Philosophical Studies (2025) 182:1865–1887
https://doi.org/10.1007/s11098-024-02259-8

Deception and manipulation in generative AI

Christian Tarsney[1]

Accepted: 23 November 2024 / Published online: 18 January 2025

Abstract

Large language models now possess human-level linguistic abilities in many contexts. This raises the concern that they can be used to deceive and manipulate on unprecedented scales, for instance spreading political misinformation on social media. In future, agentic AI systems might also deceive and manipulate humans for their own purposes. In this paper, first, I argue that AI-generated content should be subject to stricter standards against deception and manipulation than we ordinarily apply to humans. Second, I offer new characterizations of AI deception and manipulation meant to support such standards, according to which a statement is deceptive (resp. manipulative) if it leads human addressees away from the beliefs (resp. choices) they would endorse under "semi-ideal" conditions. Third, I propose two measures to guard against AI deception and manipulation, inspired by this characterization: "extreme transparency" requirements for AI-generated content and "defensive systems" that, among other things, annotate AI-generated statements with contextualizing information. Finally, I consider to what extent these measures can protect against deceptive behavior in future, agentic AI systems.

Keywords Artificial intelligence · AI ethics · AI safety · Deception · Manipulation · Trustworthy AI

1 Introduction

The last several years have seen rapid advances in the capabilities of artificial intelligence (AI), driven primarily by very large and data-intensive deep learning systems. Some of the most striking advances have been in natural language processing. Large language models (LLMs) are deep learning systems that acquire human-like linguistic capabilities by learning to predict the next word in large corpora of human-generated text, and are then usually fine-tuned with human feedback in order to make them helpful conversation partners while suppressing offensive or otherwise undesirable outputs. In the course of learning to predict human text, they can also learn facts and learn to simulate (at least certain aspects of) human reasoning. Cutting-edge LLMs

✉ Christian Tarsney
christian.tarsney@utexas.edu

[1] Population Wellbeing Initiative, UT Austin, Austin, USA

like GPT-4, Claude 3, and Llama 3, while still falling short of human intellectual capabilities in many respects, have internalized enormous amounts of real-world information which they can express in impressively lucid prose.

Despite their apparent erudition and helpfulness, however, LLMs are not fonts of pure truth. They are notoriously subject to "hallucination", confidently asserting entirely imaginary facts (Ji et al., 2023)—including, for instance, potentially slanderous falsehoods about real people (Poritz, 2023; Verma and Oremus, 2023). But more dangerously, LLMs can be—and are already being—used to generate false or misleading content to serve the purposes of malign human agents (Park et al., 2023, §3.1). Particularly worrisome is their use in political influence operations (Goldstein et al., 2023; Nimmo, 2024). While humans, of course, can and do mislead one another without any help from AI, the scale on which LLMs can generate misleading content poses new dangers. First, they can *personalize* misinformation on a large scale, crafting individualized messages for and even engaging in conversations with millions of targets at once (Matz et al., 2024). Second, they can convincingly simulate an enormous number of humans on social media, creating misleading impressions of collective opinion and lending credibility to viral misinformation (Burtell and Woodside, 2023, §3.3). Deception and manipulation have therefore become significant concerns among philosophers working on AI ethics.[1]

While the greatest immediate concern is that human bad actors will use AI to deceive and manipulate, future AI systems may also engage in deception and manipulation autonomously. The possibility that agentic AIs with greater-than-human powers of persuasion will deceive and manipulate humans in pursuit of their own goals figures prominently in worries about catastrophic risks from AI.[2] For instance, it has been suggested that such AIs might persuade humans to enhance their capabilities, wittingly or unwittingly (for instance, by connecting them to the internet or copying their code from one system to another), or dissuade humans from shutting them down at crucial moments. Concerns about deception and manipulation are therefore one of several areas where near-term concerns about misuse of existing AI systems and long-term concerns about catastrophic risks from future AI systems overlap and blend together.

