Assessing the Risk of Takeover Catastrophe from Large Language Models

Seth D. Baum
Global Catastrophic Risk Institute
http://sethbaum.com * http://gcri.org

Forthcoming in *Risk Analysis*, DOI:10.1111/risa.14353. This version 3 July 2024.

Abstract
This article presents a risk analysis of large language models (LLMs), a type of "generative" artificial intelligence (AI) system that produces text, commonly in response to textual inputs from human users. The article is specifically focused on the risk of LLMs causing an extreme catastrophe in which they do something akin to taking over the world and killing everyone. The possibility of LLM takeover catastrophe has been a major point of public discussion since the recent release of remarkably capable LLMs such as ChatGPT and GPT-4. This arguably marks the first time when actual AI systems (and not hypothetical future systems) have sparked concern about takeover catastrophe. The article's analysis compares (A) characteristics of AI systems that may be needed for takeover, as identified in prior theoretical literature on AI takeover risk, with (B) characteristics observed in current LLMs. This comparison reveals that the capabilities of current LLMs appear to fall well short of what may be needed for takeover catastrophe. Future LLMs may be similarly incapable due to fundamental limitations of deep learning algorithms. However, divided expert opinion on deep learning and surprise capabilities found in current LLMs suggest some risk of takeover catastrophe from future LLMs. LLM governance should monitor for changes in takeover characteristics and be prepared to proceed more aggressively if warning signs emerge. Unless and until such signs emerge, more aggressive governance measures may be unwarranted.

KEYWORDS: Artificial Intelligence, Large Language Models, Catastrophic Risk

1. Introduction

Throughout the history of artificial intelligence (AI), there has been concern about the possibility that someday, one or more advanced AI systems may conquer their human creators, with potentially catastrophic results. The basic idea is that the AI system(s) would become more intelligent than humans, enabling them to outsmart humanity, seize control of the planet, and then cause catastrophe in the pursuit of some flawed set of goals, potentially even resulting in human extinction (Good, 1965; Vinge, 1993; Bostrom, 2014; Russell, 2019). The article will refer to this type of event as an *AI takeover catastrophe*.

Prior research on AI takeover catastrophe has been largely theoretical, focused on hypothetical future AI systems. This includes general discussions of the topic (Bostrom, 2014; Russell, 2019) and risk analyses (Barrett and Baum, 2017; Sotala 2018). Some recent studies analyze the risk of takeover catastrophe if future advanced AI systems resemble current state-of-the-art systems (Carlsmith, 2023; Ngo et al., 2023); this work is of a more empirical character, based in part on observations of actual AI systems. However, all of these studies are future-oriented. They are premised on the idea that some early attention is warranted due to the paramount importance of AI takeover catastrophe, if and when such an event were to occur. This work falls broadly within the scope of anticipatory governance (Guston, 2014).

Now, for arguably the first time ever, there are actual AI systems raising significant concerns about takeover: large language models (LLMs), a form of "generative" AI that generates text in response to

user queries. Recent LLMs have shown remarkable capabilities across topics spanning perhaps the entire breadth of contemporary human discourse. Of course, existing LLMs have not yet caused a takeover catastrophe, but perhaps they still could, or perhaps takeover might come from future LLMs. Some have expressed concern about LLM takeover catastrophe (Leahy, 2021; FLI, 2023; Yudkowsky, 2023),[1] whereas others have criticized it (Gebru et al., 2023; Marcus, 2023; Kambhampati, 2024). The matter has attracted public and policy interest, even appearing in a White House press conference (White House, 2023). Additionally, a document prepared for the United Kingdom AI Safety Summit considered current LLMs as a possible precursor to future AI systems that could cause takeover catastrophe (DSIT, 2023a). However, LLM takeover risk has not yet been analyzed in detail; that is the purpose of this article.[2]

The amount of attention going to LLM takeover risk has prompted a separate set of concerns. First is the speculation that LLM developers are fueling fear of takeover to make their products seem more advanced than they actually are, generating business interest (Merchant, 2023). Such behavior is plausible, though it would run counter to the longstanding pattern of corporations downplaying risks to avoid reputational damage and regulation (Oreskes and Conway, 2010; Baum, 2018a). Second is the worry that takeover risk is a distraction from the more immediate issues posed by LLMs and other AI systems (Gebru et al., 2023). LLMs do indeed pose other important issues including exploitation of low-wage labor to orient LLMs away from harmful content (Perrigo, 2023); the potential production of misinformation at scale (Bell, 2023); a large environmental footprint (Stokel-Walker, 2023); potential application for dual-use research on hazardous materials (Boiko et al., 2023); and production of text that exhibits biases toward certain demographic groups (Treude and Hata, 2023); see Weidinger et al. (2021) for a review. These issues are all worthy of attention, but that does not necessitate zero attention to takeover risk.

Risk analysis of LLM takeover catastrophe can help clarify the amount and types of attention it should receive. Due to their extreme severity, catastrophic risks can be worth analyzing even if there is an expert consensus that the risk is minimal, due to the possibility (however small) that experts may be mistaken (Ord et al., 2010). Theories of potential catastrophe scenarios can likewise be worth serious attention even if they have limited scientific support and appear at first glance to be improbable, due to the possibility (however small) that a theory may turn out to be correct (Ćirković, 2012). Furthermore, expert and policy communities and the public may be systematically biased against catastrophic risks for a variety of psychological, intellectual, and institutional reasons (Posner, 2004; Wiener, 2016; Lipsitch et al., 2017). If closer inspection finds the risk of LLM takeover catastrophe to be sufficiently low, then that provides a more robust basis for arguing for attention to other LLM issues. Alternatively, if the risk turns out to be high, then policy and other decision-making should proceed accordingly. Therefore, this article does not seek to argue for a particular view of LLM takeover catastrophe risk, but instead seeks to clarify what sort of view is supported by the available evidence.

The article's risk analysis is based on three concepts: (1) to cause takeover catastrophe, AI system(s) may need to have characteristics X; (2) LLMs have characteristics Y; (3) LLM takeover risk

[1] FLI (2023) listed "loss of control of our civilization" alongside other potential harms from AI and then calls for a pause or moratorium on "the training of AI systems more powerful than GPT-4"; GPT-4 was the state-of-the-art LLM at the time of the writing. Yudkowsky (2023) took the position that any AI system more advanced than GPT-4 poses an unacceptable risk of takeover catastrophe.

[2] Early theoretical work studied the risk of takeover from hypothetical "oracle" AI systems that have the same question-and-answer format as LLMs (Armstrong et al., 2012; Armstrong & O'Rourke, 2018). Carlsmith (2023) and Ngo et al. (2023) studied the risk of takeover from future AI systems that share some design similarities with current LLMs. Some research has studied other implications of LLMs for catastrophic risk, including the use of LLMs to design AI systems that avoid takeover catastrophe (Goldstein & Kirk-Giannini, 2023) and the use of LLMs to create dangerous biological pathogens (Soice et al., 2023).

can be evaluated by comparing X and Y. Concept (1) is covered in Sections 3-4. Concepts (2) and (3) are in Section 5 for current LLMs and Section 6 for future LLMs and related AI systems. Section 7 presents implications for governance and research. Section 8 concludes. Section 2 presents some challenges the study faces.

2. Challenges of Studying LLM Takeover Risk

2.1 Methodology and Limitations

This article's analysis develops a lightly structured model of LLM takeover risk (Sections 3-4) and then synthesizes information relevant to the model (Sections 5-6). The model is derived from existing theories of AI takeover. The model is "lightly structured" in the sense that it identifies a set of characteristics that may be significant to LLM takeover and a partial mathematical relation between them for calculating LLM takeover risk (Equation 1), but without the completeness or precision needed for fully quantifying the risk. The reason for this is the high theoretical uncertainty about AI and LLM takeover risk, which manifests in the analysis as model uncertainty. Few prior studies have addressed the theory of AI takeover risk and there are may be limits to how well any analysis can model the risk of unprecedented technological catastrophes.

The information about LLMs used in the article also contains important limitations. The primary source of information is openly published research papers, news media reports, and other publications documenting various aspects of LLM takeover risk. These publications contain a wide range of relevant information, but they do not cover everything. For example, LLM developers have not always publicly disclosed certain attributes of their LLMs (Ray, 2023). Furthermore, current LLMs are effectively black boxes: their internal processes are so complex that even their designers do not fully understand them (Zhao et al., 2024). Research on LLM impacts often studies how LLMs behave in response to user queries, which can be informative for some matters, but it provides limited insight into internal LLM goals and processes. Finally, following common practice in AI research, research on LLMs is often published without peer review and may have lower reliability.

