that the ML research community will overcome them. Even the perceived agency of algorithmic systems depends heavily on (sometimes exploitative) human labor and data extraction [82, 144]. Moreover, it can be difficult to measure the rate of progress towards agentic systems. Dehghani et al. [55] provide evidence that factors other than "fundamental algorithmic superiority" may lead to the perception that a particular method is superior. Raji et al. [151] discuss several issues with benchmarking, including construct invalidity and limitations in scope.

We have no disagreements on the technical challenges of developing systems of increased agency. We are also not claiming that systems of significantly greater agency than those in development already (e.g., compared to ACT-1 [9], GATO [155]) will be coming soon. Rather, our view is that even absent significant technical breakthroughs, continued work within the current scientific paradigm [112] of scaling deep-learning seems likely to generate systems that are appreciably more agentic than current systems. Scaling laws provide predictable relationships between the amount of compute and data used to train model of a given size, and the performance of a model on some metric. Of particular interest for increasing agency is that scaling laws have been derived for reinforcement learning [73, 89, 181] and generative modeling [90, 103]. There is also initial work into developing scaling laws for robotics [40]. The upshot is that continued training of larger models with more compute and data seems likely to increase the ability of systems to act in environments of increasing scope, over longer time horizons, to achieve goals without significant designer/operator intervention.

One could also object to the need for attention to (2). Even if a system that is more agentic than those currently deployed has been developed, there might still be strong reasons against deployment, despite the incentives in Section 3.2.1. Raji et al. [153] argues that deployed AI systems often simply do not work, suffering from issues such as robustness failures, missing safety features, or being set to perform impossible tasks (such as inferring criminality from appearance). Given that increasingly agentic systems would be more capable of achieving goals without human specification of how, the failures that Raji et al. [153] highlight could disincentivize adoption of increasingly agentic systems, even if they were developed.

The likely failures of a more agentic system (relative to what has been deployed already) are certainly a barrier to deployment – given the disproportionate impact of these failures on already marginalized groups, we would hope and advocate for restrictions on deployment [37]. However, we think that this barrier is unfortunately weak relative to countervailing forces. Hype around the (claimed) functionalities of ML systems is strong [34, 132, 162], which is unsurprising given massive financial investments [78]. Continued cycles of deployment and failure [18, 38, 135, 148, 156, 190, 197] suggest that increasingly agentic systems will likely be deployed according to industry interests, and not the interest of those most likely to be harmed.

3.3.2 Techno-Determinism. An objection against our characterization of increasingly agentic systems is that it is techno-deterministic – it assumes that AI development is inevitable and determines the direction of sociocultural development [198]. This objection comes in two parts. Firstly, the perceived inevitability of ML progress nullifies accountability of those developing the systems and removes reason to regulate or stop development. Secondly, techno-determinism neglects social and cultural structures and implies a reductionist view of the harms caused by ML systems. Related to this is the adoption of discourse around ML systems that their capabilities are both scientifically impossible to explain, and yet deterministic in their societal impact [41]. Some who study the harms of more agentic systems have also been accused of techno-optimism – optimism about the potential of technology to solve major social problems – and techno-determinism [50]. The problems include a disproportionately high reliance on technological solutions and neglect of insights from structural aspects of risk-analysis.

We do not dispute the dangers of techno-determinism or techno-optimism. However, careful work on identifying and mitigating harms of increasingly agentic systems need not rely on or contribute to either. For example, one can be engaged in activism to ban specific uses or developments of increasingly agentic AI, while concurrently pursuing sociotechnical research to mitigate those systems' potential harms. In this framing, the sociotechnical work can be seen as an attempt to reduce harm in the case that one's broader attempts to change the field's course of action are not successful. While it can be argued that working on such harm reduction contributes to perceptions of inevitability or deployment incentives, being thoughtful in the framing of one's work can significantly contribute to avoiding this issue.

4 ANTICIPATED HARMS FROM INCREASINGLY AGENTIC SYSTEMS

The previous section argued for the need to anticipate the harms of increasingly agentic systems. We now delve into some of these harms and why they are of especial importance for the FATE community.

4.1 Systemic, Delayed Harms

A systemic harm is a harm that is pervasively embedded in society. A delayed harm is a harm whose cause has a non-immediate impact. Systemic, delayed harms from algorithmic systems negatively influence groups of people in non-immediate ways. While harder to analyze than immediate harms, systemic and delayed harms might also be more insidious, as they can be caused even by low-stakes decision making systems. Each action might not seem consequential on its own, but, in aggregate, the outcomes can be destructive, long-lasting, and hard to fix. For example, there has recently been evidence that a single rent-setting algorithm might have significantly contributed to an increase in housing rental costs across the US [188].

The FATE community has studied systemic and delayed harms in the past, such as environmental risks [23], concentration of power [4, 149], unfair algorithmic hiring decisions [180], and privacy infringements [62]. Another line of work focusing on the long-term fairness implications of decisions [25, 54, 94, 101, 121, 203]. More broadly, many have identified the systemic nature of general classes of harms, such as financial risk [15], racism [33], and misogyny [123].

Social media is speculated to be a contributing factor to many systemic and delayed harms, including mental health issues [91, 200], the amplification of political polarisation [195], and the spread of fake news [13]. There is evidence on both sides for many of these issues [32, 106, 117], but caution seems warranted due to the sheer scale of these platforms (e.g., Facebook has almost three-billion users [126]).

While many of these harms do not involve the use of algorithms that are trained to act over long time horizons [99], the application of reinforcement-learning based recommendation systems (RLRS) in today's social media platforms warrant additional reason for concern. In particular, Carroll et al. [42], Evans and Kasirzadeh [66], Krueger et al. [111] show that long time-horizon systems, such as RLRS, will have incentives to change or manipulate users' internal states (e.g. preferences, beliefs, and psychology) for the purposes of increasing the metrics the RLRS systems are optimizing. While some work has also investigated potential solutions [42, 69], how to practically measure and address these issues in real-world RLRS remains an open problem. Notably, such systems are not speculative: RLRS are now increasingly applied by major social media providers (such as YouTube or Facebook), as discussed in Section 3.1.2.