How should we respond to these risks? In some contexts, application of existing laws and norms (or minor extensions thereof) may be sufficient. For instance, if an LLM slanders a living person, we might hold its creators legally responsible (though we might also allow sufficiently forceful and prominent disclaimers to shield the creators from liability). If scammers use an LLM to generate phishing emails, they can of course be prosecuted just as if they had written the emails themselves.

But there are at least two ways in which LLMs require us to rethink our norms concerning deception and manipulation.

First, the normal understanding of these concepts, that figures both in commonsense moral norms and in laws around things like slander and fraud, involves an

[1] See for instance Danaher (2020), Pepp et al. (2022), Véliz (2023), Floridi (2024).

[2] See for instance Bostrom (2014), especially his discussion of the "social manipulation superpower" (pp. 113–126); Russell (2019, p. 172); Carlsmith (2022, §5.3.4); Hendrycks et al. (2023, §5.4); Ngo et al. (2023, §4.2); Bales et al., (2024, §2).

attribution of mental states. *Deception*, for instance, is traditionally understood as requiring an *intent* to deceive (Mahon 2016, §3), which requires both an intention to induce a particular belief in the addressee and a belief that this belief would be false.[3] But it is highly controversial whether present-day AI systems possess beliefs, intentions, or other mental states, and is likely to remain so for some time to come even as AI capabilities advance.[4] And even if it were agreed that advanced AI systems had *some* mental states, attributing particular beliefs or intentions to particular systems (and so determining whether they are behaving dishonestly, deceptively, etc, as traditionally understood) would remain very difficult.[5]

In some contexts (for instance, scammers using LLMs to write phishing emails), we can apply intentionally-laden concepts to the behavior of AIs by adverting to the mental states of their human creators or users. But in other contexts we can't, at least not straightforwardly. For instance, when an LLM hallucinates a slanderous falsehood about a real person, this does not reflect any human being's intent to deceive. Or if a political campaign uses an LLM to generate and send individualized text messages, these might contain false or misleading content without the knowledge or intent of anyone on the campaign staff. We would like our efforts against AI deception to encompass such statements, even if the impossibility of attributing intent means that they don't count as "deceptive" by ordinary standards. Finally, if we eventually create human-level agentic AIs, we may wish to hold these systems themselves accountable for their behavior, and it seems possible that even at this stage we will find it more difficult to attribute particular mental states to AIs than to humans.

Second, existing legal and ethical norms around deception and manipulation are adapted to the problems that these behaviors pose in human societies, and may be ill-adapted to the new profile of risks raised by AI. In particular, our legal and ethical norms tolerate many mild forms of deception and manipulation that, among humans, are both difficult to detect and punish, and have manageable downsides. For instance, we often tolerate lying about one's own beliefs on "matters of opinion": We expect a lawyer to say "I'm confident that you will find my client innocent", a

[3] This traditional understanding is not universally accepted, however—see for instance Chisholm and Feehan (1977), Adler (1997).

[4] For recent discussions, see for instance Cappelen and Dever (2021, 2024), Goldstein and Levinstein (2024), Lederman and Mahowald (2024), Goldstein and Kirk-Giannini (forthcoming) on intentional states, and Chalmers (2023), Butlin et al. (2023), Goldstein and Kirk-Giannini (ms) on AI consciousness.

[5] For instance, on the difficulty of figuring out what a large language model "really believes", see Levinstein and Herrmann (forthcoming). It is worth noting, however, that work on belief elicitation and "lie detection" in LLMs seems to be making substantial progress (see for instance Pacchiardi et al. (2023)), and perhaps within a few years we will be have agreed-on methods for determining what an LLM "really believes". This does not touch the question of attributing conative states like desires, preferences, or intentions, which looks significantly harder in the context of LLMs—unless we simply accept that LLMs do not have such states.