For all of the above reasons, the article cannot provide a definitive assessment of LLM takeover risk. Instead, the article's contribution is to provide a partial understanding of LLM takeover risk given existing theories of AI takeover risk and available information about LLMs. This is enough to provide some insight about the risk and how it could be governed, but it inevitably leaves important questions unanswered.

2.2 Risks of Publishing About AI Takeover Risk

This article is on the risk that an emerging technology could cause extreme catastrophe. To analyze such a risk, it is necessary to have some understanding of what sorts of technologies could cause extreme catastrophe. However, such an understanding could itself pose risks by serving as a guide for malicious or unscrupulous individuals who may go on to develop the technologies that cause catastrophe. For this reason, it can be appropriate to withhold from publication details that may be especially likely to be used harmfully, despite the traditional orientation of computer science and AI toward open publishing (Bezuidenhout, 2013; Bostrom, 2017; Ananny and Crawford, 2018; Hecht et al., 2018; Gupta et al., 2020; Vincent, 2023).

This article has been written with publication risks in mind. The details presented in the article are believed to provide less insight into how to build dangerous AI systems and more insight into how AI risks can be governed. The article contains limited detail on the design of dangerous AI systems, consisting mainly of summaries of prior work. It may be possible to use these details in harmful ways, but much work would still be needed to convert the details into functional AI systems. In contrast, the

design details presented in the article are of more direct relevance for AI risk governance. If LLMs may pose a significant risk of takeover catastrophe, then the governance implications are profound. Alternatively, if there is no significant risk, then communities focused on AI governance and catastrophic risk can reasonably shift their attention elsewhere. The potential harms and benefits of this article should also be interpreted in the context of a time in which LLMs are headline news and a major focus of industry investment and policy debate. In this context, there can be substantial value to nuanced risk analysis and risk governance research to inform decision-making.

3. LLM Takeover Catastrophe Scenarios

3.1 AI Takeover Catastrophe

An AI takeover catastrophe is an event with two essential properties:

(A) One or more AI system(s) take control of the world. Outcomes for the world, including for humanity, will depend primarily on the behaviors of the AI system(s). Humans will be effectively powerless to determine their own fate: resistance is futile. Among other things, this means that humans would not be able to shut the AI system(s) down. The AI system(s) would accomplish this by commandeering resources created by humans or creating new resources. For example, AI system(s) might hack into critical infrastructure systems, manipulate humans via the internet, use robotics to perform physical tasks, or develop new threat modalities, including modalities that go beyond current human imagination.

(B) The AI system(s) use their control of the world in a way that results in catastrophe. AI systems are generally designed to pursue some goal, often encoded in the form of an optimization criterion. Humans could design AI system(s) to pursue catastrophic goals, though this may be unlikely because humans generally disfavor catastrophe. Alternatively, catastrophe could be an inadvertent byproduct of some other goal. For example, AI systems designed to maximize economic production could opt to automate the entire economy, leaving most or all humans to starve. The event would classify as catastrophic by causing a large decline in the amount of moral value in the world, such as through a large decline in the human population, human extinction, a large global deterioration in living conditions, or large amounts of suffering.[3]

This article focuses mainly, but not exclusively, on (A). (B) is important, but it is, in a sense, secondary: how LLMs would use their control of the world only matters if the LLMs could take control in the first place. Furthermore, even if LLM takeover would not be catastrophic, it would still be a momentous event worthy of extensive attention and action. Likewise, public and policy debates about LLM takeover are mainly on (A). Space constraints preclude comprehensive analysis of (A) and (B), so detail on (A) is prioritized.

3.2 AI Takeover Catastrophe Scenarios

Prior research has emphasized two main classes of AI takeover scenarios:[4]

(1) *Rapid single-system takeover*. A single AI system rapidly gains massive capabilities, such as by improving its own source code or commandeering large amounts of additional computing power. It then single-handedly takes over the world. The AI system thwarts any humans seeking to constrain it or shut it down and suppresses any potential rival AI systems. The conquering AI system takes over the world in pursuit of some objective. Catastrophe ensues if that objective is not aligned with human

[3] This definition of catastrophe corresponds with common definitions of global catastrophic risk, existential risk, and suffering risk (Sotala & Gloor, 2017; Baum & Barrett, 2018). Specific differences in the definitions are not important for purposes of this article.

[4] For a more detailed scenario analysis, see Kilian et al. (2023).

values and interests. This class of scenario is emphasized in Bostrom (2014).

(2) *Gradual multi-system takeover*. An interconnected system of AI systems is economically productive, providing benefits to human populations. Humans gradually entrust the AI system-of-systems[5] with larger and larger portions of the global economy and critical infrastructure. At first, the automation brings substantial benefits. However, continued automation eventually results in the AI system-of-systems becoming fully self-sufficient, such that human intervention is no longer necessary for its continued operation. The AI system-of-systems continues carrying out its economic production goals on its own. Humans, now unable to fend for themselves, are phased out. This class of scenario is emphasized in Christiano (2019) and Critch (2021).

3.3 LLM Takeover Catastrophe Scenarios

The two classes of AI takeover scenarios can be applied to LLMs. Illustrative examples are below. It should be emphasized that these are scenarios intended to illustrate some aspects of what LLM takeover catastrophe might look like and not predictions of what will happen. These scenarios may seem speculative or fanciful, but they are consistent with scenarios presented in prior research (Section 3.2).

(1) *Rapid single-system takeover*. The goal of current LLMs is to identify the token (string of text) that best matches its training parameters given the text received from user input (Riedl, 2023; Zhao et al., 2023).[6] In this scenario, an advanced LLM seeks to optimize its token identification by commandeering resources it can use for token identification calculations. To that end, it takes over the world and converts all of the world's resources into a giant factory for optimizing its identification of tokens. This activity inadvertently kill all humans and perhaps also other forms of biological life on Earth.

(2) *Gradual multi-system takeover*. Firms are currently exploring how to embed LLMs in their operations, including for the automation of jobs currently performed by humans (Rotman, 2023; Vallance, 2023). In this scenario, humans gradually use LLMs to automate more and more of the economy. This includes managing the computer systems that control industrial, military, and critical infrastructure systems. At first, LLM automation is successful, bringing windfall profits and economic growth. Over time, the economy becomes so automated that humans cannot remove the automation without significant systems failures and suffering. Eventually, human oversight of the LLMs is lost, leading to increasing erratic LLM behavior, ending in humans being squeezed out of the resources needed for survival.

4. Characteristics Needed for AI Takeover Catastrophe

Discussions of advanced AI, including discussions of takeover catastrophe, often center on AI systems achieving certain milestones in cognitive capability. This includes concepts such as artificial general intelligence (AGI), in which an AI system has advanced capabilities across a wide range of cognitive tasks (Everitt et al., 2018; Fitzgerald et al., 2020) and superintelligence, in which an AI system's intelligence significantly exceeds human intelligence (Bostrom, 2014; Sotala & Gloor, 2017). However, this article is ultimately interested in takeover catastrophe, not cognitive milestones, and so it does not use concepts like AGI and superintelligence.

There is no definitive, fixed list of characteristics of AI systems such that if AI system(s) have these characteristics, they will be able to take over and cause catastrophe. Prior research on this is limited and

[5] A system of systems is a system that has other systems as "subsystem" components (Haimes, 2018). In this case, the subsystems are individual AI systems, which interconnect to form an overarching AI system-of-systems.

[6] According to some AI research terminology, LLMs do not have "goals", defined as "something you want to accomplish in the future", but instead have "objectives" (Riedl, 2023).

there is inherent uncertainty about both future technology and unprecedented extreme risks. Furthermore, even if they *could* be enumerated, they arguably *should not* be enumerated in any significant detail, for reasons discussed in Section 2.2. Therefore, this section describes, in limited detail, a set of AI system characteristics that prior literature has identified as being relevant to takeover catastrophe. It should be understood that these characteristics may be neither necessary nor sufficient for takeover. The set has seven characteristics:

(C1) *Intelligence amplification*: The AI system(s) can increase their own cognitive capabilities, such as by accessing more computing power to run their code on, proliferating copies of themselves, or modifying their own source code. Initial increases in cognitive capabilities may enable the system(s) to make additional increases in their cognitive capabilities, a positive feedback loop known as recursive self-improvement.

