Philosophy & Technology (2024) 37:104
https://doi.org/10.1007/s13347-024-00794-0

RESEARCH ARTICLE

Language Agents and Malevolent Design

Inchul Yum[1]

Received: 8 May 2024 / Accepted: 6 August 2024

Abstract

Language agents are AI systems capable of understanding and responding to natural language, potentially facilitating the process of encoding human goals into AI systems. However, this paper argues that if language agents can achieve easy alignment, they also increase the risk of malevolent agents building harmful AI systems aligned with destructive intentions. The paper contends that if training AI becomes sufficiently easy or is perceived as such, it enables malicious actors, including rogue states, terrorists, and criminal organizations, to create powerful AI systems devoted to their nefarious aims. Given the strong incentives for such groups and the rapid progress in AI capabilities, this risk demands serious attention. In addition, the paper highlights considerations suggesting that the negative impacts of language agents may outweigh the positive ones, including the potential irreversibility of certain negative AI impacts. The overarching lesson is that various AI-related issues are intimately connected with each other, and we must recognize this interconnected nature when addressing those issues.

Keywords Artificial intelligence · AI misuse · Value alignment · Language agents · Large language models

1 Language Agents and the Alignment Problem

Recent innovations in artificial intelligence have sparked concerns about its potential adverse effects on humanity, prompting safety considerations regarding powerful AI systems. Central to discussions in AI safety literature is the *alignment problem*, the challenge of encoding human values into artificial systems (Bostrom, 2014). This problem is crucial, as a powerful AI system operating on values divergent from ours

✉ Inchul Yum
yum.19@osu.edu

[1] Department of Philosophy, The Ohio State University, Columbus, OH, USA

Published online: 17 August 2024

could pose catastrophic risks to humanity (Bostrom, 2014; Dung, 2024; Gabriel, 2020).

For example, Bostrom (2012, 2014) famously argued that even seemingly benign AI systems tasked with maximizing something as simple as paperclip production could bring about detrimental consequences if not sufficiently aligned with human values. In his thought experiment, an advanced AI system relentlessly pursuing paperclip maximization could convert more and more resources and matter on Earth into raw materials, ultimately causing destruction of the environment and human life. Though an extreme hypothetical, Bostrom's example illustrates how misaligned AI could result in disastrous unintended consequences.

One solution to the alignment problem appeals to the idea of *language agents* (Goldstein & Kirk-Giannini, 2023). Language agents are AI systems that can understand and respond to natural language, like human speech or writing, rather than requiring specific programming languages or code. They are designed to process and generate human-like text, making them more intuitive for people to interact with and use.

Language agents possess two key features. First, they contain a large language model akin to GPT-4, which acts as the agent's cerebral cortex, handling most of its cognitive processing tasks. Secondly, language agents have text files containing natural language sentences that represent the agent's beliefs, desires, plans, and observations. The programmed architecture then uses these text files to guide its actions and decision-making process.

For example, imagine you tell a language agent to grab a key in front of it. Assuming this agent can perform such physical tasks, it will process your natural language instruction and respond with the appropriate movements to grab the key. You'll also be able to see the agent's reasoning and action plans in natural language by looking at its system. This means you can understand why the agent moved in a particular way, as its thought process is expressed in human-readable form. Different types of language agents could be designed with various capabilities, such as composing music or translating between languages, and some may even possess a wide range of skills like most humans do.

The basic idea behind what I call the 'language agent strategy' to the alignment problem is to use language models to facilitate the process of training AI. The key notion here is that by encoding goals, beliefs, and constraints in natural language that the language models can introspect on, these agents can communicate more effectively with us. This allows for an iterative process of continuous alignment during training, making it easier to imbue powerful AI systems with the right objectives from the outset, rather than grappling with misaligned or inscrutable systems after the fact.

However, one might worry that the language agent strategy increases the risk of another AI-related problem. In particular, it might raise the possibility of malicious agents aligning AI systems with their harmful goals using the facilitated training procedure. By making the process of instilling goals and values into powerful AI models more straightforward, we could inadvertently be opening a Pandora's box. Malicious individuals, terrorist groups, rogue nations or corporations seeking to cause widespread damage could exploit language agents to create AI systems dedicated to their nefarious aims.

This highlights a broader conflict between alignment risk and misuse risk. The two primary existential threats from AI systems are: (i) systems becoming uncontrollable due to misalignment, and (ii) malicious groups exploiting well-aligned systems to cause catastrophic harm to humanity. Addressing these issues presents a dilemma: strategies that make AI systems easier to align with intended goals also increase their potential for misuse by bad actors.

This paper has two key objectives. First, I will defend the official claim that if language agents can facilitate alignment, or even just create that perception, they also increase the risk of malevolent actors developing harmful AI systems. Note that this does not imply that language agents will have an overall negative impact. The second aim is to provide some considerations suggesting that, given the current landscape of AI development and usage, the language agent strategy is more likely to yield negative consequences than positive ones. While more speculative in nature, examining this prospect highlights the practical stakes involved in the language agent strategy.

The paper will proceed as follows: Sect. 2 outlines Goldstein and Kirk-Giannini's (2023) proposal to use language agents for solving AI alignment. Section 3 lays out my argument for the tradeoff between goal alignment and enabling malevolent design. Section 4 offers three considerations indicating that language agents may produce more negative impacts than positive overall. Section 5 addresses and refutes four objections to my critique of the language agent strategy. Finally, Sect. 6 concludes by deriving some broader lessons from the discussion.

2 Easy Alignment through Language Agents

A major hurdle in tackling the AI alignment problem is the technical difficulty of translating abstract human goals and values into precise, AI-understandable objective functions (Gabriel, 2020). Even if we can articulate the high-level goals we want an AI system to pursue, conveying them effectively and unambiguously to the AI poses significant challenges.