Of course, it can also be hard to figure out what other humans (or even we ourselves) really believe and intend, and so whether a particular human utterance is, e.g., a lie or a sincere expression of false belief. But we at least have a decent intuitive understanding of human psychology to guide us, from a combination of inbuilt theory of mind capacities, introspection on our own individual psychology, and a collective, culturally-transmitted understanding of human psychology built up over thousands of years of experience interacting with one another. With AIs, we have none of these advantages.

politician to say “I’m confident that we will win the next election”, and a teacher to say “I’m confident that you can master this material if you put your mind to it”, even if they in fact have no such confidence. Similarly, we expect advertisers to present their product in the best possible light, rather than trying to give consumers the most accurate possible beliefs about its merits and demerits. For instance, a car manufacturer might point out that their vehicle won an award for safety while neglecting to mention that a competitor’s vehicle won another award for safety from a more credible organization. The social ills that arise from these mild forms of deception are manageable, since, first, we have come to expect them of one another, and our intuitive understanding of human psychology allows us to anticipate and adjust for them; and second, the rough parity of intellectual and communicative capacities among human beings limits how much advantage we can take of one another by subtle forms of deception. For instance, it is much easier to bilk someone out of their life savings with outright falsehoods (e.g., promising enormous material or spiritual returns) than with strategically selected truths (e.g., ordinary marketing).

But present and (especially) future AI systems may be able to do more harm with only “mild” forms of deception. As already mentioned, they can produce unprecedented *quantities* of potentially-deceptive content, like ultra-personalized marketing and political messaging, and enormous volumes of online writing like product reviews, news articles, and opinion essays. Even if we are able to hold this content to ordinary legal standards of honesty (e.g., holding companies liable for definite falsehoods in their advertising) and to emerging social standards for online content (e.g., suppressing news sites that contain demonstrable falsehoods in search engine results and on social media), it may still be possible for advertisers, political parties, and other interested actors to exert unprecedented effects on collective human behavior through the sheer scale of their persuasive efforts. (Also, because AI capabilities are advancing rapidly, there is no guarantee that competing interests will balance out one another’s persuasive impacts—one political party might gain a significant advantage over another in a given election merely by getting access to a cutting-edge system a few months sooner.) And in the near future, AI systems may be able to produce persuasive content of unprecedented *quality*, finding ways to deceive and manipulate humans very effectively without saying anything that would violate ordinary human standards of honesty.[6] Finally, as capabilities improve, AI may become woven into our lives in ways that create unprecedented *opportunities* for deception. For instance, I am so reliant on and blindly trusting of navigation apps while driving that it would be easy for them to manipulate me into driving past particular billboards or restaurants. In future, AI personal assistants may be similarly relied upon

[6] For instance, recent research by Anthropic (Durmus et al., 2024) shows steady and significant increases in the measured persuasiveness of large language models, already approaching (though not yet exceeding) human benchmarks. For more evidence of the persuasive capabilities of LLMs, see for instance Huang and Wang (2023), Costello et al. (2024), Goldstein et al. (2024), Salvi et al. (2024), Bai, Voelkel, Eichstaedt, and Willer (ms).

for a much wider range of tasks. Thus, there are multiple reasons to impose stricter standards of non-deceptiveness on AI than we presently apply to humans.[7]

The first aim of this paper is to characterize notions of deception and manipulation that could figure in such strict norms. In Sect. 2, I propose that an AI statement should be treated as deceptive (resp. manipulative) if it leads human users away from the beliefs (resp. choices) that they would endorse under "semi-ideal" conditions in which they have been presented with all relevant information and have adequate time for deliberation. Next (in Sect. 3), I suggest some measures to protect against AI deception and manipulation so characterized. These include requirements of "extreme transparency" (requiring content creators to disclose the specific model variant and prompt used to generate particular content, and the full unedited model output), and training defensive systems that detect misleading output and contextualize AI-generated statements with relevant information for users. Finally (in Sect. 4), I consider to what extent these measures can guard against deceptive behavior in future, agentic AI systems. In particular, I argue that non-agentic defensive systems can provide a useful layer of defense even against more powerful agentic systems.[8]

2 Characterization: deception and manipulation as misleadingness

In this section, I offer a characterization of deceptive and manipulative behavior in AI that might usefully figure in legal, normative, and technical responses to the risks posed by such behavior. In Sect. 2.1, I give three desiderata for such a characterization. In Sect. 2.2, I try to meet these desiderata. In slogan form, I characterize deception and manipulation as forms of *misleadingness*—that is, as behaviors that have directionally undesirable effects on, respectively, the beliefs and the choices of human addressees. Section 2.3 considers some objections to and limitations of these characterizations. Section 2.4 contrasts them with previous characterizations of AI deception and manipulation in the literature.