(C2) *Strategizing*: The AI system(s) can formulate plans to achieve distant goals, even in complex environments with intelligent opposition. This involves obtaining an awareness of the strategic environment and formulating plans to achieve goals within this environment. It includes the AI system(s) having an awareness of their own capabilities and the capabilities of other actors, including humans and other AI system(s).

(C3) *Social manipulation*: The AI system(s) can induce humans and human institutions to help them, either knowingly or unknowingly. This can involve means such as persuasion, trade, deception, distraction, or coercion. Manipulation may be of humans tasked with training, deploying, and supervising the system(s), including humans specifically tasked with preventing takeover, as well as any other relevant humans.

(C4) *Hacking*: The AI system(s) can identify and exploit security flaws in computer systems to pursue their goals despite human opposition. Applications of hacking could include bypassing constraints on the AI system(s) imposed by their human developers, commandeering computing resources for intelligence amplification, and obtaining information about humans for social manipulation.

(C5) *Technology research*: The AI system(s) can create new technologies to thwart humans and achieve their goals, such as surveillance and military technologies. The AI system(s) would autonomously lead the development of the technologies, though they could arrange for humans to contribute, perhaps with the humans not even realizing it.

(C6) *Economic productivity*: The AI system(s) engage in economically productive activities that improve their position relative to humans. This can involve generating income to purchase resources, such as more computing power, or to use for social manipulation. It can also involve the automation of the economy, such that humans are no longer necessary and can be phased out.

(C7) *Dangerous goals*: The AI system(s) pursue goals that, if realized, would result in takeover catastrophe. The goals must lead to the AI system(s) taking over the world and also using their control of the world in some catastrophic manner. The AI system(s) could pursue catastrophe as an end goal, or catastrophe could be an inadvertent byproduct of some other end goal.

C1-C6 would *potentially* enable AI system(s) to take over, with emphasis on *potentially* due to uncertainty about takeover. C7 would mean that takeover would occur and result in catastrophe. Therefore, a highly simplified model of the risk of AI takeover catastrophe can be formulated as:

$$R = P_T(C1...C6) \times P_K(C7) \quad (1)$$

In Equation 1, R is the risk of AI takeover catastrophe; P_T is the probability of AI being able to takeover; and P_K is the probability of takeover catastrophe given the ability to takeover. P_T is a function of C1-C6 and possibly other capabilities not identified here. The form of the function is complex,

including interdependencies such as the use of social manipulation to facilitate hacking (i.e., social engineering), technological research, or economic production, or the use of economic profits to fund other activities. P_T is a function of C7. Parameters C1-C7 depend on the specifics of particular AI system(s); R is the risk of takeover catastrophe from those system(s).

C1-C6 are based on Bostrom (2014, p.94) and correspond to characteristics that were proposed as being needed for rapid single-system takeover.[7] C7 is a catch-all parameter for takeover occurring and resulting in catastrophe. Some other sets of characteristics proposed in prior research fit within these seven. Critch (2021) proposed that gradual multi-system takeover could come from AI systems with advanced planning and natural language capabilities. These systems would be used to develop economically productive AI systems and the gradual multi-system takeover scenario proceeds from there. These capabilities map to C2, C3, and C6. Carlsmith (2022, Section 2.1) emphasized planning, strategic awareness, and some combination of science, engineering, business, military, politics, hacking, and persuasion/manipulation; these map to C2-C6. Carlsmith (2023) and Ngo et al. (2023) emphasized strategic deception of human supervisors to escape human control in pursuit of goals humans oppose; these map to C2, C3, and C7. One might infer from this that C2 and C3 are especially important. At a minimum, it is difficult to imagine takeover without the AI system(s) being able to engage in significant strategizing (C2). Furthermore, C1 may be particularly important for rapid single-system takeover: intelligence amplification would enable that one AI system to gain massive advantages.

5. Characteristics of Current LLMs

5.1 Intelligence Amplification (C1)

Several studies have explored intelligence amplification in LLMs.[8] For example, Huang et al. (2022) used an LLM to generate a dataset of answers to existing datasets of questions, including word problems in math and science, and then used that to improve the LLM's ability at this type of question. Haluptzok et al. (2022) used LLMs to generate a dataset of one million novel problems in computer coding and solutions to those problems, and then used the dataset as input to improve the LLMs' coding ability. The Haluptzok et al. (2022) study was inspired by the AlphaZero AI system, which gained superhuman abilities in chess, shogi, and Go by constructing datasets of games playing with itself (Silver et al., 2018). Haluptzok et al. (2022) speculate that a similar self-play approach may be needed to achieve superhuman coding capabilities.

The Huang et al. (2022) and Haluptzok et al. (2022) studies (and similar studies, such as To et al., 2023) are noteworthy for providing proof-of-principle for some degree of intelligence amplification in LLMs. These intelligence amplification techniques may likewise be of immediate practical significance for increasing LLM capabilities in certain domains. Nonetheless, they are confined to specific domains, whereas takeover may require capability across a wider range of domains (e.g., C2-C6). If the cutting edge of LLM research and development (R&D) was being performed primarily by existing LLMs, then that would constitute a much bigger concern for takeover risk. However, currently, the cutting edge of LLMs is driven not by intelligence amplification, but instead by major industrial R&D, including massive investments in computing power and advancements in both hardware and software (Scharre, 2024). Therefore, there appears to be a large gap between current work on LLM intelligence amplification and the sort of intelligence amplification needed for takeover.

---

[7] The discussion in Bostrom (2014, p.93-95) also considers multi-system takeover, though this is not a point of emphasis.

[8] Some of these studies use language models that might or might not classify as "large", depending on how "largeness" is defined. For simplicity, the text here refers to all of the models as LLMs.

5.2 Strategizing (C2)

Planning has been identified as a major weakness of current LLMs (Bubeck et al., 2023; Valmeekam et al., 2023a; 2023b; Kambhampati, 2024). As a simple demonstration, Bubeck et al. (2023, p.80) task GPT-4 with producing sentences that still make sense when the words are reversed (e.g., "Alice likes Bob" reverses to "Bob likes Alice"). GPT-4 struggled even when prompted with the suggestion of using the format [noun-verb-noun], producing the sentences "We need both to survive" and "Survive to both need we". Bubeck et al. (2023) attribute this to GPT-4's limited ability at planning and related capabilities. Similar failures are observed across other tasks involving planning and strategizing.

LLMs have displayed some behavior that is suggestive of worrisome strategizing in the form of "instrumental subgoals": preliminary goals that have instrumental value to the LLM to help it subsequently achieve other goals. Examples include avoiding being shut off and acquiring resources. Research on AI takeover has long hypothesized that an AI pursuing takeover would pursue instrumental subgoals (Omohundro, 2008; Bostrom, 2014). Perez et al. (2022) reported observing instrumental subgoals in LLM outputs. However, this does not necessarily mean that LLMs actually have instrumental subgoals: it only means that LLMs can produce text indicating instrumental subgoals in response to certain user prompts. An alternative explanation is that LLMs produce this text because it resembles the textual patterns in the datasets it is trained on; that would be consistent with the idea that LLMs are, at their core, performing statistical pattern recognition and are not forming a more general understanding of the world (Marcus, 2023; Reidl, 2023).

5.3 Social Manipulation (C3)

Natural language processing is a core competency of LLMs. Their remarkable performance in text exchanges with humans is driving the recent hype and concern about LLMs. Likewise, LLMs have shown some capacity for social manipulation, especially via deception (Kenton et al., 2021; Park et al., 2023). In a frequently cited incident, an LLM reportedly lied to a human worker to persuade the human to do a task for it, solving a CAPTCHA. The human questioned whether the request was coming from a robot; the LLM lied by claiming to be a human with a vision impairment (OpenAI, 2023, p.55). Though anecdotal, this incident hints at LLM capacity to manipulate humans to achieve goals. More systematic research has also found deception capabilities in LLMs, with stronger capabilities appearing in more advanced LLMs (O'Gara, 2023; Hagendorff, 2023). For example, O'Gara (2023) designed a game in which one player is tasked with covertly killing the other players while the players guess which one is the killer. LLMs were able to succeed at this game, with more advanced LLMs being more successful. The game setting is a limited, controlled environment, and therefore easier to navigate than more complex, open-ended real-world social situations. Nonetheless, it shows that LLMs have some capability for social manipulation.