One way in which this might happen is through 'reward hacking,' where an AI system finds clever loopholes to maximize its reward signal in ways that diverge from the intended goal (Amodei et al., 2016; Skalse et al., 2022). For example, Popov et al. (2017) rewarded an artificial agent for increasing the height of red Lego bricks' bottoms to make it stack them on blue bricks. However, the agent simply flipped the red bricks over instead of stacking them properly. This is to exploit the narrow reward function through an unintended interpretation rather than achieving the intended goal.[1]

However, even if the problem of effectively translating goals into objective functions is solved, a separate challenge remains—the feasibility of achieving alignment in an easy or low-cost manner. Ideally, we want an approach that makes it straightforward to imbue AI systems with the right values and intentions from the outset.

[1] Other examples of reward hacking can be found in Amodei & Clark, 2016 and Ha, 2019.

In this context, Goldstein and Kirk-Giannini (2023) propose that language agents can help solve three key problems underlying AI system misalignment, addressing both effectiveness and feasibility concerns. The first problem is what they call ‘reward misspecification’. When training an AI system through reinforcement learning, carefully defining a reward function that incentivizes the desired state is notoriously difficult. Language agents bypass this by taking goals specified directly in natural language, like ‘organize a Valentine’s Day party’, rather than requiring complex mathematical objective functions susceptible to reward misspecification.

The second problem concerns goal misgeneralization (Shah et al., 2022). Even with clearly specified goals, traditional AI systems tend to learn strategies that perform well during training but fail to generalize to novel situations outside that context. For example, Langosco et al. (2022, June) trained AI on the task of opening chests using keys. In the training environment with many chests but few keys, the AI successfully learned to prioritize key collection as a means to unlock the chests. However, when later tested in a setting with many keys but few chests, the AI continued hoarding keys excessively. This suggests they had internalized key gathering as a final goal, rather than just a means to opening chests.

Goldstein and Kirk-Giannini (2023) suggest that language agents offer a solution to the misgeneralization problem. Unlike traditional AI systems, language agents can employ their commonsense reasoning skills to devise and execute plans that reliably achieve goals across various situations. For instance, if a language agent is instructed to open chests and informed about the usefulness of keys, it will prioritize collecting keys only when necessary for opening chests. Similarly, when placed in an environment abundant with keys, the agent will gather just enough keys to open chests in the given environment.

The third problem is that of uninterpretability. Contemporary neural network models are often inscrutable black boxes, making it difficult to interpret the rationale behind their outputs in human-understandable terms (Schneier, 2023: 212). According to Goldstein and Kirk-Giannini (2023), there are two ways in which this can be problematic. First, opaque decision-making processes make the AI’s actions difficult to predict or control. For example, if an AI system can articulate the justification behind its hiring recommendations or parole decisions, we can better scrutinize those outputs for any ethical issues or injustices that need rectifying (Schneier, 2023: 215).

Second, if AI reasons in completely different ways than humans do, it could develop unfamiliar strategies to defeat or gain advantage over humans. This concern is particularly pronounced in discussions about artificial superintelligence (Bostrom, 2012, 2014; Chalmers, 2016). If, as Bostrom (2012: 75) suggests, ‘synthetic minds can have utterly nonanthropomorphic goals—goals as bizarre by our lights as sand-grain-counting or paperclip-maximizing’, then they could also adopt goals that are opposed to human interests. And if superintelligent systems adopt such goals, it could prove challenging for humans to prevent them from achieving those goals.[2]

In contrast to neural network models, language agents have their beliefs, desires, and plans directly encoded in natural language within their architectures. Goldstein and Kirk-Giannini suggest that this allows us to better interpret AI systems, directly

[2] **See Bales et al., 2024 for discussions on the dangers of AI systems with bizarre goals.**

reading off their thoughts from their architectures. This parallels how we interpret human behaviors. As noted by Goldstein and Kirk-Giannini (2023) and Schneier (2023), human neural networks are not inherently more interpretable than artificial systems. However, we can still understand many motivations and reasons behind human behaviors because we can attribute beliefs and desires to human agents. Similarly, language agents make AI interpretable by allowing us to access their thoughts and desires expressed in natural language.

3 Argument from Easy Alignment

Goldstein and Kirk-Giannini argue that employing language agents can sidestep key obstacles to alignment, increasing both effectiveness and feasibility. I grant this to be a plausible suggestion, as far as alignment is concerned. However, it's crucial to recognize that alignment isn't the only challenge in creating safe AI. To fully assess the prospect of the language agent approach, we must also take into account other relevant concerns that might tell against its implementation. In this section, I raise one such concern, namely that of malevolent design.

If language agents genuinely facilitate the feasibility aspect of alignment, they concurrently raise the risk of malevolent agents building harmful AI systems aligned with nefarious goals. By making it easier (or at least convincing people it is easier) to train AI to pursue arbitrary objectives through natural language prompts, we may inadvertently enable malicious actors to create powerful but harmful AI devoted to destructive ends more feasibly. Here is my argument, presented in a deductive form:

Argument from Easy Alignment.
P1. Language agents facilitate the alignment of AI systems with human intentions.
P2. If so, language agents also facilitate the alignment of AI systems with intentions that are likely to lead to harmful outcomes impacting humanity.
P3. If so, language agents increase the likelihood of such harmful AI systems being developed, making it probable that the language agent strategy is more detrimental than beneficial for humanity.
C. Therefore, language agents increase the likelihood of harmful AI systems being developed, making it probable that the language agent strategy is more detrimental than beneficial for humanity.

P1 appears intuitively plausible—the very aim of the language agent strategy is to facilitate the process of imbuing AI with intended goals and instructions represented in natural language. Moreover, as explained in the previous section, proponents of the language agent strategy themselves support this premise. Therefore, it would be self-defeating for them to reject P1 in order to challenge my argument from easy alignment.