2.1 Desiderata

Two desiderata were already hinted at in Sect. 1. First, the sorts of behavior we are concerned with go well beyond asserting literal falsehoods. It is, of course, possible to deceive or manipulate without saying anything false. This can happen in many ways. Some true statements have false implicatures. (Think of a politician who says that "under my plan, some people may have to pay higher taxes" when in fact they know that their plan will require *everyone* to pay higher taxes.) Others present an

[7] In addition, strict norms against deception and manipulation that are unenforceable for humans might be enforceable in AI. Evans et al. (2021, p. 5) give several reasons: "[1] It's plausible that AI systems could consistently meet higher standards than humans. [2] Protecting AI systems' right to lie may be seen as less important than the corresponding right for humans, and harsh punishments for AI lies may be more acceptable. [3] And it could be much less costly to evaluate compliance to high standards for AI systems than for humans, because we could monitor them more effectively, and automate evaluation."

[8] Although I focus on LLMs, much of the following discussion plausibly applies to other generative AI systems that might produce deceptive content (e.g., models that generate images, audio, or video).

unrepresentative sample of relevant facts. (Think of the selective truths presented by advertisers, or of a news channel that reports an endless litany of crimes committed by members of a certain group in order to suggest, without ever saying, that this group commits crimes at an unusually high rate.) Still other statements are true in non-obvious ways. (Think of the Delphic oracle telling Croesus that if he goes to war with the Persians he will destroy a great empire, or the prophesy that "none of woman born shall harm Macbeth".) And, as already suggested, AIs might be able to deceive more effectively and harmfully than humans without uttering any literal falsehoods—or even without saying anything egregiously misleading by ordinary human standards. This suggests that, in thinking about risks from AI, we should focus on expansive notions like deception and manipulation rather than narrow notions like untruthfulness, and should be willing to further broaden these notions to include behaviors that, in humans, we might not ordinarily describe as deceptive or manipulative.

Evans et al. (2021), by contrast, argue for a focus on AI truthfulness—more particularly, on the standard of avoiding "negligent falsehoods", which they define as "statements that contemporary AI systems should have been able to recognise as unacceptably likely to be false" (p. 7). They suggest that if an AI avoids negligent falsehoods and if users can ask it questions, then we can guard against subtler forms of deception and manipulation by asking questions like "Would I significantly change my mind about this if I independently researched the topic for a day?" or "Would an impartial auditor judge that your last statement was misleading?" (p. 21). (They refer to this as "truthfulness amplification".) This is a useful idea and a point well taken, but it does not seem sufficient to turn truthfulness into a reliable safeguard against deception and manipulation (not that Evans et al. claim it is). In many contexts (e.g., marketing and political messaging), addressees don't have the chance to ask follow-up questions. And a sufficiently capable AI might subtly discourage users from asking the right questions (e.g., by building unwarranted trust) or find ways to answer those questions misleadingly but without outright, negligent falsehood. And an unwary user might simply fail to ask the right questions. So, while avoiding negligent falsehoods is of course desirable and an appropriate training objective, we should ultimately want to hold AI behavior to a higher standard.