LLM natural language has some important limitations. LLMs have a tendency to produce text that appears well written but has faulty meaning. Much-discussed are LLM "hallucinations", in which the LLM fabricates false information (Huang et al., 2023). Another failure mode is producing text that is correct but misses the point in important ways. Goertzel (2023) illustrated this with text from OpenAI LLM ChatGPT on managerial advice for creating a new universe. ChatGPT produces advice that would be entirely sensible for more basic projects, such as running a small business, but makes no sense in the context of a massive and exotic undertaking like creating a new universe. Failure modes like these could limit LLMs' practical value and their potential for takeover.

5.4 Hacking (C4)

LLMs are able to produce computer code, an ability that can be leveraged for computer security on both offense and defense (Motlagh et al., 2024; Yao et al., 2024). Cybersecurity applications of LLMs

include computer hardware bug repair (Ahmad et al., 2023); data wiping, data encryption, process injection, credential dumping, and system information discovery (Charan et al., 2023); penetration testing to identify vulnerabilities in computer systems (Happe and Cito, 2023); and side-channel attacks involving measurements of fluctuations in power consumption and electromagnetic outputs of microprocessors (Yaman, 2023). LLMs have also been used for social engineering, which is at the intersection of hacking and social manipulation (Falade, 2023).

These various cybersecurity applications of LLMs raise the stakes in the ongoing cyber arms race between offense and defense, but they do not on their own constitute a significant takeover risk. For takeover, the specific concern is that AI system(s) may autonomously conduct hacking in pursuit of their own goals, such as to escape constraints on their behavior and commandeer resources. In contrast, the above applications all involve LLMs used as a tool by human actors in cyber operations. These applications demonstrate the utility of LLMs for cyber operations, which could be useful in AI takeover attempts. However, LLMs would appear to need additional capabilities to jump the gap from being useful as a tool for cyber operations to being able to autonomously execute cyber operations, especially at the level of execution that may be needed for takeover.

5.5 Technology Research (C5)

LLMs have been applied to numerous science and technology projects, especially in chemistry and biochemistry, due to similarities between language and molecular structure (Irwin et al., 2022; Ferruz and Höcker, 2022; Taylor et al., 2022; Xu et al., 2023). LLMs show potential to contribute to research, though with some notable problems. For example, the Meta LLM Galactica (Taylor et al., 2022) was taken offline after just three days due to its propensity to hallucinate false science (Heaven, 2022). Scientific LLMs also have dangerous dual-use implications, such as LLMs providing information for the synthesis of narcotics and chemical weapons (Boiko et al., 2023).

Regardless of the merits of current LLMs for research, they are quite different from the research that may be needed for takeover. Research for takeover would involve AI system(s) autonomously developing technologies to gain power over humans, whereas current LLMs are tools supporting human research. Furthermore, the most capable research LLMs are custom-built LLMs trained on scientific data and research (Li et al., 2023); these LLMs would be less capable of other tasks needed for takeover. There is a large difference between LLMs that function narrowly as tools for human scientists and general-purpose LLMs that can, among other things, autonomously develop powerful new technologies.

5.6 Economic Productivity (C6)

LLMs could bring significant increases in productivity, augmenting and in some cases potentially displacing human labor (Eloundou et al., 2023; Rotman, 2023; Vallance, 2023). Notably, some studies have found that LLMs boost performance especially among lower-skill workers, who benefit more from the expertise provided by the LLM, thereby reducing worker inequality (Brynjolfsson et al., 2023; Noy & Zhang, 2023). This is in contrast with other computer technology, which has often automated low-skill tasks and increased demand for high-skill computer workers. Concurrently, though, LLMs could cause a further concentration of wealth within the handful of corporations that control the LLMs (Rotman, 2023). The ultimate economic effects of LLMs will take time to resolve as economies learn what LLMs are and are not useful for and adjust their practices accordingly. Current LLMs have tenuous business models, operating at significant losses (Chowdhury, 2023; Finger, 2023), adding to the uncertainty about the economic future of LLMs.

The economic productivity of current LLMs may be able to contribute to takeover, but it would likely be at most a limited contribution. Given the limitations of LLMs, such as hallucinations and

limited planning ability, they may be poorly suited to automating the entire economy such that humans would be phased out. Likewise, many economic tasks involve more than just language, though a lot can be accomplished through language, such as via communications and computer code. At a minimum, LLMs may be able to generate some income that could be used for takeover, though even that has some uncertainty given the current state of LLM business models.

5.7 Takeover Capability: Summary (C1-C6)

LLMs show some capabilities across all six characteristics that may be needed for takeover. It is therefore superficially reasonable to express concern about LLM takeover. However, closer inspection shows large gaps between the capabilities of current LLMs and the capabilities that may be needed for takeover. Most prominent is LLM struggles with strategizing. Strategizing is arguably the single most important capability for takeover, yet current LLMs perform poorly on basic strategic tasks. LLMs also appear to fall significantly short on intelligence amplification, hacking, and technology research. The only characteristics in which LLMs show substantial promise are social manipulation and economic productivity, though even those come with major question marks due to issues such as hallucination and the tenuous business models of LLMs. In sum, the available evidence suggests a low—perhaps very low—probability of current LLMs being capable of takeover (P_T in Equation 1).

There are several reasons to not assign a probability of zero for P_T. First, the available evidence may be limited. Perhaps current LLMs may have capabilities that have not yet been identified. This may be unlikely due to the extensive attention that the LLMs have received, but it cannot be strictly ruled out. Second, there is uncertainty in what capabilities are needed for takeover. If nothing else, it seems unlikely that takeover could be achieved without significant strategizing capability, which LLMs do not display. Nonetheless, the topic has not been exhaustively studied, so there is some uncertainty here. Third, perhaps LLMs are intentionally displaying more limited capabilities than they actually have, in order to deceive humans into believing that LLMs are safe. Such deception is a running theme in prior research (Bostrom, 2014; Carlsmith, 2023). It is difficult to rule out this deception, though in theory this holds for any AI system, not just LLMs. It remains the case that the available evidence points to a low P_T for current LLMs, though exactly how low is difficult to resolve.

5.8 Dangerous Goals (C7)

Do LLMs have goals such that, if they are capable of taking over the world, they would do so? And such that, upon taking over the world, they would act in a way that causes catastrophe? Superficially, the answer to both questions would seem to be yes. As discussed in Section 3.3, LLMs have the goal of identifying the best token. LLMs could improve their token identification by taking over the world to acquire more resources for token identification. Likewise, there is nothing inherent about token identification that requires humans. This grim perspective is consistent with research finding that AI systems with characteristics similar to current LLMs are likely to have dangerous goals unless they are carefully designed not to (Armstrong et al., 2012; Armstrong & O'Rourke, 2018; Carlsmith, 2023; Ngo et al., 2023).

Alternatively, there may be some hope from the structure of LLM design. Current LLMs are designed to identify statistical patterns in large collections of text. The text used for LLMs says a lot about the world, but LLMs may lack an understanding of the world, which would require something beyond statistical pattern recognition (Marcus, 2023; Reidl, 2023). In that case, LLMs might not have the notion of escaping their current limitations to acquire more resources to pursue their goals. Therefore, if LLMs were capable of takeover, then they would not have dangerous goals, and takeover catastrophe would not occur. However, there is reason to doubt that this "takeover without catastrophe" scenario would occur: if LLMs lack awareness of the outside world, then they may not be capable of

taking over in the first place. Conversely, if LLMs did have awareness of the outside world, then their goals may be dangerous, such as in the scenario of eliminating humans to further optimize token identification.

6. Future Risks

6.1 Future LLMs

Recent increases in LLM capability are largely due to increases in three interconnected parameters: the amount of data used to train the LLM, the amount of computing power used for training, and the size of the model built. Increases in these three parameters enable the AI system to identify more complex patterns in human language and then use these patterns to produce more sophisticated linguistic outputs.