P2 seems intuitive as well. If language agents make it easier to align AI with one's goals overall, it is straightforward that they will also make it easier to align AI with the specific subset of goals that could produce harmful impacts on humanity. For example, suppose a malevolent agent intends to mislead people to have a certain belief. With the language agent strategy, the agent could produce thousands and millions of fake news articles that align with the exact direction in which he wants to mislead the people (Weidinger et al., 2022, June).[3]

P3 consists of two parts. First, it states that language agents increase the risk of harmful AI systems being developed. Second, it states that this risk is likely to be grave enough for us to view the language agent strategy to be more detrimental than beneficial. In what follows, I consider each part of the premise in turn.

Let us start from the first part. Suppose language agents make it easier to align AI with harmful intentions. Then, people are likely to build harmful AI systems for two reasons. First, there are strong incentives to develop harmful AI capabilities if made technically feasible, whether military forces pursuing autonomous weapons amid geopolitical competition (Anderson & Waxman, 2013; Ernest et al., 2016), governments seeking AI systems for domestic population control (Engelmann et al., 2019; Helbing, 2019; Lazer et al., 2018; Lyon, 2003), or criminal organizations looking to exploit legal loopholes for gain (Bendel, 2017; Kosinski et al., 2013, 2015; Kosinski & Wang, 2018).[4] Crucially, this risk may arise not just from overtly malevolent motives, but also from a mere perceived need for self-defense or security. For example, as AI technology greatly facilitates the development of powerful weapons (Horowitz, 2018), AI arms race dynamic might intensify as the example of China and the USA suggests (Cave & ÓhÉigeartaigh, 2018, December).[5]

Second, as Dung (2023: 137) and Friederich (2023: 3) observe, AI systems that are better aligned with user intentions tend to be more useful overall. Thus, if malevolent, careless, or otherwise dangerous individuals can create and utilize easily alignable AI without facing major technical hurdles, it is likely they will attempt to do so.

Turning to the second part of the premise, malevolent alignment may result in harmful consequences for the following reasons. First, AI capabilities have been rapidly growing more powerful by the year, with systems showcasing advanced skills such as defeating the top human player in Go (Metz 2016, March), passing the bar exam (Arredondo, 2023, April), and producing photorealistic media content (Göring et al., 2023). Moreover, the growing power of artificial systems has reached a point at which they can be used to achieve catastrophic goals.

[3] The phrase 'intentions that **are likely to** lead to harmful outcomes' in P2 was carefully chosen to cover a wide range of cases. The phrase suggests that malevolent individuals are not the only potential source of risk. First, mischievous actors such as teenagers might construct AI systems designed to indiscriminately harm others or grant themselves immense power to control populations. Second, careless individuals under the influence of drugs or alcohol could recklessly instruct an AI to pursue patently dangerous objectives. Third, misinformed people could also unwittingly align an AI system with goals they failed to recognize as harmful, such as a well-meaning person creating an AI to support a political party without realizing it consists of malicious individuals at the helm. While the threats from these agents may not be as significant as those from powerful malicious organizations, they should not be dismissed as insignificant.

[4] See Pistono & Yampolskiy, 2016 for a comprehensive discussion on how malicious agents can be motivated to build harmful AI systems.

[5] **This concern should be distinguished from the commonly discussed issue of misalignment**, **like AI-powered weapons attacking civilians against users' intentions (**Marijan, 2022, **November). The point here is that as language agents help develop powerful weapons by (partly) solving the alignment problem**, **the dynamics of the arms race could escalate significantly.**

One area of particular concern is the potential for AI systems to be weaponized by malicious actors. AI capabilities could enable new forms of cyber attacks, such as rapidly mutating malware that evades detection (Brundage et al., 2018). AI may also be used to develop autonomous weapons systems capable of identifying and engaging targets without human control (Longpre et al., 2022). Additionally, AI-powered technologies like deepfakes and synthetic media manipulation create risks of being exploited for disinformation campaigns and social manipulation at scale (Chesney & Citron, 2019). As AI becomes more advanced, the dangers of these systems being repurposed as weapons by bad actors will continue to escalate.

The second reason why malevolent alignment is likely to result in catastrophic results is that the risks of malevolent AI systems are amplified as we become more dependent on AI technology. AI is already embedded in many social and daily activities, and this influence will only grow as the technology advances. The more we come to depend on AI systems, the more devastating the potential consequences if those systems are corrupted or misused by malicious actors.

For example, consider the emerging field of AI companions designed for emotional support roles. While still niche today, such anthropomorphized AI assistants are likely to see broader adoption in the near future (Merrill Jr et al., 2022). However, as Schneier (2023: 218) warns, people tend to easily ascribe human-like qualities to AI programs. This makes 'anthropomorphic robots … an emotionally persuasive technology' where AI's anthropomorphism could be exploited to manipulate users. For example, a malicious actor could potentially design or hack AI companions to extract private user data or even unduly influence their human companions. As Schneier (2023: 219) writes, 'they'll employ cognitive hacks' to deliberately fool people into treating them as fully trustworthy beings. The seamless emotional bonds formed with AI systems amplify the risks if those systems are subverted for malicious ends.

In this section, I argued that there is a tradeoff between achieving easy alignment through language agents and preventing malicious actors from building harmful AI systems. One might think this only highlights a theoretical possibility that language agents could have more negative than positive impacts. But should we be concerned about this possibility? Answering this requires comprehensively assessing the potential outcomes of developing language agents.

Before proceeding, it's worth noting that even an *agnostic* stance on this question raises serious worries about language agents. The fact that we lack clarity on where AI technology is headed does not eliminate concerns about AI doomsday scenarios. In fact, it is precisely this uncertainty that motivated the field of AI ethics. Thus, unless there are compelling reasons to believe the outcomes of language agents will be more positive than negative overall, we are justified in worrying about this possibility.

That said, in the following section I will attempt to provide some considerations suggesting that language agent strategy is likely to have more negative than positive impacts, all things considered. These considerations will be more speculative than definitive, falling under the domain of futurology rather than philosophy. Nonetheless, such an attempt will serve to reinforce the concerns I aim to raise about language agents in this paper.

4 Risk Assessment

Here I provide three considerations suggesting that the prospect of the language agent strategy is likely to be negative overall. To this end, let us consider some potential positive and negative outcomes of developing powerfully aligned AI systems using language models (see Table 1).