4.2 Collective Disempowerment

We take **collective self-governance** to be the capacity and ongoing act of deciding collectively how to govern one's community, whether it be a small, local community, a state, or human societies at large [45]. Collective self-governance is about power, which is a core theme in FATE work [4, 18, 26, 29, 104, 201]. Indeed, the FATE community has extensively studied the ways in which automated decision-making can disempower individuals, by impairing human decision-making [84, 85] or subjecting individuals to oppressive institutions [18, 19, 83, 205]. We extend this ongoing discussion by pointing to some ways in which increasingly agentic systems can result in collective disempowerment. A key underlying point will be that increasingly agentic systems will likely seem more capable of handling more important societal functions without significant operator or designer intervention, as we discussed in Section 3.2.1. We discuss two possibilities: a situation in which power diffuses away from all humans, and a situation in which power concentrates in the hands of a few.

4.2.1 Diffusion of Power Away from Humans. As systems become increasingly agentic, they have increasing control of societal functions in many ways. At one end, humans in a particular social structure may decide to cede decision-making power to an particular system, such as one that decides taxation policy [204]. At the other end, power may gradually be ceded, as separate groups are incentivized to delegate more central functions to increasingly agentic systems per Section 3.2.1. Cooper et al. [48], Nissenbaum [134] examine the erosion of accountability that externalizes algorithmic harms. Even if collective disempowerment is a risk, it might not be a large enough risk for a single party to be concerned. In either case, ceding decision-making power to such systems is *not* inevitable; it would be a result of collective human decisions.

Regardless of how power is ceded, any group might have increasing difficulty in controlling increasingly agentic systems. Specifying a correct objective function is quite difficult [110, 170]. Even a system successfully trained under a correctly specified objective function may do something completely different in a different environment [114, 165]. Additional problems remain in understanding how to manage the interests of multiple stakeholders [52]. As well, it would likely be extremely difficult to understand the decisions of the controlling system(s). Analysis of a single decision is likely insufficient for understanding the reasons for a series of long-term decisions (i.e., the overall plan). Collective self-governance requires not just having decisions be made, but understanding why those decisions are made, which Lazar [115] terms the publicity requirement. Lazar [115] argues that failure to satisfy this requirement delegitimizes the exercise of political authority, by nullifying the moral effectiveness of consent.

4.2.2 Exacerbating the Extreme Concentration of Power Amongst the "Coding Elite". The FATE community has highlighted the concerning ways in which the deployment of algorithmic systems has concentrated power in the hands of designers and/or operators. Kasy and Abebe [104] argues that common notions of fairness legitimize hierarchies that are the result of historical injustice. They also provide a framework to reason about the impact of algorithmic decisions on the distribution of power. Burrell and Fourcade [39, p. 217] identifies the **coding elite** – a nebula of software developers, tech CEOs, investors, and computer science and engineering professors, among others, often circulating effortlessly between these influential roles – as a main beneficiary of the concentration of power. According to Burrell and Fourcade [39], the coding elite concentrates power by controlling the algorithms underlying the modern digital world, using that power to affect politics for their own gains. The amount of control exerted is already substantial with existing algorithmic systems, considering the centrality of the products of a handful of tech companies in our daily lives.

Increasingly agentic systems threaten to exacerbate an already extreme concentration of power. First, Ganguli et al. [72, p. 11] show that the proportion of large-scale ML results from industry has dominated in the past few years. The importance of large-scale results for increasing agency is that, as we discussed in Section 3, scaling up the compute, data, and parameters of a system provides a significant way to increase its agency, and is in some sense easier than deriving fundamental algorithmic insights. It therefore seems plausible that large industrial labs will continue to be the ones who deploy and profit the most from increasingly agentic systems. Second, increasingly agentic systems would likely enable the coding elite to integrate algorithms into more of society. There are many tasks now that are yet outside the reach of algorithmic systems, such as deciding national economic policy or running a business. Increasingly agentic systems seem more likely to be able to assume many of those tasks than current systems.

4.3 Harms Yet To Be Identified

In Section 3.2.5, we identified emergent behaviours as a possible cause of increasingly agentic algorithmic systems. Here, we explain some emergent behaviours that could be the source of harms that have yet to be identified and the link of those behaviours with increasing agency.

4.3.1 Reward Hacking. An RL system trained to maximize its score in the video game CoastRunners will drive off-track and keep turning in circles forever, thus achieving a high score despite not completing the race-track as intended by the programmers [46]. This kind of failure is called **reward hacking** [110, 170], which is when a system exploits a reward signal to achieve a goal in an unforeseen, perhaps undesirable way. As an instance of Goodhart's law [81], reward hacking is a common problem in ML systems that involve elements associated with increasing agency, in particular goal-directedness.[3] Increased model size or training time can result in abrupt increases in reward hacking, because a more capable model is better able find unforeseen maxima of its reward function [141].

If increasingly agentic systems are deployed in consequential domains like finance, health care, and law, reward hacking could result in extremely negative outcomes. Even with knowledge of the possibility of reward hacking, designers might still deploy systems

[3] A large number of examples of reward hacking are compiled in this online spreadsheet.

anyway if the harms from their systems are externalized, or if they judge the immediate likelihood of reward hacking to be low.

4.3.2 Instrumental Goals. An **instrumental goal** is a goal that is useful as a subobjective in pursuit of a specified goal. A **convergent instrumental goal** is a goal that would be useful in pursuit of a wide range of possible goals. For example, acquiring money is a convergent instrumental goal since money increases economic power and optionality. Many convergent instrumental goals involve gaining some sort of power over the environment and other actors within it [137].