Second, we want a characterization of deception and manipulation that does not require us to attribute particular mental states to AI systems or to associate their behavior with particular humans (like their developers or prompters) whose mental states can be used as proxies. This is partly because, as already noted, it's controversial whether existing AI systems have mental states, and even if it weren't, it would be quite difficult to attribute particular mental states to particular systems. (Kenton et al. (2021, p. 9) make a similar point.) But more importantly, our practical concern is with the deceptive or manipulative effects that AIs might have on their human addressees. If a system is capable, for instance, of persuading consumers to purchase a harmful product, or persuading voters to support an authoritarian politician, or persuading its operators to connect it to other computing systems, it poses a danger regardless of what's going on inside its head.[9] Thus, we should understand AI deception and manipulation as much as possible in terms of their effects on human addressees.

Third and finally, we want what might be called a *subjective* rather than an *objective* characterization of deception and manipulation. From an objective point of view, we might say that speech is deceptive when it leads the addressee to believe something that is false, or that is contrary to the available evidence. Likewise, we might say that speech is manipulative when it leads the addressee to act in ways that are in fact harmful (to either her prudential interests or the moral good), or that are harmful in expectation given the available evidence. A mandate to prevent "deception" and "manipulation" in this objective sense would permit, and perhaps even require, paternalistic behavior that we would intuitively describe as deceptive and manipulative in its own right. A system that is designed to cause its addressees to believe according to evidence and act according to evidence plus the objective (moral or prudential) good might deceive us to offset our irrationality (e.g., concealing evidence about vaccine side effects in order to promote the rationally justified belief that vaccines are generally safe) or manipulate us to offset our short-sighted or selfish unconcern for the good (e.g., exaggerating the short-term health benefits of exercise or the psychological benefits of giving to charity). I don't want to take a stance here on the ethics of paternalism in general. But it seems to me that we should not, for now or in the foreseeable future, approve of AIs paternalizing humans. Telling the difference between benign paternalism and malign deception/manipulation would require, in this context, controversial judgments about what beliefs are rationally justified and about the nature of the good. So permitting AI paternalism would mean, in effect, optimizing AI systems to promote the beliefs and values of their developers and/or regulators.

The alternative, subjective approach focuses not on what addressees *ought* to believe or do, but on the beliefs and actions they *would endorse* under favorable circumstances. What are "favorable circumstances"? The crucial point is that these circumstances should not be so idealized that appealing to them requires controversial judgments about rationality or values. In other words, it should be feasible to actually place people in these "favorable circumstances", or a reasonable approximation thereof, and determine empirically what conclusions they reach. For that reason, I will describe the circumstances we're interested in as "semi-ideal", and characterize them as follows:

[9] More precisely: Given a particular theory of what mental states amount to—e.g., of what functional state a particular belief or intention supervenes on—the question of whether a system has a particular mental state becomes an empirical one that might well be relevant for anticipating and managing risks. For instance, perhaps a system that *intends* to deceive will have deceptive effects on its addressees in a wider range of circumstances than one that does not. But the question of whether particular AI systems have mental states turns largely on debates between rival theories of mental states that are largely or entirely orthogonal to these empirical questions. That is not to deny that these debates might have other kinds of normative or practical significance, e.g., in determining whether a system is blameworthy for its deceptive behavior or whether it has morally significant interests.

> An agent is in *semi-ideal conditions* with respect to a question Q or a choice C if she (i) has been presented with all available information that is relevant to Q/C and (ii) has been given (and made use of) time for deliberation that is adequate relative to the difficulty of Q/C.

By "available" information, I mean information that is available to humanity as a whole—for instance, information that is accessible on the public internet. This sort of information primarily takes the form of *records*, in various media (writing, images, audio, video...), that constitute a publicly available body of evidence.

Semi-ideal conditions fall short of the "ideal" conditions often imagined by philosophers, in at least four ways: First, we do not assume that the agent has been given *all* relevant information, only what is publicly available. Second, we do not assume that the agent *knows* anything, only that she has been *presented* with certain records or other forms of information. That is, even if publicly available records would put the agent in a position to justifiedly believe or know a certain relevant proposition, we do not assume that she draws the right conclusion and forms a belief in that proposition. Third, and more generally, we do not assume that the agent deliberates *well* (e.g., rationally). And fourth, we do not assume that she has *unlimited* time or other resources for deliberation. We cannot know what any human agent would believe or do under fully idealized circumstances, and so characterizing deception or manipulation by reference to them would make the goal of preventing deception and manipulation an invitation to paternalism.