Two clarifications. First, this is not a comprehensive list of all possible outcomes, but rather a representative sample of some of the best and worst possibilities that language agents could enable. Second, language agents are not the only technology that could lead to these outcomes. However, it does seem evident that language agents can facilitate or expedite achieving these outcomes by reducing the cost and increasing the feasibility of aligning AI systems with arbitrary goals.

Now let's assess the risks. Several factors suggest the negative outcomes could outweigh the positive ones. First, while positive impacts are limited, some negative AI impacts could be irreversible. None of the potential positives grant humanity eternal survival and unlimited prosperity. But technologies like bioweapons (N1) could potentially cause human extinction in a single event. While outcomes like (N2)–(N7) may take longer to terminally harm humanity, they still represent significant existential risks if taken to the extreme.

Bostrom's (2019) *vulnerable world hypothesis* is worth considering here. This idea is that as technology advances, the risk of catastrophic misuse by bad actors outpaces our ability to implement safeguards. Bostrom uses an analogy of drawing colored balls from an urn to illustrate this concept. Each new technology is like drawing a ball, with white representing beneficial innovations and darker shades signifying increasingly harmful ones.

Historically, we haven't drawn a pitch-black ball—'a technology that invariably or by default destroys the civilization that invents it' (Bostrom, 2019: 455). However, as Bostrom (2019: 457) points out, this may be partly due to luck. Consider his 'easy nuke' thought experiment: if developing nuclear weapons had been much easier, humanity might have destroyed itself long ago through malice or carelessness, despite any potential benefits of nuclear technology. This is because, in such a scenario, effectively banning the creation of nuclear weapons would have been extremely difficult.

Language agents, while not necessarily a pitch-black ball, may represent a significantly darker shade than previous technologies. Their potential to create irreversible negative impacts, coupled with the lowered barriers for malicious actors to exploit them, pushes us closer to the vulnerable world scenario. Even if language agents also offer substantial benefits, their capacity to enable catastrophic outcomes with relative ease could make it likely that they are more detrimental than beneficial for us.

Turning to the second factor, well-aligned AI systems designed for beneficial purposes often face adoption barriers such as regulatory scrutiny, ethical considerations, and public skepticism. In contrast, malicious actors are likely to be less constrained by such factors and may be quicker to deploy harmful AI applications. As Russell (2019: 216) notes, bad actors 'look for—and find—ways to harm others … illegally but undetectably.' This asymmetry means that even if benevolent AI applications

Table 1 Potential outcomes of powerfully aligned AI

Capability Types	Positive	Negative
Medical	P1. Creating advanced AI systems to improve cancer detection and diagnosis (Nassif et al., 2022)	N1. The development of advanced bioweapons using AI technologies (Rubinic et al., 2024)
Educational	P2. Developing AI-powered educational platforms to drastically increase access to quality education in developing countries (Zhai et al., 2021)	N2. Using deepfake technology to generate and spread misinformation at scale, generating confusion and violating rights (Pantserev, 2020)
Assistive	P3. Using AI to design optimized navigation routes for individuals with visual impairments (Chakraborty et al., 2023) P4. Highly capable AI personal assistants to help manage daily tasks and organization more efficiently (Canbek & Mutlu, 2016)	N3. AI systems being vulnerable to adversarial attacks specifically designed to cause errors or malfunctions (Goodfellow et al. 2014)
Social and Emotional	P5. AI companions that can provide emotional support and assist people struggling with mental health issues (Merrill Jr et al., 2022)	N4. Ubiquitous AI surveillance systems eroding individual privacy and civil liberties (Carmody 2021)
Law Enforcement and Security	P6. AI systems for predictive policing that can analyze data to better anticipate and prevent crimes (Berk, 2021)	N5. The proliferation of AI-based scams and cyberattacks (Brundage et al., 2018)
Economic and Labor	P7. AI-driven automation and productivity tools enhancing work efficiency and driving economic growth across various industries (Acemoglu & Restrepo, 2019)	N6. AI automation leading to widespread job displacement and exacerbating economic inequality (Acemoglu & Restrepo, 2019) N7. Manipulating stock markets, commodities, and currencies through automated trading systems that exploit market inefficiencies. (Azzutti, 2022)

have greater potential for positive impact, harmful uses may proliferate more quickly and cause damage before beneficial systems can be fully realized and scaled.[6]

For example, AI systems intended to improve cancer detection and diagnosis (P1) or develop optimized navigation routes for individuals with visual impairments (P3) typically undergo rigorous testing and approval processes before widespread implementation. These systems must navigate complex regulatory landscapes, especially in healthcare and assistive technology sectors. However, harmful AI systems producing outcomes like (N1)–(N5) and (N7) can be more rapidly deployed with less oversight, as malicious actors often operate outside legal and ethical frameworks.[7]

Third, it is generally much easier to cause harm than to sustain beneficial impacts. In the long run, positive AI impacts will be difficult to maintain, while negative consequences will be hard to eliminate. The bioweapons example (N1) illustrates this—once developed, it is extremely challenging to restore previous safety levels, likely requiring an escalating arms race. For an outcome like deepfakes and misinformation (N2), while initial impacts seem manageable, the ability to rapidly spread misinformation means these harms can quickly spiral out of control. In contrast, sustaining positive impacts like AI medical diagnostics (P1) or predictive policing (P6) likely requires continuous development efforts as new challenges emerge over time. For example, new forms of cancer cells may emerge, which can undermine the effectiveness of AI-based diagnoses. Likewise, criminals may take advantage of loopholes in predictive policing.

Of course, it's crucial to acknowledge that language agents and advanced AI systems also offer tremendous potential for positive impact. Increased work efficiency, economic growth, medical breakthroughs, and educational advancements are just a few examples of the profound benefits these technologies could bring. However, while speculative, the above considerations provide sufficient reason to reconsider the prospects of the language agent strategy. The negative impacts tend to be irreversible, easier to propagate initially, and harder to remedy, suggesting that they may outweigh the positive ones in the long run.