An algorithmic system that sought to gain power over other actors, such as through manipulation or threats [107], would be concerning. An additional concern would be if the same thing were to happen without explicit, malicious instructions from their designer(s) or operator(s) to perform such behaviour. While this possibility remains uncertain, some initial evidence does not dismiss it. Perez et al. [145] show that increased training of a LLM with RL techniques can increase the proportion of the time that the LLM expresses the pursuit of convergent instrumental goals, such as gaining wealth and persuading the operator not to shut it off, without any apparent designer or operator instruction to do so. To recall, RL is about the construction of agents, by training systems to act over long time horizons to achieve goals without explicit human intervention. Training LLMs with RL techniques plausibly increases their agency; therefore, Perez et al. [145] provides some evidence that increasing the agency of LLMs can be associated with an increase in the expression of convergent instrumental goals. It is important not to overstate this early evidence; expressing a desire to pursue a goal is different from actually pursuing the goal in the world. Yet, this evidence should be taken as an additional reason for caution regarding increasingly agentic systems.

5 PATHS TO PREVENTING HARMS

Much work remains in figuring out how to address the present need we have highlighted throughout our piece. We provide a preliminary discussion of some directions and tie them to existing work from the FATE community.

5.1 Investigating the Sociotechnical Attributes of Increasingly Agentic Systems

Several landmark works in the FATE community have involved audits [154] of algorithmic systems [27, 38, 135, 156]. Audits have motivated action from designers to reduce the harms of their systems [152].

Since one typically audits a deployed system, it will be difficult to perform thorough audits of increasingly agentic systems before they are widely deployed. Nevertheless, there are a variety of ways to reason about the potential impacts of a system's deployment. Assuming either that we have the system in question or that we can simulate it faithfully [11, 63, 143], we can formulate and test hypotheses in simple experiments or simulation. Small-scale studies and simulations will almost certainly fail to capture perfectly what would happen if the system was actually deployed on a large. Nevertheless, such investigations can also highlight potentially concerning phenomena for further investigation. Failing to observe harm should not necessarily be taken to mean that a system is safe; on the other hand, observation of a potential harm in a pre-deployment study should motivate further research into understanding to what extent the harm would appear in practice.

More broadly, it might be possible to take inspiration from policy-making techniques such as scenario planning [189], which involve thinking ahead about how to make effective policy decisions when uncertain about what the world will look like in the future. The emerging science of forecasting [182] may also provide insights into anticipating the impacts of emerging systems.

Other tools from the FATE community may be helpful for characterizing the sociotechnical attributes of increasingly agentic systems, even before widespread deployment. For example, datasheets [75] and model cards [128] can highlight sources of harm, such as accountability gaps [134], in a way that does not depend upon a particular application. In the same vein, Gilbert et al. [79] introduces reward reports to document what it appears that systems are optimizing for, which may help to reduce the likelihood of unintended negative consequences from system operation. Interpretability work may also help us understand how a system is achieving a goal [8, 136].

Another promising line of work is to propose both quantitative and qualitative metrics that build upon our characterization of agency. For instance, we could have metrics for measuring the degree to which an AI system is capable of accomplishing tasks in the real world. Having metrics for agency would facilitate the study of when agency causes or is correlated with observed negative impacts. Some existing work already measures aspects of agency, such as long-term planning [185] and goal-directedness [142].

5.2 Regulatory and Institutional Interventions

Stronger regulations could prevent some harms of increasingly agentic systems from occurring. Compute limits enforced by compute usage tracking [37, 166], while a source of serious privacy risks, could help to control the pace at which systems become increasingly agentic and permit more time to develop mitigations. Along this line, it might be collectively beneficial to decide upon a threshold of agency as a deployment bar. If an AI system surpassed this level of agency, it could be forbidden from application in certain consequential sectors, like energy, the military, finance, health care, and criminal justice. The FATE community has previously rallied around deeming certain applications off-limits for particular technologies, such as the use of deep learning to predict criminality [71].

Efforts to improve democratic control and oversight over AI development could help to address the incentives for the development of increasingly agentic systems that we highlighted in Section 3.2.1. Tutt [184] proposes an "FDA for algorithms", which would scrutinize each algorithmic system before permitting its deployment, just as drugs are regulated in the United States. Huang and Siddarth [92], Jernite et al. [97] highlight the imbalance of power between AI developers and the rest of society, and propose frameworks and vehicles for collective data governance. Given the importance of data for training state-of-the-art systems [90], democratic control over data usage could be an important check on the development of increasingly agentic systems.

6 CONCLUSION

Our work focused on the increasing prevalence of agency in machine-learning systems and associated harms. We situated our characterization of increasing agency in the context of diverse work on the meaning of agency. We argued that there is a need to anticipate the harms from increasingly agentic systems, given a strong track record of, and incentives for, technical developments and increasing deployment. We described some anticipated harms from increasingly agentic systems, namely that they could cause systemic and delayed harms, disempower human decision-making, exacerbate extreme concentrations of power, and be a source of additional unknown threats through emergent capabilities.

Addressing the harms of increasingly agentic systems shares commonalities with central lines of work in the FATE community on anticipating the harms of algorithmic decision-making systems. Future work, such as investigations into the sociotechnical attributes of increasingly agentic systems and interventions upon the structural factors underlying their harms, readily follows from ongoing efforts. Immense pressure to develop and deploy emerging technologies should be met with similarly strong attempts to guide and constrain their impact.

ACKNOWLEDGMENTS

In no particular order, we would like to thank Robert Harling, Max Kaufmann, Fernando Diaz, Seth Lazar, Janarthanan Rajendran, Nicolas Le Roux, Shahar Avin, Kayla Matteucci, Thomas Gilbert, Herbie Bradley, and Usman Anwar for insightful comments about the direction of the project and parts of the paper. All mistakes remain our own.

AC acknowledges funding from Open Philanthropy and the Effective Ventures Foundation. AW acknowledges support from a Turing AI Fellowship under EPSRC grant EP/V025279/1, and the Leverhulme Trust via CFI. KC acknowledges funding from the Marshall Scholarship and Cambridge Trust.