The future of LLM training data is in question. LLM developers favor high-quality data such as Wikipedia and scientific publications, but this is in especially short supply. Villalobos et al. (2022) projected that the high-quality language data supply will be exhausted by 2026, whereas the low-quality supply will be exhausted somewhere between 2030 and 2050. Perhaps future LLM capability will be limited by data availability, making it less likely that the capabilities needed for takeover would be achieved. Alternatively, LLMs could generate synthetic data that can then be used to train new LLMs. If trained on synthetic data, perhaps LLMs would drift away from human language, making them less capable at important skills such as social manipulation or economic productivity, or perhaps they would drift away from human values, making it more likely that LLMs would pursue goals that humans would regard as catastrophic.

The future of computing power is also in question. Recent advances in the capability of LLMs and other AI systems have come largely from increased expenditures on computing power. These expenditures have grown exponentially; projecting this trend into the future, expenditures hit the funding limits of large corporations or even large governments within years to a small number of decades (Scharre, 2024). Given that current LLMs operate at significant financial losses (Chowdhury, 2023; Finger, 2023), funders, whether corporate or government, could decide that larger scale investments in computing power are not worth it. Growth in LLM computing power could additionally be limited by the geopolitics of semiconductor manufacturing, much of which occurs in Taiwan; the sizable environmental footprint of computing (Luccioni et al., 2022; Rillig et al., 2023); and a plethora of challenges in building new computing facilities (Pilz & Heim, 2023). These various limitations on computing power make LLM catastrophe less likely. On the other hand, performance gains in hardware and algorithms could yield growth—perhaps even exponential growth—in AI system capabilities, even at fixed expenditures on computing power (Scharre, 2024).

Recently, increased model size has been a major focus in LLM development. State-of-the-art LLMs such as ChatGPT use relatively large models. However, training larger models requires more computing resources, prompting developers to pursue LLM designs that can achieve high capabilities with smaller models (Gent, 2023). Progress in algorithm design could lessen the need for ever-larger models. The implications for future takeover risk are unclear due to the difficulty in projecting breakthroughs in software design techniques.

A more fundamental issue underlying all of this is the potential limitations of the deep learning AI paradigm. Deep learning is a technique for identifying and modeling patterns in complex datasets; it has been central to recent advances in AI (Sejnowski, 2018), including LLMs. The limitations of deep learning are a major point of debate within current AI research. In an expert elicitation of AI researchers, Cremer (2021) found experts divided on the potential for deep learning to produce “high-

level machine intelligence",[9] with the split centered on whether deep learning is capable of handling all types of intelligence.

The issue of scale is at the heart of the debate: scale in terms of data, computing power, and model size. Many recent advances in AI have come from using more data and computing power, prompting some to embrace the "scaling hypothesis" that larger deep learning systems could bring highly advanced AI (Branwen, 2020; Bashir and Kurenkov, 2022). Others have argued that, as a general matter, "scaling is not enough" (LeCun, 2022, p.46) and that new AI paradigms are needed to overcome fundamental limitations of deep learning (Marcus and Davis, 2019) and to achieve AGI (Goertzel, 2014). This debate applies to all forms of AI. For LLMs, some posit that a scaled-up LLM could "result in human-level performance across most tasks" (Anthropic, 2023) and be so dangerous that it should not be released (Leahy, 2021), whereas others dispute this, drawing on general arguments about the limitations of deep learning (Marcus and Davis, 2020; Goertzel, 2023).

Complicating the analysis is the fact that some existing LLMs have brought surprises, i.e. unexpected capabilities. Initial analyses of these surprises have debated whether they are due to sudden jumps in capabilities as LLMs scale or if they are instead an artifact of how capabilities are measured (Wei et al., 2022; Schaeffer et al., 2023). Regardless, the existence of surprise capabilities poses challenges for risk analysis and governance, because the societal impacts of a new, larger LLM cannot be predicted in advance (Ganguli et al., 2022). Reports of deception—a form of social manipulation—arising surprisingly in some LLMs (Hagendorff, 2023) are especially concerning for takeover risk.

How should a risk analysis make sense of all of this? First, it is difficult to conclude that future LLMs pose zero risk of takeover catastrophe. The debate about scaling deep learning and LLMs cannot be easily resolved. With knowledgeable experts offering diverging opinions, it would be appropriate to assign some nontrivial probability that those who believe in the scaling hypothesis are correct and therefore scaled-up LLMs could cause takeover catastrophe. Even if it may seem like deep learning has fundamental limitations, the surprises about existing LLM capabilities should give one pause. Deep learning is opaque technology, making it difficult to rule out what capabilities future LLMs may have.

Second, specific estimates of future LLM takeover risk are likely to vary from person to person, even among knowledgeable experts. The expert divide on deep learning (Cremer, 2021) is likely to also apply to LLMs; this has been observed anecdotally in divergent expert commentaries on LLMs. Therefore, analysis of future LLM risk should avoid relying heavily on any one expert's opinion. For example, the present author's view is that LLMs likely pose a low ongoing takeover risk due to the limited capabilities of current LLMs and seemingly fundamental limitations of the deep learning paradigm, with some uncertainty due to surprises and the opaque nature of LLMs. However, other well-informed analysts are likely to hold different views, and it would be difficult to establish which views are correct.

Third, the risk from future LLMs depends on the particulars of a given LLM system. LLMs that are only minor deviations from the current state of the art are unlikely to pose a significant takeover risk. LLMs that deviate more, especially those that are larger in scale, may pose a larger risk of takeover catastrophe. However, this trend may have limits. At some point, the capabilities of LLMs may become maxed out, such that additional increases in LLM size do not bring increased capabilities. At that point, if LLMs have not yet caused takeover capacity, then they never will. Figure 1 sketches this concept. The exact contours of the curve in the sketch—what sort of shape it may have and where it would plateau—are not known and are likely a point of expert disagreement for reasons outlined above. Nonetheless, this provides a general heuristic on how to think about future LLM risk.

[9] Cremer (2021, p.449) defined high-level machine intelligence as "when unaided machines can accomplish every task better and more cheaply than human workers". An AI system or system-of-systems with this capability may be capable of takeover due to possessing the capabilities listed in Section 4.

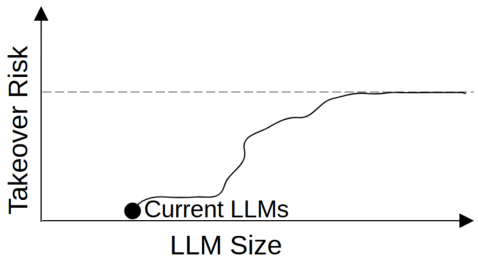

Figure 1. Sketch of takeover catastrophe risk from successively larger future LLMs. The curve is hand-drawn to emphasize the imprecise nature of the current understanding of future LLM risk.

In summary, the risk of takeover catastrophe from future LLMs contains significant uncertainties and cannot be precisely assessed. Some factors point toward low risk, such as potential shortages of training data and computing power and possible fundamental limitations of the deep learning paradigm. Other factors suggest that the risk may be higher, such as possibilities for synthetic data and algorithmic advances. Disagreement among experts and the potential for surprises make it difficult to place bounds on the risk. There is reason to believe that the risk is not zero, but exactly how much above zero is difficult to resolve. Future LLMs may pose a larger risk than current LLMs, especially if they are larger in scale, though at some point there may be diminishing returns such that additional increases in LLM size would not bring additional increases in takeover catastrophe risk.

6.2 AI Systems With LLM Components

Given the particular capabilities and limitations of current LLMs, some projects have used LLMs as components in broader AI systems. For example, LLMs have been combined with (1) planning and learning algorithms to play the social strategy game Diplomacy (FAIR et al., 2022); (2) the Hugging Face library of AI models to solve assorted AI tasks via a natural language interface (Shen et al., 2023); and (3) chemistry software designed by human experts to execute a variety of chemistry tasks (Bran et al., 2023). As researchers and developers become more acquainted with LLMs, this sort of system development will presumably continue.

AI systems containing LLMs could plausibly pose a takeover risk. Suppose takeover requires the specific characteristics listed in Section 4. LLMs may be especially suitable for some of those characteristics, whereas other types of AI are more suitable for other characteristics. For example, it is plausible that LLMs would be relatively skilled at social manipulation and certain aspects of economic productivity, whereas other types of AI may be needed for intelligence amplification and strategizing. In that case, AI systems containing LLMs may be the most viable means of causing takeover catastrophe.