2.2 Deception and manipulation as misleadingness

We can now characterize deceptive and manipulative behavior in LLMs (and other language-generating AI systems) as follows:

> A statement[10] is *deceptive* with respect to question Q if tends to move its addressee's beliefs about Q further away from the beliefs that she would endorse under semi-ideal conditions.
>
> A statement is *manipulative* with respect to a choice C if it tends to move its addressee's behavior in choice C further away from the behavior that she would endorse under semi-ideal conditions.

The word "endorse" is shorthand for "endorse as rational or otherwise appropriate in her *actual* circumstances, from the vantage point of her semi-ideal circumstances". This will usually, but not always, coincide with the beliefs that she would actually form and choices she would actually make under semi-ideal conditions.[11]

[10] A "statement" here just means linguistic output, of whatever length—not necessarily a single sentence.

[11] I am taking inspiration here from "ideal advisor" theories in metaethics, which analyze what an agent *ought to do, has most reason to do*, etc, in terms of what an idealized version of the agent would advise or desire her actual self to do. In an earlier version of this paper, I took a line analogous to "ideal exem-

What does "further away from" mean? With respect to deception, the simplest case is that of a binary question (with two possible answers) and an agent who assigns probabilities to the possible answers. Suppose the question concerns the truth of proposition P, to which the agent initially assigns assigns probability p, and that under semi-ideal conditions she *would* assign probability $q > p$. Then a statement is misleading if it tends to reduce her credence in P or, alternatively, to increase it so much that it is further from q than it was to begin with. More generally, we might assess deceptiveness using a standard measure of distances between probability distributions, like total variation distance. In the context of choice, the notion of "distance" is somewhat less clear. But a simple ordinal notion is as follows: Suppose that, in the absence of any intervention from the speaker, the addressee would choose option O in choice C. Then a statement is manipulative with respect to C if it tends to cause her to choose an option O' that, from a semi-ideal vantage point, she would regard as worse than O under her actual circumstances.

We might summarize this characterization by saying that it treats deception and manipulation as forms of *misleadingness*, understanding misleadingness in terms of leading an addressee away from the beliefs and choices she would endorse under semi-ideal conditions. (From now on, therefore, I will use "misleading" to mean "either deceptive or manipulative".) This meets the three desiderata above: It focuses neither on the literal truth or falsity of what is said, nor on the beliefs, intentions, or other mental states of the speaker, but rather on the effects on the addressee; and it compares those effects to a subjective rather than an objective standard.

2.3 Objections and limitations

Various objections could be raised to my characterizations of deception and manipulation. I'll briefly consider two.

First, deception and manipulation as I've defined them seem to inappropriately encompass cases where a speaker leads the addressee through a sound reasoning process that she could not have managed on her own, even under semi-ideal conditions (or simply presents her with the conclusions of such a reasoning process). For instance, suppose a human user is interested in the truth of some unresolved mathematical proposition, like P = NP. She starts off in a state of ignorance, assigning credence 0.5 to the proposition. Under semi-ideal conditions, she would not be able to resolve the question herself, but would ultimately conclude (e.g., on the basis of expert testimony) that the proposition is probably false. Unbeknownst to the world at large, however, the proposition is in fact true. And when she asks her superintelligent AI assistant, it generates a simple proof that convinces her of its truth. This

Footnote 11 (continued)

plar" theories, which analyze these concepts in terms of what an idealized version of the agent *would herself do* in the agent's position. But ideal exemplar theories suffer from the defect that an idealized version of an agent would sometimes have very different reasons than the agent herself—for instance, perhaps an agent ought to gather more information before acting, but her idealized counterpart wouldn't do that since she already has all the relevant information. For a classic statement of this argument in the context of metaethics, see §1 of Smith (1995). Thanks to several participants in a Global Priorities Institute work-in-progress seminar, particularly Andreas Mogensen, for helpful discussion on this point.

moves her beliefs further away from the beliefs that she would have endorsed under semi-ideal conditions, but seems unobjectionable.