REFERENCES

[1] 2022. *Defence Artificial Intelligence Strategy*. Technical Report. Ministry of Defence. https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy
[2] 2022. *Establishing a pro-innovation approach to regulating AI*. Technical Report. Office for Artificial Intelligence. https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement
[3] 2023. Auto-GPT: An Autonomous GPT-4 Experiment. https://github.com/Significant-Gravitas/Auto-GPT original-date: 2023-03-16T09:21:07Z.
[4] Mohamed Abdalla and Moustafa Abdalla. 2021. The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3461702.3462563
[5] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372871
[6] Rediet Abebe and Kira Goldner. 2018. Mechanism design for social good. *AI Matters* 4, 3 (Oct. 2018), 27–34. https://doi.org/10.1145/3284751.3284761
[7] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3461702.3462624
[8] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. 2022. Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=xNOVfCCvDpM
[9] Adept. 2022. ACT-1: Transformer for Actions. https://www.adept.ai/act
[10] M. Mehdi Afsar, Trafford Crump, and Behrouz Far. 2022. Reinforcement learning based recommender systems: A survey. https://doi.org/10.48550/arXiv.2101.06286 arXiv:2101.06286 [cs].
[11] Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2022. Using Large Language Models to Simulate Multiple Humans. *arXiv preprint arXiv:2208.10264* (2022).
[12] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. https://doi.org/10.48550/arXiv.2204.14198 arXiv:2204.14198 [cs].
[13] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–236. https://doi.org/10.1257/jep.31.2.211
[14] Jacob Andreas. 2022. Language Models as Agent Models. https://doi.org/10.48550/arXiv.2212.01681 arXiv:2212.01681 [cs].
[15] John Armour and Jeffrey N. Gordon. 2014. Systemic Harms and Shareholder Value. *Journal of Legal Analysis* 6, 1 (Oct. 2014), 35–85. https://doi.org/10.1093/jla/lau004
[16] Association for Computing Machinery (ACM). 2019. "Reinforcement Learning for Recommender Systems: A Case Study on Youtube," by Minmin Chen. https://www.youtube.com/watch?v=HEqQ2_1XRTs
[17] Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (Dec. 2022), 1067–1074. https://doi.org/10.1126/science.ade9097 Publisher: American Association for the Advancement of Science.
[18] Chelsea Barabas, Colin Doyle, J. B. Rubinovitz, and Karthik Dinakar. 2020. Studying up. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372859
[19] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 62–76. https://proceedings.mlr.press/v81/barabas18a.html
[20] Joslyn Barnhart. 2022. Emerging Technologies, Prestige Motivations and the Dynamics of International Competition. (Jan. 2022). https://www.governance.ai/research-paper/emerging-technologies-prestige-motivations-and-the-dynamics-of-international-competition
[21] Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. 2023. A Survey of Meta-Reinforcement Learning. https://doi.org/10.48550/arXiv.2301.08028 arXiv:2301.08028 [cs].
[22] G.A. Bekey. 2005. *Autonomous Robots: From Biological Inspiration to Implementation and Control*. MIT Press. https://books.google.ca/books?id=3xwfia2DpmoC
[23] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922 event-place: Virtual Event, Canada.
[24] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463
[25] Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. 2016. Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. https://papers.ssrn.com/abstract=2846909
[26] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The Forgotten Margins of AI Ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533157
[27] Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho. 2022. Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533204
[28] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association

for Computational Linguistics, Online, 1004–1015. https://doi.org/10.18653/v1/2021.acl-long.81
[29] William Boag, Harini Suresh, Bianca Lepe, and Catherine D'Ignazio. 2022. Tech Worker Organizing for Power and Accountability. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533111
[30] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258 arXiv:2108.07258 [cs].
[31] Samuel Bowman. 2022. The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 7484–7499. https://doi.org/10.18653/v1/2022.acl-long.516
[32] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. 2017. Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences* 114, 40 (2017), 10612–10617. Publisher: National Acad Sciences.
[33] Paula A. Braveman, Elaine Arkin, Dwayne Proctor, Tina Kauh, and Nicole Holm. 2022. Systemic And Structural Racism: Definitions, Examples, Health Damages, And Approaches To Dismantling. *Health Affairs* 41, 2 (Feb. 2022), 171–178. https://doi.org/10.1377/hlthaff.2021.01394 Publisher: Health Affairs.
[34] Meredith Broussard, Nicholas Diakopoulos, Andrea L Guzman, Rediet Abebe, Michel Dupagne, and Ching-Hua Chuan. 2019. Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly* 96, 3 (2019), 673–695. Publisher: SAGE Publications Sage CA: Los Angeles, CA.
[35] Noam Brown and Tuomas Sandholm. 2019. Superhuman AI for multiplayer poker. *Science* 365, 6456 (Aug. 2019), 885–890. https://doi.org/10.1126/science.aay2400 Publisher: American Association for the Advancement of Science.
[36] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
[37] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J. B. Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. https://doi.org/10.48550/arXiv.2004.07213 arXiv:2004.07213 [cs].
[38] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html
[39] Jenna Burrell and Marion Fourcade. 2021. The Society of Algorithms. *Annual Review of Sociology* 47, 1 (2021), 213–237. https://doi.org/10.1146/annurev-soc-090820-020800 _eprint: https://doi.org/10.1146/annurev-soc-090820-020800.
[40] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. 2023. Broken Neural Scaling Laws. https://doi.org/10.48550/arXiv.2210.14891 arXiv:2210.14891 [cs].
[41] Alexander Campolo and Kate Crawford. 2020. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society* 6 (Jan. 2020), 1–19. https://doi.org/10.17351/ests2020.277
[42] Micah D. Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. 2022. Estimating and Penalizing Induced Preference Shifts in Recommender Systems. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2686–2708. https://proceedings.mlr.press/v162/carroll22a.html ISSN: 2640-3498.
[43] Harrison Chase. 2022. LangChain 0.0.77 Docs. https://langchain.readthedocs.io/en/latest/modules/agents/getting_started.html
[44] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision Transformer: Reinforcement Learning via Sequence Modeling. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 15084–15097. https://proceedings.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html
[45] Tom Christiano and Sameer Bajaj. 2022. Democracy. In *The Stanford Encyclopedia of Philosophy* (spring 2022 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/democracy/
[46] Jack Clark and Dario Amodei. 2016. Faulty Reward Functions in the Wild. https://openai.com/blog/faulty-reward-functions/
[47] Katherine M. Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Joshua B. Tenenbaum. 2022. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. (May 2022). https://doi.org/10.48550/arXiv.2205.05718
[48] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533150
[49] Rémi Coulom. 2002. Reinforcement Learning Using Neural Networks, with Applications to Motor Control. (June 2002).
[50] Carla Zoe Cremer and Luke Kemp. 2021. Democratising Risk: In Search of a Methodology to Study Existential Risk. https://papers.ssrn.com/abstract=3995225
[51] Allan Dafoe. 2018. AI Governance: A Research Agenda. (Aug. 2018). https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf
[52] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. https://doi.org/10.48550/arXiv.2012.08630 arXiv:2012.08630 [cs].
[53] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. https://doi.org/10.48550/arXiv.2011.03395 arXiv:2011.03395 [cs, stat].
[54] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is not static. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372878
[55] Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The Benchmark Lottery. https://doi.org/10.48550/arXiv.2107.07002 arXiv:2107.07002 [cs].
[56] Daniel Clement Dennett. 1981. *The Intentional Stance*. MIT Press.
[57] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A Survey for In-context Learning. https://doi.org/10.48550/arXiv.2301.00234 arXiv:2301.00234 [cs].
[58] Hubert L Dreyfus. 1965. *Alchemy and artificial intelligence*. Technical Report. RAND CORP SANTA MONICA CA.
[59] Lilian Edwards. 2022. *The EU AI Act: a summary of its significance and scope*. Technical Report. https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf
[60] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The Algorithmic Imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency*.