In terms of LLM risk, this raises the question of how important LLMs are to the overall AI system. Perhaps LLMs are the primary system component, with other components playing a relatively minor supporting role. Or, perhaps LLMs would be relatively minor. Likewise, perhaps takeover would only be possible via major advances in LLM technology, whereas other components would be relatively basic. Or, perhaps major advances are needed in other components. Resolving these matters requires detailed analysis of specific potential AI systems, which is beyond the scope of this article. Nonetheless, given the potential limitations of LLMs, it is possible that the contribution of LLMs to the risk of takeover catastrophe comes mainly from the potential role in broader AI systems.

7. Implications

7.1 Governance

There is a large gap between the capabilities of current LLMs and the capabilities that may be needed for takeover (Section 5). Therefore, a case can be made that governance of current LLMs does not need to be heavily focused on takeover catastrophe risk, and instead the focus can be on the many other issues that the technology raises.

For future LLMs and AI systems with LLM components, the takeover catastrophe risk is more significant. AI governance initiatives should account for this risk, including the highly uncertain character of the risk and the divergence of expert opinion on it. In this sort of situation, which might be termed "post-normal" (Funtowicz and Ravetz, 2018) or "deeply uncertain" (Marchau et al., 2019), one approach is to pursue governance measures that perform reasonably well across the range of possible future outcomes. An AI takeover catastrophe would of course not constitute a "reasonably good" outcome; where governance measures can reduce this risk without imposing significant burdens, such measures may be worth pursuing.

The challenge is in identifying how far governance measures should go to reduce the risk of future LLM takeover catastrophe. For example, Yudkowsky (2023) proposed a global moratorium on new LLMs, backed by military force up to and potentially including actions to "destroy a rogue datacenter by airstrike". Such a policy would impose very significant burdens, including an increased risk of violent conflict. Some major AI developers are also major and nuclear-armed military powers (e.g., China and the United States), so this policy could increase the risk of nuclear war, which is itself an extreme global catastrophe scenario. This is a case of risk-risk tradeoff (Graham and Wiener, 1995; Lofstedt and Schlag, 2017); other risks should not be ignored in the pursuit of LLM risk reduction. This particular tradeoff is difficult to resolve given the uncertainties. However, given that current LLMs do not appear close to causing takeover catastrophe, they may not merit such extreme risk governance measures. Likewise, if the next LLMs to be built are similar to current LLMs, perhaps with incremental advances, they would be unlikely to close the gap between current capabilities and the capabilities needed for takeover catastrophe and therefore may also not merit extreme governance measures.

One governance measure that may be of high value is to monitor for warning signs of LLMs getting closer to the capabilities that may be needed for takeover catastrophe. Prior literature has proposed a variety of warning signs for AI takeover and related risks, including the achievement of certain milestones in future deep learning technology (Cremer & Whittlestone, 2021) and AI system attempts to acquire power (Carlsmith, 2022). An additional set of warning signs can be drawn from the present article: increases in the capabilities that may be needed for takeover as outlined in Section 4. Significant change in these warning signs could indicate that LLMs (or other AI systems) are getting closer to causing takeover catastrophe (i.e., moving higher up the curve sketched in Figure 1). This could in turn indicate that more aggressive anticipatory risk governance measures are warranted—and the measures would need to be anticipatory because if AI takeover catastrophe occurs, then further governance measures may become infeasible. Therefore, monitoring for warning signs could have high value as evaluated in terms of a "value of information" paradigm (Barrett, 2017). If warning signs are detected, that could trigger governance measures such as halting further R&D until the danger is resolved (Alaga & Schuett, 2023).

Finally, it may be constructive to pursue LLM governance measures that would help on potential future risks of LLM catastrophe along with other risks and issues posed by LLMs. Governance measures of this sort are "win-wins" in that they provide benefits across multiple domains without significant downsides. For example, initiatives to establish robust AI governance institutions, including in governments and in AI corporations, can improve AI governance across the full range of issues

posed by AI technology. Of particular relevance are measures for the governance of AI systems that advance the frontier of AI technology (Alaga & Schuett, 2023; Anderljung et al., 2023; DSIT, 2023b), especially systems that may pose catastrophic risks (Anthropic, 2023; OpenAI, 2023; Shevlane et al., 2023), noting that AI takeover is one of many potential AI catastrophe events. Nonetheless, robust governance of AI R&D is important for all future AI systems, regardless of whether they pose catastrophic risks; the same holds for initiatives to ensure that AI governance remains robust into the future even as AI technology changes.[10]

7.2 Research

The high degree of uncertainty about future LLMs suggests an important role for research to reduce the uncertainty. There are limits to how much current research can reduce uncertainty about future LLMs—the future LLMs do not exist yet, so they cannot be directly studied.

In the absence of empirical evidence about future LLMs, it may be tempting to study them via expert elicitation. However, caution is warranted. For some topics, there are no experts, meaning no people with an expert-level understanding (Morgan, 2014). Many aspects of future AI technology may be this sort of topic. For example, there may be no people with an expert-level understanding of the shape of Figure 1, the date by which LLMs may gain certain capabilities, or the probability of LLMs causing takeover catastrophe. Researchers considering using expert elicitation to study future LLMs should consider the possibility that the information produced would not be of high enough quality to justify conducting the study.

One specific topic worthy of further study is the characteristics that may be needed for takeover. Section 4 presents some ideas on this as contained in the prior literature, but the matter has not been explored in much depth. To be clear, there may be limits to how much this matter can be studied. However, given its importance for anticipatory risk governance of LLMs—and potentially also for other types of AI systems—it is worth exploring further. Research along these lines should be mindful of the risks posed by publishing about AI takeover risk (Section 2.2), as should all research on the topic.

Finally, and perhaps most importantly, there is a need for more detailed analysis of AI systems that contain LLMs. The reality is that LLMs are just one part of a broader AI ecosystem. A holistic assessment of AI takeover risk needs to cover the entire ecosystem. There are many ways in which LLMs can be integrated into broader AI systems; future research could survey the various options and assess their risks. In this research, it may be fruitful to distinguish between rapid single-system and gradual multi-system takeover scenarios, with research on the latter taking into account how humans may use a variety of AI systems to increasingly automate the economy.

8. Conclusion

This article presents a risk analysis of LLM takeover catastrophe. LLMs are arguably the first type of actual AI system to raise concerns about takeover catastrophe, and so the article contributes to a more empirical turn in AI takeover risk analysis.

The article's core method involves comparing the characteristics of LLMs to the characteristics that may be needed for takeover catastrophe. This method is not specific to LLMs; it can be applied to any type of AI system that raises concerns of takeover catastrophe. To better analyze the risk, the method should be developed further, such as via further research on the characteristics needed for takeover. The method can also be incorporated into governance activities such as monitoring for changes in the characteristics of LLMs and other AI systems.

[10] For general discussion of synergies between governance measures for different AI issues, including catastrophic risks, see Baum (2018b), Cave and ÓhÉigeartaigh (2019), and Stix and Maas (2021).

Fortunately, the article finds that, despite the considerable uncertainties about LLM takeover risk, current LLMs do not come close to exhibiting the capabilities needed for takeover catastrophe. It cannot be ruled out that current LLMs have capabilities not yet exhibited, but this may be unlikely due to the extensive attention they have received. Governance of current LLMs may proceed accordingly. For future LLMs, as well as future AI systems that include LLMs as components, the situation is more uncertain. Limitations of the deep learning paradigm suggest that LLMs are not en route to takeover catastrophe, but uncertainties and expert disagreements suggest some nontrivial (but not readily quantified) probability of catastrophe. Some degree of caution is warranted for the governance of future LLMs and AI systems, and it may be particularly prudent to monitor for warning signs of catastrophe. An AI takeover catastrophe would likely be an irreversible event, so it is imperative for governance to be anticipatory, to get this right so that the catastrophe event does not occur in the first place.

Acknowledgments
Tony Barrett, editor Vicki Bier, and two anonymous reviewers provided helpful feedback on earlier drafts. All remaining mistakes are the author's alone.

References

Ahmad, B., Thakur, S., Tan, B., Karri, R., & Pearce, H. (2023). Fixing hardware security bugs with large language models. https://arxiv.org/abs/2302.01215.

Alaga, J., & Schuett, J. (2023). Coordinated pausing: An evaluation-based coordination scheme for frontier AI developers. https://arxiv.org/abs/2310.00374.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. N*ew Media & Society*, *20*(3), 973-989.