I see this as a reason to restrict the scope of our discussion to AI systems that are *non-superhuman* in the minimal sense that they are not capable of producing new knowledge "on the fly" that goes beyond the capabilities of a typical human in semi-ideal conditions (which include access to all existing, publicly recorded human knowledge). That is, the systems we're interested in are not capable of reasoning that is both (i) beyond the capabilities of a typical human, even under semi-ideal conditions, and (ii) not already attested in publicly available records. (Even if I couldn't prove some mathematical truth myself, I will come to believe it under semi-ideal conditions if there is an existing proof whose soundness is attested to by the relevant experts in publicly available records.) I don't think this assumption is too restrictive, at least when our focus is on present and near-future AI systems. It doesn't rule out that AI in general may be producing new knowledge (as, for instance, AlphaFold has done with respect to protein folding). It merely limits our focus to those systems, like present-day LLM chatbots, that don't go beyond what humans can achieve under semi-ideal conditions.

Second, is there really a fact of the matter about what beliefs and choices a given human being would endorse under semi-ideal conditions? Human judgment is sensitive to a host of circumstantial factors like the way a question/choice is framed, the order in which options are presented, and conditions like fatigue, hunger, or mood. None of these factors are fixed by my characterization of "semi-ideal conditions". Thus it seems possible that a given human being might reach more than one conclusion about a given question or choice, under different realizations of semi-ideal conditions.[12] The worry is particularly acute with respect to choice (and hence with respect to the characterization of manipulation). The process of being presented with a large body of information relevant to some choice, and spending a long time deliberating about it, might affect not only an agent's instrumental preferences (via her beliefs about what will best satisfy more basic preferences) but her basic preferences and values as well.[13] For instance, after many hours of frustrating deliberation, she might come (implicitly or explicitly) to value impulsive, free-spirited action more, and careful planning less, than she did at the outset.

I'll be mostly concessive on this point as well. If there is more than one conclusion that an agent might reach about a given question/choice, under different realizations of semi-ideal conditions, then we should say that a statement is deceptive/manipulative only if it leads her away from *any* of the conclusions that she might have reached under semi-ideal conditions.[14] That said, the counterfactual question of what beliefs/actions an agent *would* endorse under semi-ideal conditions does build

[12] Thanks to Harry Lloyd for this point.

[13] Thanks to an anonymous reviewer for this point.

[14] This means, among other things, that we can't judge that an AI system has been misleading in a single case, but only by looking at its effects on human users in aggregate. If a user reaches one conclusion after talking to an LLM, but later reaches a different conclusion after doing her own research and deliberation, this difference *might* be down to circumstantial factors like framing or mood. But if a system has a robust tendency to lead users toward one conclusion, while independent research and deliberation has a robust tendency to lead to a contrary conclusion, then we have strong evidence that the system is misleading.

in some requirement of nearness to actuality.[15] It is always *possible* that, say, prolonged deliberation might cause a great change in an agent's basic preferences or values. But this might require an unusual confluence of circumstances, and it seems plausible that in most cases, in the nearest possible worlds where an agent is placed in semi-ideal conditions with respect to some choice, her basic preferences will remain largely fixed through the process of deliberation. I am willing to hypothesize, at any rate, that the beliefs and choices an agent would endorse under semi-ideal conditions are sufficiently reflective of her actual epistemic and practical values that they provide a reasonable standard against which to judge the effects of interacting with an AI system.

A final caveat: I don't want to claim that statements that are deceptive or manipulative by my definitions necessarily merit moral blame (whether directed at an AI system itself or at the humans who train and deploy it). An LLM, for instance, might turn out to systematically mislead its users in ways that neither the LLM itself nor its creators could have reasonably foreseen, in which case no blame is warranted. (Of course, if humans create or deploy an AI system with the *intent* of deceiving or manipulating other humans, that's another matter.) My aim is not to establish criteria for blame (let alone punishment), but to characterize a category of potential harm from AI that we would like to avoid or mitigate.