ACM. https://doi.org/10.1145/3531146.3533186

[61] Kathleen M. Eisenhardt. 1989. Agency Theory: An Assessment and Review. *The Academy of Management Review* 14, 1 (1989), 57–74. https://doi.org/10.2307/258191 Publisher: Academy of Management.

[62] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 172–186. https://proceedings.mlr.press/v81/ekstrand18b.html

[63] Hadi Elzayn and Benjamin Fish. 2020. The effects of competition and regulation on error inequality in data-driven markets. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372842

[64] Mustafa Emirbayer and Ann Mische. 1998. What is Agency? *American journal of sociology* 103, 4 (1998), 962–1023. Publisher: The University of Chicago Press.

[65] Spotify Engineering. 2021. Shifting Consumption towards Diverse content via Reinforcement Learning. https://research.atspotify.com/2021/03/shifting-consumption-towards-diverse-content-via-reinforcement-learning/ Section: Algorithmic Responsibility.

[66] Charles Evans and Atoosa Kasirzadeh. 2022. User Tampering in Reinforcement Learning Recommender Systems. https://doi.org/10.48550/arXiv.2109.04083 arXiv:2109.04083 [cs].

[67] Richard Evans and Jim Gao. 2016. DeepMind AI Reduces Google Data Centre Cooling Bill by 40%. https://www.deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40

[68] Tom Everitt, Ryan Carey, Eric D. Langlois, Pedro A. Ortega, and Shane Legg. 2021. Agent Incentives: A Causal Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (May 2021), 11487–11495. https://doi.org/10.1609/aaai.v35i13.17368 Number: 13.

[69] Sebastian Farquhar, Ryan Carey, and Tom Everitt. 2022. Path-Specific Objectives for Safer Agent Incentives. https://doi.org/10.48550/arXiv.2204.10018 arXiv:2204.10018 [cs, stat].

[70] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533193

[71] Coalition for Critical Technology. 2020. Abolish the #TechToPrison-Pipeline. https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16

[72] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and Surprise in Large Generative Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533229

[73] Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling Laws for Reward Model Overoptimization. https://doi.org/10.48550/arXiv.2210.10760 arXiv:2210.10760 [cs, stat].

[74] Jason Gauci, Edoardo Conti, Yitao Liang, Kittipat Virochsiri, Yuchen He, Zachary Kaden, Vivek Narayanan, Xiaohui Ye, Zhengxing Chen, and Scott Fujimoto. 2019. Horizon: Facebook's Open Source Applied Reinforcement Learning Platform. https://doi.org/10.48550/arXiv.1811.00260 arXiv:1811.00260 [cs, stat].

[75] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Dec. 2021), 86–92. https://doi.org/10.1145/3458723

[76] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[77] Jenny Gesley, Tariq Ahmad, Edouardo Soares, Ruth Levush, Gustavo Guerra, James Martin, Kelly Buchanan, Laney Zhang, Sayuri Umeda, Astghik Grigoryan, Nicolas Boring, Elin Hofverberg, Clare Feikhert-Ahalt, Graciela Rodriguez-Ferrand, George Sadek, and Hanibal Goitom. 2019. Regulation of Artificial Intelligence in Selected Jurisdictions. *Copyright, Fair Use, Scholarly Communication, etc.* (Jan. 2019). https://digitalcommons.unl.edu/scholcom/177

[78] Charlie Giattino, Edouard Mathieu, Julia Broden, and Max Roser. 2022. Artificial Intelligence. *Our World in Data* (2022).

[79] Thomas Krendl Gilbert, Sarah Dean, Nathan Lambert, Tom Zick, and Aaron Snoswell. 2022. Reward Reports for Reinforcement Learning. https://doi.org/10.48550/arXiv.2204.10817 arXiv:2204.10817 [cs].