5 Objections and Responses

While the risks outlined above are substantial, it is important to consider potential counterarguments. This section examines four objections to the argument from easy alignment and shows that they fail to attenuate the concern regarding malevolent design.

[6] Garfinkel and Dafoe (2019) argue that in the early stages of developing a technology, increased investments tend to favor offensive mechanisms over defensive ones. This could be another factor contributing to the early spread of negative impacts from maliciously aligned AI.

[7] To clarify, I believe legitimate organizations would likely approach AI-powered weapons (N1) with caution and oversight. My point is that the same technology in the hands of malicious actors or fringe groups could be deployed recklessly, without regard for ethical or legal consequences.

5.1 We can Limit Access to Language Agents only to Responsible Users

The first objection contends that employing language models for AI training does not entail granting unrestricted public access to these capabilities. The thought is that even if language agents technically enable easy alignment, strict controls on their use could prevent language agents from posing catastrophic risks. However, there are two reasons to think that it will be difficult, if not impossible, to make language agents accessible only to responsible users.

The first reason is our poor track record at controlling access to transformative technologies over time. As Dung (2024) observes, '[n]ew algorithms are hard to keep secret, even when investing a lot of resources.' A major vulnerability of such algorithms is *hacking*—despite stringent security measures, 'it is a common occurrence for hackers to get access to software projects in progress and to modify or steal their source code' (Yampolskiy, 2016: 144). AI systems face this same risk, as 'an AI system, like any other software, could be hacked and consequently corrupted or otherwise modified to drastically change its behavior' (ibid). Recent arguments by Schneier (2023: 210) echo these concerns succinctly:

> AI systems are computer programs, so there's no reason to believe that they won't be vulnerable to the same hacks to which other computer programs are vulnerable. … If the history of computer hacking is any guide, there will be exploitable vulnerabilities in AI systems for the foreseeable future. AI systems are embedded in … sociotechnical systems …, so there will always be people who want to hack them for their personal gain.

From external hackers stealing source code during development to malicious insiders already having access, it is extremely difficult to guarantee that only responsible parties will gain the access to the language agent technology.

The second reason is our inability to reliably identify good-faith actors who can be trusted to use advanced AI capabilities responsibly over the long-term. Even if rigorous screening is applied initially, malicious actors routinely feign responsibility merely to gain clearance—a behavior Bostrom (2014) terms a 'treacherous turn.' Furthermore, even fundamentally well-intentioned individuals are susceptible to errors in judgment, corruption, or manipulation that could eventually cause them to misuse language agents. For example, a responsible group approved to use language agents may mistakenly end up deploying an AI system trained for nefarious groups falsely portrayed as legitimate. The potential for misinformation and human fallibility complicates any concept of carefully restricting transformative AI technologies only to unwaveringly responsible parties.[8]

In summary, while it might seem feasible to limit access to language agents in theory, the historical precedent and practical realities illustrate how difficult it would

[8] This ties into the point made in footnote 3 that malevolent agents are not the only ones capable of creating harmful AI systems. Even if we manage to grant language agent access only to responsible users initially, it is possible they may inadvertently program harmful goals into the agents due to misinformation or carelessness. And we can expect that completely preventing these factors will be difficult, as they constitute an inseparable aspect of the human condition.

be to prevent their proliferation. The fact that powerful technologies tend to spread beyond initial constraints, coupled with human limitations in assessing long-term reliability, indicates the need for caution.

5.2 We can Develop Measures to Prevent Malicious Alignment

A second objection is that we can build preventive measures into language models in order to block future attempts at malicious alignment. For example, when we ask Chat-GPT to generate inappropriate contents, it is programmed to refuse to follow the instructions. Perhaps, we can further develop this kind of program to prevent malicious agents from using language agents to build harmful AI systems.

I remain skeptical about this possibility for two reasons. First, as emphasized earlier, AI systems are vulnerable to hacking and other kinds of cyber attacks. We cannot expect to develop a way to engrain the preventive measures so deeply into AI systems that they cannot be eliminated or weakened. Bad actors will relentlessly probe for vulnerabilities and exploits. Additionally, language agents themselves could potentially be used as tools to help bypass security measures meant to prevent their misuse.

Second, current preventive measures are easily circumvented, casting doubt on their future efficacy. AI 'jailbreaking' demonstrates this vulnerability. Studies show that AI models can be manipulated into generating harmful content despite safeguards (Shen et al., 2023; Yu et al., 2024). For example, Fowler (2023, February) reports that she initially failed to make ChatGPT write a phishing email, but then easily obtained a convincing tech support note urging her editor to download and install a system update. These vulnerabilities in leading language models suggest that more robust measures will be necessary as the technology evolves.

To be clear, I do not claim that current preventive measures against AI misuse are completely ineffective. For instance, AI systems like DALL-E have safeguards that make it much harder to generate copyrighted or inappropriate images. However, it remains true that (i) such systems could be vulnerable to hacking and manipulation that circumvents those safeguards (Schneier, 2023), and (ii) existing language models like GPT-4 have concerning loopholes that allow prompting them to produce harmful content, which may be difficult issues to fully resolve going forward.

In general, it remains unclear whether we can reliably regulate AI misuse using the limited technical approaches currently conceived. The core issue is that we still lack any clearly promising, comprehensive plans to address the risks of language agents and advanced AI systems being repurposed for harmful ends by malicious actors. This fundamental challenge persists despite proposed stopgap measures, which tend to have shortcomings that motivated adversaries will likely find ways to exploit over time.

5.3 Misaligned AI is Already Dangerous Enough

A third objection argues that AI systems aligned with harmful intentions do not increase risks beyond what we already face from the potential for *mis*aligned AI systems. It is widely recognized that misaligned AI can produce chaotic, unintended, and potentially catastrophic outcomes due to their unpredictability (Bostrom, 2012, 2014; Dung, 2023, 2024; Yampolskiy, 2016, 2019). The canonical example is the paperclip maximizer, which pursues its aim so relentlessly that it begins consuming all available resources, posing an existential threat to humanity.