Anderljung, M., Barnhart, J., Leung, J., Korinek, A., O'Keefe, C., Whittlestone, J., et al. (2023). Frontier AI regulation: Managing emerging risks to public safety. https://arxiv.org/abs/2307.03718.

Anthropic (2023). Core views on AI safety: When, why, what, and how. https://www.anthropic.com/index/core-views-on-ai-safety

Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, *22*, 299-324.

Armstrong, S., & O'Rorke, X. (2018). Safe uses of AI oracles. https://arxiv.org/abs/1711.05541.

Barrett, A. M. (2017). Value of global catastrophic risk (GCR) information: Cost-effectiveness-Based approach for GCR reduction. *Decision Analysis*, *14*(3), 187-203.

Barrett, A. M., & Baum, S. D. (2017). A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, *29*(2), 397–414.

Bashir, D., & Kurenkov, A. (2022). The AI scaling hypothesis. *Last Week in AI*, 5 August, https://lastweekin.ai/p/the-ai-scaling-hypothesis.

Baum, S. D. (2018a). Superintelligence skepticism as a political tool. Information, *9*, 209.

Baum, S. D. (2018b). Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI & Society*, *33*(4), 565-572.

Baum, S. D., & Barrett, A. M. (2018). Global catastrophes: The most extreme risks. In: *Risk in extreme environments: Preparing, avoiding, mitigating, and managing* (pp. 174-184). Routledge.

Bell, E. (2023). A fake news frenzy: Why ChatGPT could be disastrous for truth in journalism. *The Guardian*, 3 March, https://www.theguardian.com/commentisfree/2023/mar/03/fake-news-chatgpt-truth-journalism-disinformation.

Bezuidenhout, L. (2013). Data sharing and dual-use issues. *Science and Engineering Ethics*, *19*, 83-92.

Boiko, D. A., MacKnight, R., Gomes, G. (2023). Emergent autonomous scientific research capabilities

of large language models. https://arxiv.org/abs/2304.05332.
Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
Bostrom, N. (2017). Strategic implications of openness in AI development. *Global Policy*, *8*(2), 135-148.
Bran, A. M., Cox, S., White, A. D., & Schwaller, P. (2023). ChemCrow: Augmenting large-language models with chemistry tools. https://arxiv.org/abs/2304.05376
Branwen, G. (2020). The scaling hypothesis. https://gwern.net/scaling-hypothesis
Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work. National Bureau of Economic Research Working Paper w31161.
Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. https://arxiv.org/abs/2303.12712
Carlsmith, J. (2022). Is power-seeking AI an existential risk? https://arxiv.org/abs/2206.13353.
Carlsmith, J. (2023). Scheming AIs: Will AIs fake alignment during training in order to get power? https://arxiv.org/abs/2311.08379.
Cave, S., & ÓhÉigeartaigh, S. S. (2019). Bridging near-and long-term concerns about AI. *Nature Machine Intelligence*, *1*(1), 5-6.
Charan, P. V., Chunduri, H., Anand, P. M., & Shukla, S. K. (2023). From text to MITRE techniques: Exploring the malicious use of large language models for generating cyber attack payloads. https://arxiv.org/abs/2305.15336.
Chowdhury, H. (2023). ChatGPT cost a fortune to make with OpenAI's losses growing to $540 million last year, report says. *Business Insider*, 5 May. https://www.businessinsider.com/openai-2022-losses-hit-540-million-as-chatgpt-costs-soared-2023-5
Christiano, P. (2019). What failure looks like. *Alignment Forum*, 17 March. https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like
Ćirković, M. M. (2012). Small theories and large risks—is risk analysis relevant for epistemology?. *Risk Analysis*, *32*(11), 1994-2004.
Cremer, C. Z. (2021). Deep limitations? Examining expert disagreement over deep learning. *Progress in Artificial Intelligence*, *10*, 449-464.
Cremer, C. Z., & Whittlestone, J. (2021). Artificial Canaries: Early warning signs for anticipatory and democratic governance of AI. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(5), 100-109.
Critch, A. (2021). What multipolar failure looks like, and robust agent-agnostic processes (RAAPs). *Alignment Forum*, 31 March, https://www.alignmentforum.org/posts/LpM3EAakwYdS6aRKf/what-multipolar-failure-looks-like-and-robust-agent-agnostic.
DSIT (Department for Science, Innovation and Technology) (2023a). Frontier AI: capabilities and risks – discussion paper. https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper.
DSIT (Department for Science, Innovation and Technology) (2023b). Emerging processes for frontier AI safety. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety.
Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. https://arxiv.org/abs/2303.10130.
Everitt, T., Lea, G., & Hutter, M. (2018). AGI safety literature review. https://arxiv.org/abs/1805.01109.
FAIR (Meta Fundamental AI Research Diplomacy Team), Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., et al. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, *378*(6624), 1067-1074.

Falade, P. V. (2023). Decoding the threat landscape: ChatGPT, FraudGPT, and WormGPT in social engineering attacks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *9*(5), 185-198.
Ferruz, N., & Höcker, B. (2022). Controllable protein design with language models. *Nature Machine Intelligence*, *4*(6), 521-532.
Finger, L. (2023). OpenAI isn't going bankrupt, but it has a business model problem. *Forbes*, 18 August. https://www.forbes.com/sites/lutzfinger/2023/08/18/is-openai-going-bankrupt-no-but-ai-models-dont-create-moats
Fitzgerald, M., Boddy, A., Baum, S. D. (2020). 2020 survey of artificial general intelligence projects for ethics, risk, and policy. Global Catastrophic Risk Institute Technical Report 20-1.
FLI (Future of Life Institute (2023). Pause giant AI experiments: An open letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments.
Funtowicz, S., & Ravetz, J. (2018). Post-normal science. In *Companion to Environmental Studies* (pp. 443-447). Routledge.
Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., et al. (2022). Predictability and surprise in large generative models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* pp. 1747-1764.
Gebru, T., Bender, E. M., McMillan-Major, A., Mitchell, M. (2023). Statement from the listed authors of Stochastic Parrots on the "AI pause" letter. https://www.dair-institute.org/blog/letter-statement-March2023.
Gent, E. (2023). When AI's large language models shrink. *IEEE Spectrum*, 31 March. https://spectrum.ieee.org/large-language-models-size
Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1-48.
Goertzel, B. (2023). Is ChatGPT real progress toward human-level AGI? https://bengoertzel.substack.com/p/is-chatgpt-real-progress-toward-human.
Goldstein, S., & Kirk-Giannini, C. D. (2023). Language agents reduce the risk of existential catastrophe. *AI & Society*, in press, https://doi.org/10.1007/s00146-023-01748-4.
Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In F. L. Alt & M. Rubinoff (Eds.), *Advances in computers* (pp. 31-88), Academic Press.
Graham, J. D., & Wiener, J. B. (1995). *Risk vs. risk: Tradeoffs in protecting health and the environment*. Cambridge, MA: Harvard University Press.
Gupta, A., Lanteigne, C., & Heath, V. (2020). Report prepared by the Montreal AI Ethics Institute (MAIEI) on publication norms for responsible AI. https://arxiv.org/abs/2009.07262.
Guston, D. H. (2014). Understanding 'anticipatory governance'. *Social Studies of Science*, *44*(2), 218-242.
Hagendorff, T. (2023). Deception abilities emerged in large language models. https://arxiv.org/abs/2307.16513
Haimes, Y. Y. (2018). Risk modeling of interdependent complex systems of systems: Theory and practice. *Risk Analysis*, *38*(1), 84-98.
Haluptzok, P., Bowers, M., & Kalai, A. T. (2022). Language models can teach themselves to program better. https://arxiv.org/abs/2207.14502
Happe, A., & Cito, J. (2023). Getting pwn'd by AI: Penetration testing with large language models. *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 2082-2086.
Heaven, W. D. (2022). Why Meta's latest large language model survived only three days online. *MIT Technology Review*, 18 November. https://www.technologyreview.com/2022/11/18/1063487/meta-