[80] Trystan S. Goetze. 2022. Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533106

[81] Charles Goodhart. 1975. Problems of monetary management: the UK experience in papers in monetary economics. *Monetary Economics* 1 (1975).

[82] Mary L. Gray and Siddharth Suri. 2019. *Ghost work: how to stop Silicon Valley from building a new global underclass.* Houghton Mifflin Harcourt, Boston.

[83] Ben Green. 2020. The false promise of risk assessments. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372869

[84] Ben Green and Yiling Chen. 2019. Disparate Interactions. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287563

[85] Nancy Green. 2021. An AI Ethics Course Highlighting Explicit Ethical Agents. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3461702.3462552

[86] Dylan Hadfield-Menell and Gillian K. Hadfield. 2019. Incomplete Contracting and AI Alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3306618.3314250

[87] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering Diverse Domains through World Models. https://doi.org/10.48550/arXiv.2301.04104 arXiv:2301.04104 [cs, stat].

[88] Stacy A. Hickox. 2010. Employer Liability of Negligent Hiring of Ex-Offenders. *Saint Louis University Law Journal* 55, 3 (2010), 1001–1046. https://heinonline.org/HOL/P?h=hein.journals/stlulj55&i=1029

[89] Jacob Hilton, Jie Tang, and John Schulman. 2023. Scaling laws for single-agent reinforcement learning. https://doi.org/10.48550/arXiv.2301.13442 arXiv:2301.13442 [cs, stat].

[90] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=iBBcRUlOAPR

[91] Yubo Hou, Dan Xiong, Tonglin Jiang, Lily Song, and Qi Wang. 2019. Social media addiction: Its impact, mediation, and intervention. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 13, 1 (Feb. 2019). https://doi.org/10.5817/CP2019-1-4 Number: 1.

[92] Saffron Huang and Divya Siddarth. 2023. Generative AI and the Digital Commons. https://cip.org/research/generative-ai-digital-commons

[93] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner Monologue: Embodied Reasoning through Planning with Language Models. https://doi.org/10.48550/arXiv.2207.05608 arXiv:2207.05608 [cs].

[94] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 1617–1626. https://proceedings.mlr.press/v70/jabbari17a.html ISSN: 2640-3498.

[95] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901 event-place: Virtual Event, Canada.

[96] Michael C. Jensen and William H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 4 (Oct. 1976), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

[97] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Gérard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Isaac Johnson, Dragomir Radev, Somaieh Nikpoor, Jörg Frohberg, Aaron Gokaslan, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2206–2222. https://doi.org/10.1145/3531146.3534637 arXiv:2206.03216 [cs].

[98] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* (Nov. 2022), 3571730. https://doi.org/10.1145/3571730 arXiv:2202.03629 [cs].

[99] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate Feedback Loops in Recommender Systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3306618.3314288

[100] Deborah G. Johnson and Mario Verdicchio. 2017. Reframing AI Discourse. *Minds and Machines* 27, 4 (Dec. 2017), 575–590. https://doi.org/10.1007/s11023-017-9417-6

[101] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc. https://papers.nips.cc/paper/2016/hash/eb163727917cbba1eea208541a643e74-Abstract.html

[102] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (Aug. 2021), 583–589. https://doi.org/10.1038/s41586-021-03819-2 Number: 7873 Publisher: Nature Publishing Group.

[103] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. https://doi.org/10.48550/arXiv.2001.08361 arXiv:2001.08361 [cs, stat].

[104] Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 576–586. https://doi.org/10.1145/3442188.3445919 event-place: Virtual Event, Canada.

[105] Joe Kava. 2014. Better data centers through machine learning. https://blog.google/inside-google/infrastructure/better-data-centers-through-machine/

[106] Betul Keles, Niall McCrae, and Annmarie Grealish. 2020. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth* 25, 1 (Dec. 2020), 79–93. https://doi.org/10.1080/02673843.2019.1590851 Publisher: Routledge _eprint: https://doi.org/10.1080/02673843.2019.1590851.

[107] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of Language Agents. *arXiv:2103.14659 [cs]* (March 2021). http://arxiv.org/abs/2103.14659 arXiv: 2103.14659.

[108] Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. 2022. Discovering Agents. https://doi.org/10.48550/arXiv.2208.08345 arXiv:2208.08345 [cs].

[109] Raffi Khatchadourian. 2015. The Doomsday Invention. *The New Yorker* (Nov. 2015). https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom

[110] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. 2020. Specification gaming: the flip side of AI ingenuity. *DeepMind Blog* (2020).

[111] David Krueger, Tegan Maharaj, and Jan Leike. 2020. Hidden Incentives for Auto-Induced Distributional Shift. https://doi.org/10.48550/arXiv.2009.09153 arXiv:2009.09153 [cs, stat].

[112] T.S. Kuhn and I. Hacking. 2012. *The Structure of Scientific Revolutions*. University of Chicago Press. https://books.google.co.uk/books?id=3eP5Y_OOuzwC

[113] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences* 40 (2017), e253. https://doi.org/10.1017/S0140525X16001837 Publisher: Cambridge University Press.

[114] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. 2022. Goal Misgeneralization in Deep Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12004–12019. https://proceedings.mlr.press/v162/langosco22a.html

[115] Seth Lazar. 2022. Legitimacy, Authority, and the Political Value of Explanations. https://doi.org/10.48550/arXiv.2208.08628 arXiv:2208.08628 [cs].

[116] Daniel Leufer. 2020. Why We Need to Bust Some Myths about AI. *Patterns* 1, 7 (Oct. 2020), 100124. https://doi.org/10.1016/j.patter.2020.100124

[117] Gideon Lewis-Kraus. 2022. How harmful is social media? https://www.newyorker.com/culture/annals-of-inquiry/we-know-less-about-social-media-than-we-think Publication Title: The New Yorker.