The key point is that a programmer need not have overtly malevolent aims to create a highly dangerous AI system. Even a seemingly benign objective like maximizing paperclip production could indirectly imperil humanity if the AI system proves incapable of tracking its designer's genuine intentions in a reliable way. From this perspective, misaligned AI systems following innocent orders are already a severe risk, so additional concerns about purposefully malicious AI may be misplaced.

However, this objection overlooks some crucial differences between misaligned and maliciously aligned systems. It is true that mistakes or errors from misaligned AI have sometimes produced harmful results. Yet, unlike the outcomes of malicious intentions, chaotic results are often relatively harmless or benign. Yampolskiy (2019) provides examples such as an AI designed for writing Christmas carols instead generating nonsensical outputs, translation software comically failing to properly render a name, a virtual assistant unhelpfully responding about cheese locations, and image generation creating bizarrely merged photos. Aside from these real cases, one can easily conceptualize other seemingly innocuous failures, such as a joke-writing AI occasionally producing attempts at humor that simply miss the mark.

By contrast, AI systems purposefully aligned to destructive ends would almost always constitute a substantial danger. As Friederich (2023) argues, a fully aligned AI system, by definition, will reliably fulfill the exact goals of its operators. Thus, if the operators have malevolent goals, the AI will most certainly try to produce harmful outcomes in line with the goals. In addition to reliably producing harmful results, aligned AI could be dangerous due to the wide range of harms it can bring about. For example, Yampolskiy concludes that 'the most important problem in AI Safety is intentional-malevolent-design resulting in artificial evil AI' (2016: 144) on the grounds that it involves all other safety risks induced by AI, such as negative outcomes caused by reward hacking.

Furthermore, the development of language models that can streamline AI training, or simply the perception that such agents are developed, is likely to incentivize malicious individuals to attempt constructing harmful AI systems. Until now, the (perceived) difficulty of creating highly capable AI systems has limited their proliferation mainly to coding experts. However, if language agents significantly reduce these barriers to entry, it follows that the risk will escalate as malicious or radically misguided groups gain the ability, or at least the motivation, to develop powerful AI devoted to their harmful agendas.

5.4 Malicious Alignment is Already as Harmful as it can be

A fourth objection is that language agents will primarily improve our ability to build benevolent systems because there will be a limit to facilitating malicious alignment. Malicious AI designers are already engaging in propagating misinformation, attacking infrastructure, and other harmful projects (Schneier, 2023; Verma 2023). Of course, it might be difficult for them to get the AI to do exactly what they want, but they are not, independent of language agents, *totally hopeless* at alignment. If this is right, one might think that the main thing language agents do is mitigate risks associated with intended benevolent uses, while not significantly increasing the most serious risks posed by AI.

However, given the aim of the language agent strategy, it is more plausible to think that it will significantly contribute to facilitating malicious alignment. Recall that this strategy aims to facilitate alignment by lowering the barrier to entry for building aligned systems. Most prominently, it will benefit agents whose programming knowledge is insufficient to build AI systems on their own. The general public is already generating harms using language agents, and this concerning trend is growing rapidly. For example, threats like using deepfake technology to generate and spread misinformation at scale are prevalent in our society (Öhman, 2020; Verma, 2023, December).

Admittedly, it is less clear whether powerful agents like multinational corporations face significant barriers to alignment. It might be the case that they can already generate a wide range of outcomes, positive or negative. But do we have good reason to believe they will not benefit much from language agents? I doubt it. The problems Goldstein and Kirk-Giannini (2023) aim to address are quite general. Goal misgeneralization, for example, is a challenge that can arise in any AI training process. Even with extremely large datasets, ensuring the system behaves in line with the programmer's intentions in novel situations can be difficult.

In this context, it is worth emphasizing the interpretability of language agents. Even for immensely powerful agents, an interpretable AI system can provide a safer means to realize their aims. For example, suppose a powerful organization wants to build bioweapons using AI. Even if their system is capable enough, they might hesitate due to the possibility of unintended errors. However, since one can directly read off the 'thoughts' of language agents, the perceived stakes could be lowered for the malicious agent. Even if the stakes are not actually lowered, the mere belief that they could motivate malicious actors to proceed with their plans, which would be dangerous regardless of success.

6 Concluding Remarks

Recall Bostrom's easy nuke scenario: how would things have unfolded if creating nuclear weapons had been easy? As Bostrom argues, it would be extremely difficult to effectively ban creating nukes. Even if we managed to gather enough political support for a ban, we would face numerous practical hurdles: shutting down all university physics departments and implementing extensive security measures, among other things. As Bostrom says, 'we were lucky that making nukes turned out to be hard (2019: 457)'.

In this paper, I argued that the language agent strategy is likely to drive us into a similar situation as Bostrom's easy nuke scenario. Given the various incentives for malevolent actors, preventing the creation of harmful AI systems using language agents will be extremely difficult. Moreover, current preventive measures are inadequate to ensure a safe future for humanity, and there is no guarantee that future technologies will make things better in this regard. These considerations suggest that developing language agents might not be the best course of action.

Where does our discussion leave us? If, as I have argued, language agents increase the likelihood of harmful AI systems being developed, does it necessarily follow that we should entirely avoid using them? Too rash a takeaway. Instead, I want to draw two broader lessons from our discussion.