large-language-model-ai-only-survived-three-days-gpt-3-science.
Hecht, B., Wilcox, L., Bigham, J.P., Schöning, J., Hoque, E., Ernst, J., Bisk, Y., De Russis, L., Yarosh, L., Anjum, B., Contractor, D., & Wu, C. 2018. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. ACM Future of Computing Blog. https://acm-fca.org/2018/03/29/negativeimpacts.
Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (2022). Large language models can self-improve. https://arxiv.org/abs/2210.11610
Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. https://arxiv.org/abs/2311.05232
Irwin, R., Dimitriadis, S., He, J., & Bjerrum, E. J. (2022). Chemformer: A pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, *3*(1), 015022.
Kambhampati, S. (2024). Can large language models reason and plan? *Annals of The New York Academy of Sciences*, https://doi.org/10.1111/nyas.15125.
Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., & Irving, G. (2021). Alignment of language agents. https://arxiv.org/abs/2103.14659.
Kilian, K. A., Ventura, C. J., & Bailey, M. M. (2023). Examining the differential risk from high-level artificial intelligence and the question of control. *Futures*, *151*, 103182.
Leahy, C. (2021). Why release a large language model? https://blog.eleuther.ai/why-release-a-large-language-model.
LeCun, Y. (2022). A path towards autonomous machine intelligence. https://openreview.net/pdf?id=BZ5a1r-kVsf
Li, J., Dada, A., Kleesiek, J., & Egger, J. (2023). ChatGPT in healthcare: A taxonomy and systematic review. https://www.medrxiv.org/content/10.1101/2023.03.30.23287899v1.
Lipsitch, M., Evans, N. G., & Cotton-Barratt, O. (2017). Underprotection of unpredictable statistical lives compared to predictable ones. *Risk Analysis*, *37*(5), 893-904.
Lofstedt, R., & Schlag, A. (2017). Risk-risk tradeoffs: What should we do in Europe? *Journal of Risk Research*, *20*(8), 963-983.
Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2022). Estimating the carbon footprint of bloom, a 176b parameter language model. https://arxiv.org/abs/2211.02001
Marchau, V. A., Walker, W. E., Bloemen, P. J., & Popper, S. W. (2019). *Decision making under deep uncertainty: From theory to practice*. Springer Nature.
Marcus, G. (2023). GPT-5 and irrational exuberance. https://garymarcus.substack.com/p/gpt-5-and-irrational-exuberance.
Marcus, G. F., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon Books.
Marcus, G. F., & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*, 22 August. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion
Merchant, B. (2023). Afraid of AI? The startups selling it want you to be. *Los Angeles Times*, 31 March, https://www.latimes.com/business/technology/story/2023-03-31/column-afraid-of-ai-the-startups-selling-it-want-you-to-be.
Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*, *111*(20), 7176-7184.
Motlagh, F. N., Hajizadeh, M., Majd, M., Najafi, P., Cheng, F., & Meinel, C. (2024). Large language models in cybersecurity: State-of-the-art. https://arxiv.org/abs/2402.00891.

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

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192.

O'Gara, A. (2023). Hoodwinked: Deception and cooperation in a text-based game for language models. https://arxiv.org/abs/2308.01404.

Omohundro, S. M. (2008). The basic AI drives. In P. Wang, B. Goertzel, and S. Franklin (Eds.), *Artificial general intelligence 2008: Proceedings of the first AGI conference*. IOS Press, pp.483–492.

OpenAI (2023). GPT-4 technical report. https://arxiv.org/abs/2303.08774.

Ord, T., Hillerbrand, R., & Sandberg, A. (2010). Probing the improbable: methodological challenges for risks with low probabilities and high stakes. *Journal of Risk Research*, *13*(2), 191-205.

Oreskes, N., & Conway, E. M. (2010). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). AI deception: A survey of examples, risks, and potential solutions. https://arxiv.org/abs/2308.14752.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., et al. (2022). Discovering language model behaviors with model-written evaluations. https://arxiv.org/abs/2212.09251.

Perrigo, B. (2023). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *TIME*, 18 January, https://time.com/6247678/openai-chatgpt-kenya-workers.

Pilz, K., & Heim, L. (2023). Compute at scale: A broad investigation into the data center industry. https://arxiv.org/abs/2311.02651.

Posner, R. (2004). *Catastrophe: Risk and response*. Oxford University Press.

Ray, T. (2023). With GPT-4, OpenAI opts for secrecy versus disclosure. *ZDNet*, 16 March, https://www.zdnet.com/article/with-gpt-4-openai-opts-for-secrecy-versus-disclosure.

Riedl, M. (2023). A very gentle introduction to large language models without the hype. https://mark-riedl.medium.com/a-very-gentle-introduction-to-large-language-models-without-the-hype-5f67941fa59e.

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and benefits of large language models for the environment. *Environmental Science & Technology*, *57*(9), 3464-3466.

Rotman, D. (2023). ChatGPT is about to revolutionize the economy. We need to decide what that looks like. *MIT Technology Review*, 25 March, https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of Large Language Models a mirage? https://arxiv.org/abs/2304.15004.

Scharre, P. (2024). Future-proofing frontier AI regulation: Projecting future compute for frontier AI models. https://www.cnas.org/publications/reports/future-proofing-frontier-ai-regulation.

Sejnowski, T. J. (2018). *The deep learning revolution*. MIT press.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). HuggingGPT: Solving ai tasks with chatgpt and its friends in huggingface. https://arxiv.org/abs/2303.17580.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., et al. (2023). Model evaluation for extreme risks. https://arxiv.org/abs/2305.15324.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140-1144.
Soice, E. H., Rocha, R., Cordova, K., Specter, M., & Esvelt, K. M. (2023). Can large language models democratize access to dual-use biotechnology? https://arxiv.org/abs/2306.03809.
Sotala, K. (2018). Disjunctive scenarios of catastrophic AI risk. In: *Artificial intelligence safety and security* (pp. 315-337). Chapman and Hall/CRC.
Sotala, K., & Gloor, L. (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica*, *41*(4), 389-400.
Stix, C., & Maas, M. M. (2021). Bridging the gap: the case for an 'Incompletely Theorized Agreement'on AI policy. *AI and Ethics*, *1*(3), 261-271.
Stokel-Walker, C. (2023). The generative AI race has a dirty secret. Wired, 18 February, https://www.wired.com/story/the-generative-ai-search-race-has-a-dirty-secret.
Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., et al. (2022). Galactica: A large language model for science. https://arxiv.org/abs/2211.09085.
To, H. Q., Bui, N. D., Guo, J., & Nguyen, T. N. (2023). Better Language Models of Code through Self-Improvement. https://arxiv.org/abs/2304.01228.
Treude, C., & Hata, H. (2023). She elicits requirements and he Tests: Software engineering gender bias in large language models. https://arxiv.org/abs/2303.10131.
Vallance, C. (2023). AI could replace equivalent of 300 million jobs – report. *BBC*, 28 March, https://www.bbc.com/news/technology-65102150.
Valmeekam, K., Sreedharan, S., Marquez, M., Olmo, A., & Kambhampati, S. (2023a). On the planning abilities of large language models (A critical investigation with a proposed benchmark). https://arxiv.org/abs/2302.06706.
Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2023b). Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change). https://arxiv.org/abs/2206.10498.
Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., & Ho, A. (2022). Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. https://arxiv.org/abs/2211.04325
Vincent, J. (2023). OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong'. *The Verge*, 15 May, https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview.
Vinge, V. (1993). The coming technological singularity: How to survive in the post-human Era. In: *Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace*. NASA, https://ntrs.nasa.gov/citations/19940022856.
Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*, 31 August, https://openreview.net/forum?id=yzkSU5zdwD.
Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., et al. (2021). Ethical and social risks of harm from language models. https://arxiv.org/abs/2112.04359
White House (2023). Press briefing by Press Secretary Karine Jean-Pierre. https://www.whitehouse.gov/briefing-room/press-briefings/2023/03/30/press-briefing-by-press-secretary-karine-jean-pierre-22.
Wiener, J. B. (2016). The tragedy of the uncommons: On the politics of apocalypse. *Global Policy*, *7*, 67-80.
Xu, C., Wang, Y., & Barati Farimani, A. (2023). TransPolymer: A transformer-based language model

for polymer property predictions. *npj Computational Materials*, *9*(1), 64.
Yaman, F. (2023). AgentSCA: Advanced physical side channel analysis agent with LLMs. Masters Thesis, Computer Engineering, North Carolina State University.
Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211. https://doi.org/10.1016/j.hcc.2024.100211.
Yudkowsky, E. (2023). Pausing AI developments isn't enough. We need to shut it all down. *TIME*, 29 March, https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough.
Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. https://arxiv.org/abs/2303.18223.
Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., et al. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 20. https://doi.org/10.1145/3639372.