[118] Bo-hu Li, Bao-cun Hou, Wen-tao Yu, Xiao-bing Lu, and Chun-wei Yang. 2017. Applications of artificial intelligence in intelligent manufacturing: a review. *Frontiers of Information Technology & Electronic Engineering* 18, 1 (Jan. 2017), 86–96. https://doi.org/10.1631/FITEE.1601885

[119] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2022. Fairness in Recommendation: A Survey. https://doi.org/10.48550/arXiv.2205.13619 arXiv:2205.13619 [cs].

[120] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

[121] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158.

[122] Michele Loi and Matthias Spielkamp. 2021. Towards Accountability in the Use of Artificial Intelligence for Public Administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3461702.3462631

[123] Kate Manne. 2017. *Down Girl: The Logic of Misogyny*. Oxford University Press.

[124] maraoz. 2021. Interviewing Albert Einstein via GPT-3. https://maraoz.substack.com/embed

[125] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. https://doi.org/10.48550/arXiv.2203.11147 arXiv:2203.11147 [cs].

[126] Meta. 2023. Meta Reports Fourth Quarter and Full Year 2022 Results. https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-Fourth-Quarter-and-Full-Year-2022-Results/default.aspx

[127] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *AI & SOCIETY* 35, 4 (Dec. 2020), 957–967. https://doi.org/10.1007/s00146-020-00950-y

[128] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287596

[129] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. https://doi.org/10.48550/arXiv.1312.5602 arXiv:1312.5602 [cs].

[130] AI Myths. [n. d.]. Myth: AI has agency. https://www.aimyths.org/ai-has-agency

[131] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. https://doi.org/10.48550/arXiv.2112.09332 arXiv:2112.09332 [cs].

[132] Simone Natale and Andrea Ballatore. 2020. Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence* 26, 1 (2020), 3–18. Publisher: SAGE Publications Sage UK: London, England.

[133] Pandu Nayak. 2019. Understanding searches better than ever before. https://blog.google/products/search/search-language-understanding-bert/

[134] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and Engineering Ethics* 2, 1 (March 1996), 25–42. https://doi.org/10.1007/BF02639315

[135] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287593

[136] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context Learning and Induction Heads. *Transformer Circuits Thread* (2022).

[137] Stephen M Omohundro. 2008. The Basic AI Drives. In *AGI*, Vol. 171. 483–492.

[138] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/

[139] OpenAI. 2023. ChatGPT plugins. https://openai.com/blog/chatgpt-plugins

[140] Laurent Orseau, Simon McGregor McGill, and Shane Legg. 2018. Agents and Devices: A Relative Definition of Agency. https://doi.org/10.48550/arXiv.1805.12387 arXiv:1805.12387 [cs, stat].

[141] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=JYtwGwIL7ye

[142] Alexander Pan, Chan Jun Shern, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. https://doi.org/10.48550/arXiv.2304.03279 arXiv:2304.03279 [cs].

[143] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association

for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3526113.3545616

[144] Matteo Pasquinelli and Vladan Joler. 2021. The Nooscope manifested: AI as instrument of knowledge extractivism. *AI & SOCIETY* 36, 4 (Dec. 2021), 1263–1280. https://doi.org/10.1007/s00146-020-01097-6

[145] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. https://doi.org/10.48550/arXiv.2212.09251 arXiv:2212.09251 [cs].

[146] Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. 2022. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science* 378, 6623 (Dec. 2022), 990–996. https://doi.org/10.1126/science.add4679 Publisher: American Association for the Advancement of Science.

[147] Billy Perrigo. 2023. Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer. *Time* (Jan. 2023). https://time.com/6247678/openai-chatgpt-kenya-workers/

[148] Steven T. Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. https://doi.org/10.48550/arXiv.2208.02957 arXiv:2208.02957 [cs].

[149] By Pratyusha. 2020. World view. *Nature* 583 (2020), 169.

[150] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. https://doi.org/10.48550/arXiv.2112.11446 arXiv:2112.11446 [cs].

[151] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1 (Dec. 2021). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html

[152] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3306618.3314244

[153] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533158

[154] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372873

[155] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A Generalist Agent. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=1ikK0kHjvj

[156] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372879

[157] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 101–108.

[158] Stuart J. Russell and Peter Norvig. 2021. *Artificial Intelligence: A Modern Approach* (4 ed.).

[159] Markus Schlosser. 2019. Agency. In *The Stanford Encyclopedia of Philosophy* (winter 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/

[160] Ron Schmelzer. 2019. Amazon Dives Deep into Reinforcement Learning. *Forbes* (June 2019). https://www.forbes.com/sites/cognitiveworld/2019/06/14/amazon-dives-deep-into-reinforcement-learning/

[161] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* 588, 7839 (Dec. 2020), 604–609. https://doi.org/10.1038/s41586-020-03051-4 Number: 7839 Publisher: Nature Publishing Group.

[162] Anne Schulz, P Howard, and R Nielsen. 2019. Industry, Experts, or Industry Experts? Academic Sourcing in News Coverage of AI. (2019). Publisher: Reuters Institute for the Study of Journalism.

[163] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3287560.3287598

[164] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box". In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372827

[165] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. https://doi.org/10.48550/arXiv.2210.01790 arXiv:2210.01790 [cs].

[166] Yonadav Shavit. 2023. What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring. https://doi.org/10.48550/arXiv.2303.11341 arXiv:2303.11341 [cs].

[167] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2022. Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction. https://doi.org/10.48550/arXiv.2210.05791 arXiv:2210.05791 [cs].

[168] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (Jan. 2016), 484–489. https://doi.org/10.1038/nature16961 Number: 7587 Publisher: Nature Publishing Group.

[169] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. https://doi.org/10.48550/arXiv.1712.01815 arXiv:1712.01815 [cs].

[170] Joar Max Viktor Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and Characterizing Reward Hacking. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://arxiv.org/abs/2209.13085

[171] Elizabeth S Spelke and Katherine D Kinzler. 2007. Core knowledge. *Developmental science* 10, 1 (2007), 89–96. Publisher: Wiley Online Library.