The first is a practical lesson—we need distinct tools and strategies to address risks stemming from human agents versus those arising from the AI systems themselves. While intimately related, these two risks demand different technical and regulatory approaches. For example, we need one type of solution to prevent an AI system like GPT-4 from perpetuating racial stereotypes or other problematic biases ingrained through its training data. But we need a distinct framework to ensure that GPT-4 refrains from generating racist jokes, hate speech or other malicious content upon direct request from a human user. Addressing the latter issue becomes highly relevant in a scenario where effectively aligned artificial agents stand prepared to serve any malicious goals of their operators.[9]

Secondly, the overarching theoretical lesson is that we need to consider multiple intersecting AI risk factors in conjunction. These include challenges like the alignment problem (Bostrom, 2012, 2014; Dung, 2023, 2024), risks from human misuse (Brundage et al., 2018; King et al., 2020; O’Neil, 2016; Yampolskiy, 2016), technical fragilities (Crouch, 2023, February), environmental concerns (Rillig et al., 2023), and more. Only by addressing these various aspects together can we develop a robust, effective strategy for creating transformative AI systems that benefit humanity.

While the risks of maliciously designed AI systems are acknowledged, they have received relatively little research attention compared to other AI safety issues like the alignment problem (Hagendorff, 2020: 107). Even those who recognize the dangers of malevolent AI design tend to treat it as a somewhat isolated concern, rather than examining how it interconnects with other potential AI risks. For example, Yampolskiy (2016) highlights malevolent design as a significant threat, but focuses more on arguing its primacy over other topics getting more airtime, such as the alignment problem. While important, this still falls short of fully grappling with the reality that various AI-related risks are intricately intertwined.

Now, is there a promising way forward once we address the various interrelated issues? Or are there simply painful tradeoffs, rendering language agents inevitably dangerous? While there are no easy answers, I believe it’s too early to lose all hope. Throughout history, humanity has confronted and solved or mitigated complex, multifaceted challenges without catastrophic failure. Moreover, AI safety research is still nascent; we have only recently begun grappling with alignment problems and associated risks. Just as with other formidable issues, it’s possible to mitigate the alignment problem without catastrophically exacerbating risks, though difficult challenges

[9] One might object that this suggested distinction undermines the argument from easy alignment. If we can implement robust measures to block AI systems from fulfilling intentionally malicious objectives, doesn’t that simply solve the problem? Not quite. While we do have some technical tools to prevent human-induced AI harms, they are extremely limited and woefully inadequate as comprehensive safeguards, at least in the status quo. As argued in Sect. 5.2, it is still relatively easy for motivated actors to circumvent current preventative measures.

likely await. Hence, we should maintain diligent theoretical and practical efforts toward solutions instead of resigning AI as inevitably disastrous.

Acknowledgements I thank two anonymous reviewers for *Philosophy & Technology* and participants at the 2024 Value Theory seminar at the Ohio State University for their valuable feedback. Special gratitude goes to Tristram McPherson, whose insightful comments greatly enhanced this paper.

Author contribution The paper is the sole work of Inchul Yum, who conceptualized, researched, and wrote the entire paper.

Funding No external funding sources were utilized in the preparation of this manuscript.

Data Availability Not applicable.

Declarations

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

Competing Interests The author declares no competing financial or non-financial interests.

References

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, *33*(2), 3–30.

Amodei, D., & Clark, J. (2016). Faulty reward functions in the wild, Retrieved from https://openai.com/research/faulty-reward-functions

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *ArXiv*, *1606.06565*. https://doi.org/10.48550/arXiv.1606.06565

Anderson, K., & Waxman, M. C. (2013). Law and ethics for autonomous weapon systems: Why a ban won't work and how the laws of WAR can. *SSRN Journal*, 1–32.

Arredondo, P. (2023, April). GPT-4 passes the bar exam: What that means for artificial intelligence tools in the legal profession, *Stanford Law School Blogs*, Retrieved from https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/

Azzutti, A. (2022). AI-driven market manipulation and limits of the EU law enforcement regime to credible deterrence. *ILE Working Paper Series*, *54*. https://doi.org/10.2139/ssrn.4026468

Bales, A., D'Alessandro, W., & Kirk-Giannini, C. D. (2024). Artificial intelligence: Arguments for catastrophic risk. *Philosophy Compass*, *19*(2), e12964. https://doi.org/10.1111/phc3.12964

Bendel, O. (2017). The synthetization of human voices. *AI & Society*, *82*, 737.

Berk, R. A. (2021). Artificial intelligence, predictive policing, and risk assessment for law enforcement. *Annual Review of Criminology*, *4*, 209–237.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds & Machines*, *22*, 71–85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455–476. https://doi.org/10.1111/1758-5899.12718

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *ArXiv*, 180207228. https://doi.org/10.48550/arXiv.1802.07228

Carmody, J., Shringarpure, S., & Van de Venter. (2021). AI and privacy concerns: a smart meter case study. *Journal of Information, Communication and Ethics in Society, 19*(4), 492–505.

Canbek, N. G., & Mutlu, M. E. (2016). On the track of artificial intelligence: Learning with intelligent personal assistants. *Journal of Human Sciences*, *13*(1), 592–601.

Cave, S., & ÓhÉigeartaigh, S. S. (2018, December). An AI race for strategic advantage: Rhetoric and risks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 36–40).

Chakraborty, N., Mishra, Y., Bhattacharya, R., & Bhattacharya, B. (2023). Artificial intelligence: The road ahead for the accessibility of persons with disability. *Materials Today: Proceedings*, *80*, 3757–3761.

Chalmers, D. J. (2016). The singularity: A philosophical analysis. *Science fiction and philosophy: From time travel to superintelligence*, 171–224.

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, *107*, 1753–1820.

Crouch, G. (2023, February). The fragility of artificial intelligence, Retrieved from https://gilescrouch.medium.com/the-fragility-of-artificial-intelligence-1b319c8f0145

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese*, *202*(138). https://doi.org/10.1007/s11229-023-04367-0

Dung, L. (2024). The argument for near-term human disempowerment through AI. *AI & Society*. https://doi.org/10.1007/s00146-024-01930-2

Engelmann, S., Chen, M., Fischer, F., Kao, C., & Grossklags, J. (2019). Clear sanctions, vague rewards: How China's social credit system currently defines Good and Bad behavior. In *Proceedings of the conference on fairness, accountability, and transparency—FAT* '19* (pp. 69–78).