[172] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem,

Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. https://doi.org/10.48550/arXiv.2206.04615 arXiv:2206.04615 [cs, stat].

[173] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533177

[174] steven t. piantadosi [@spiantado]. 2022. Yes, ChatGPT is amazing and impressive. No, @OpenAI has not come close to addressing the problem of bias. Filters appear to be bypassed with simple tricks, and superficially masked. And what is lurking inside is egregious. @Abebab @sama tw racism, sexism. https://t.co/V4fw1fY9dY. https://twitter.com/spiantado/status/1599462375887114240

[175] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. 2020. A Deep Learning Approach to Antibiotic Discovery. *Cell* 180, 4 (Feb. 2020), 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021

[176] Yulia W. Sullivan and Samuel Fosso Wamba. 2022. Moral Judgments in the Age of Artificial Intelligence. *Journal of Business Ethics* 178, 4 (July 2022), 917–943. https://doi.org/10.1007/s10551-022-05053-w

[177] Richard Sutton. 2022. https://twitter.com/richardssutton/status/1575619651563708418

[178] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[179] Richard S. Sutton, Michael Bowling, and Patrick M. Pilarski. 2022. The Alberta Plan for AI Research. https://doi.org/10.48550/arXiv.2208.11173 arXiv:2208.11173 [cs].

[180] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2021. Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3461702.3462602

[181] Adaptive Agent Team, Jakob Bauer, Kate Baumli, Satinder Baveja, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor, Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Hannah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez-Nieves, Nemanja Rakicevic, Tim Rocktäschel, Yannick Schroecker, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl, and Lei Zhang. 2023. Human-Timescale Adaptation in an Open-Ended Task Space. https://doi.org/10.48550/arXiv.2301.07608 arXiv:2301.07608 [cs].

[182] Philip E Tetlock and Dan Gardner. 2016. *Superforecasting: The Art and Science of Prediction*. Random House.

[183] Robert Trager. 2022. Deliberating Autonomous Weapons. *Issues in Science and Technology* XXXVIII, 4 (2022). https://issues.org/autonomous-weapons-russell-forum/

[184] Andrew Tutt. 2017. An FDA for Algorithms. *Administrative Law Review* 69, 1 (2017), 83–124. https://heinonline.org/HOL/P?h=hein.journals/admin69&i=95

[185] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). https://doi.org/10.48550/arXiv.2206.10498 arXiv:2206.10498 [cs].

[186] Agnes Schim van der Loeff, Iggy Bassi, Sachin Kapila, and Jevgenij Gamper. 2019. AI Ethics for Systemic Issues: A Structural Approach. https://doi.org/10.48550/arXiv.1911.03216 arXiv:1911.03216 [cs].

[187] Lee Vinsel. 2021. You're Doing It Wrong: Notes on Criticism and Technology Hype. https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5

[188] Heather Vogell, Haru Coryne, and Ryan Little. 2022. How a secret rent algorithm pushes rents higher. https://www.propublica.org/article/yieldstar-rent-increase-realpage-rent Publication Title: ProPublica.

[189] Axel Volkery and Teresa Ribeiro. 2009. Scenario planning in public policy: Understanding use, impacts and the role of institutional context factors. *Technological forecasting and social change* 76, 9 (2009), 1198–1207. Publisher: Elsevier.

[190] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2153–2162. https://doi.org/10.18653/v1/D19-1221

[191] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=yzkSU5zdwD

[192] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://doi.org/10.48550/arXiv.2201.11903 arXiv:2201.11903 [cs].

[193] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533088

[194] Sean Welsh. 2019. Regulating Lethal and Harmful Autonomy. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3306618.3314295

[195] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta. 2021. Recommender systems and the amplification of extremist content. *Internet Policy Review* 10, 2 (June 2021). https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content

[196] Maranke Wieringa. 2020. What to account for when accounting for algorithms. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372833

[197] Robert Wolfe and Aylin Caliskan. 2022. American == White in Multimodal Language-and-Image AI. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3514094.3534136

[198] S. Wyatt. 2008. Technological determinism is dead; Long live technological determinism. In *Handbook of Science and Technology Studies*, E. Hackett, O. Amsterdamska, M. Lynch, and J. Wajcman (Eds.). MIT Press, Cambridge, 165–180.

[199] Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. 2021. Mastering Atari Games with Limited Data. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 25476–25488. https://proceedings.neurips.cc/paper/2021/hash/d5eca8dc3820cad9fe56a3bafda65ca1-Abstract.html

[200] Sunkyung Yoon, Mary Kleinman, Jessica Mertz, and Michael Brannick. 2019. Is social network site usage related to depression? A meta-analysis of Facebook–depression relations. *Journal of Affective Disorders* 248 (April 2019), 65–72. https://doi.org/10.1016/j.jad.2019.01.026

[201] Meg Young, Michael Katell, and P. M. Krafft. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3533194

[202] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. https://doi.org/10.48550/arXiv.2204.00598 arXiv:2204.00598 [cs].

[203] Xueru Zhang and M. Liu. 2020. Fairness in Learning-Based Sequential Decision Algorithms: A Survey. *ArXiv* abs/2001.04861 (2020).

[204] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. 2021. The AI Economist: Optimal Economic Policy Design via Two-level Deep Reinforcement Learning. https://doi.org/10.48550/arXiv.2108.02755 arXiv:2108.02755 [cs, econ, q-fin].

[205] Miri Zilka, Holli Sargeant, and Adrian Weller. 2022. Transparency, Governance and Regulation of Algorithmic Tools Deployed in the Criminal Justice System: a UK Case Study. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. https://doi.org/10.1145/3514094.3534200

[206] Remco Zwetsloot and Allan Dafoe. 2019. Thinking about risks from AI: Accidents, misuse and structure. *Lawfare. February* 11 (2019), 2019.