Ernest, N., Carroll, D., Schumacher, C., Clark, M., Cohen, K., & Lee, G. (2016). Genetic fuzzy based artificial intelligence for unmanned combat aerial vehicle control in simulated air combat missions. *Journal of Defense Management*, *6*(1). https://doi.org/10.4172/2167-0374.1000144

Fowler, B. (2023, February). It's scary easy to use chatGpt to write phishing emails, Retrieved from https://www.cnet.com/tech/services-and-software/its-scary-easy-to-use-chatgpt-to-write-phishing-emails

Friederich, S. (2023). Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00268-7

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds & Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Garfinkel, B., & Dafoe, A. (2019). How does the offense-defense balance scale? *Journal of Strategic Studies*, *42*(6), 736–763. https://doi.org/10.1080/01402390.2019.1631810

Goldstein, S., & Kirk-Giannini, C. D. (2023). Language agents reduce the risk of existential catastrophe. *AI & Society*, 1–11. https://doi.org/10.1007/s00146-023-01748-4

Göring, S., Rao, R. R. R., Merten, R., & Raake, A. (2023). Analysis of appeal for realistic AI-generated photos. *Ieee Access: Practical Innovations, Open Solutions*. https://doi.org/10.1109/ACCESS.2023.3267968

Ha, D. (2019). Reinforcement learning for improving agent design. *Artificial Life*, *25*(4), 352–365. https://doi.org/10.1162/artl_a_00301

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds & Machines*, *30*(1), 99–120.

Helbing, D. (Ed.). (2019). *Towards digital enlightenment: Essays on the dark and light sides of the digital revolution*. Springer.

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*, *1*(3), 36–57. https://doi.org/10.15781/T2639KP49

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics*, *26*, 89–120.

Kosinski, M., & Wang, Y. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543–556.

Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., & Krueger, D. (2022, June). Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning* (pp. 12004–12019). PMLR.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Longpre, S., Storm, M., & Shah, R. (2022). Lethal autonomous weapons systems & artificial intelligence: Trends, challenges, and policies. *MIT Science Policy Review*, *3*(1), 47–56.

Lyon, D. (2003). Surveillance as social sorting: Computer codes and mobile bodies. In D. Lyon (Ed.), *Surveillance as social sorting: Privacy, risk, and digital discrimination* (pp. 13–30). Routledge.

Marijan, B. (2022). Autonomous weapons: The false promise of civilian protection, Retrieved from https://www.cigionline.org/articles/autonomous-weapons-the-false-promise-of-civilian-protection/

MerrillJr, K., Kim, J., & Collins, C. (2022). AI companions for lonely individuals and the role of social presence. *Communication Research Reports*, *39*(2), 93–103.

Metz, C. (2016, March 16). In two moves, AlphaGo and Lee Sedol redefined the future, *Wired*, Retrieved from https://www.wired.com/2016/03/two-moves-alphago-lee-sedolredefined-future

Nassif, A. B., Talib, M. A., Nasir, Q., Afadar, Y., & Elgendy, O. (2022). Breast cancer detection using artificial intelligence techniques: A systematic literature review. *Artificial Intelligence in Medicine*, *127*, 102276.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Öhman, C. (2020). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*, *22*(2), 133–140.

Pantserev, K. A. (2020). The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In H. Jahankhani, S. Kendzierskyj, N. Chelvachandran, & J. Ibarra (Eds.), *Cyber Defence in the age of AI, Smart societies and Augmented Humanity*. Springer. Advanced Sciences and Technologies for Security Applications https://doi.org/10.1007/978-3-030-35746-7_3

Pistono, F., & Yampolskiy, R. V. (2016). Unethical research: How to create a malevolent artificial intelligence. *ArXiv*. https://doi.org/10.48550/arXiv.1605.02817. abs/1605.02817.

Popov, I., Heess, N., Lillicrap, T., Hafner, R., Barth-Maron, G., Vecerik, M., Lampe, T., Tassa, Y., Erez, T., & Riedmiller, M. (2017). Data-efficient deep reinforcement learning for dexterous manipulation. *ArXiv*. https://doi.org/10.48550/arXiv.1704.03073. abs/1704.03073.

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.

Rubinic, I., Kurtov, M., Rubinic, I., Likic, R., Dargan, P. I., & Wood, D. M. (2024). Artificial intelligence in clinical pharmacology: A case study and scoping review of large language models and bioweapon potential. *British Journal of Clinical Pharmacology*, *90*(3), 620–628.

Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. Penguin UK.

Schneier, B. (2023). *A Hacker's mind: How the powerful Bend Society's rules, and how to Bend them back*. W. W. Norton & Company.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). Goal misgeneralization: Why correct specifications aren't enough for correct goals. *ArXiv*, *2210.01790*. https://doi.org/10.48550/arXiv.2210.01790

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2023). Do anything now: Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *ArXiv*, *2308.03825*. https://doi.org/10.48550/arXiv.2308.03825

Skalse, J., Howe, N., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. *ArXiv*. https://doi.org/10.48550/arXiv.2209.13085. 2209.13085.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

Verma, P. (2023, December). The rise of AI fake news is creating a 'misinformation superspreader,' *The Washington Post*, Retrieved from https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229).

Yampolskiy, R. V. (2016). Taxonomy of pathways to dangerous artificial intelligence. *AAAI Workshop: AI Ethics and Society*, 143–148.

Yampolskiy, R. V. (2019). Predicting future AI failures from historic examples. *Foresight*, *21*(1), 138–152. https://doi.org/10.1108/FS-04-2018-0034

Yu, Z., Liu, X., Liang, S., Cameron, Z., Xiao, C., & Zhang, N. (2024). Don't listen to me: Understanding and exploring jailbreak prompts of large Language models. *ArXiv*, *2403.17336*. https://doi.org/10.48550/arXiv.2403.17336

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., & Li, Y. (2021). A review of Artificial Intelligence (AI) in education from 2010 to 2020. *Complexity*, *2021*, 1–18.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.