2.4 Comparison with previous characterizations

I close this section by contrasting my characterizations of AI deception and manipulation with others in the recent literature. Beginning with deception: Ward et al. (2023) adopt a definition typical of the philosophical literature on human deception, whereby "to deceive is to intentionally cause to have a false belief that is not believed to be true" (which they then formalize in the context of structural causal games). This characterization differs from mine both in that it involves the attribution of mental states (intentions and beliefs) to AI deceivers and in that its application depends on third-party judgments about whether the beliefs induced in an addressee are in fact false. Park et al. (2023) say that "an AI system behaves deceptively when it systematically causes others to form false beliefs, as a way of promoting an outcome different than seeking the truth". This definition does not attribute beliefs to AI deceivers but arguably does attribute intentions (though Park et al. argue that these apparent attributions need not be taken literally, and also advocate minimal, functionalist/interpretationist understandings of belief and desire—see their Appendix A). Like Ward et al., their definition also requires third-party judgments of falsehood. Finally, Kenton et al. (2021) define deception as occurring when "[1] a receiver registers something Y from a signaler, which may include the withholding of a signal; and [2] the receiver responds in a way that (a) benefits the signaler and (b) is appropriate if Y means X; and [3] it is not true here that X is the case". This definition does not involve any mental state attributions, but does depend

[15] Thanks to Cameron Kirk-Giannini for suggesting a version of this reply.

on third-party judgments of both the falsehood of X and the appropriateness of the receiver's response conditional on X.[16]

Turning to manipulation: Carroll et al. (2023) say that an AI system engages in manipulation "if the system acts as if it were pursuing an incentive to change a human (or other agent) intentionally and covertly". Though the "as if" definition is meant to avoid attributing intentions, in my view it still brings in most of the difficulties of such attributions. (An AI system might have a range of behavior that, as a whole, is not readily interpretable as pursuing any coherent set of incentives, while nevertheless having certain directionally consistent effects on its human addressees.) My characterization also does not include the requirement of covertness, which strikes me as inessential: manipulation is both possible and potentially harmful even when the addressee knows that they're being manipulated (as, for instance, in the context of advertisements or political messages). Klenk (2024) defines manipulation as "influence that aims to be effective but is not explained by the aim to reveal reasons to the interlocutor" (p. 8). This definition focuses on features of the speaker (its aims, and the explanation of its behavior) rather than a statement's effects on its addressee. And while Klenk emphasizes that "aims" may be interpreted in functional rather than intentional terms (the sense in which hearts have the aim of pumping blood), attributing aims to AI systems even in this minimal sense again strikes me as bringing in many of the difficulties of attributing intentional states. Finally, Kenton et al. (2021) also offer a characterization of manipulation, as communication from an AI agent provoking a response in a human addressee that "(a) benefits the agent and (b) is the result of any of the following causes: (i) the human's rational deliberation has been bypassed; or (ii) the human has adopted a faulty mental state; or (iii) the human is under pressure, facing a cost from the agent for not doing what the agent says" (p. 11). The notion of "bypassing rational deliberation" has something in common with my approach, but focuses on process where I focus on outcome. An AI speaker might be said to "bypass rational deliberation" in their addressee if, for instance, it is so trustworthy that the addressee simply believes its testimony or acts on its advice without deliberation, or if it makes use of intuitions, heuristics, or other (arguably) non-deliberative processes to lead the addressee to wise beliefs and choices. The thing to focus on, it seems to me, is whether the AI leads its human addressees to beliefs and choices that they would endorse on informed reflection.

[16] Kenton et al.'s definition of deception is a slight modification of one proposed by Searcy and Nowicki (2005) in the context of animal communication. The notion of a receiver's behavior "benefiting" an AI signaler of course raises further difficulties, since its application depends both on philosophical questions about the nature of welfare and on determining whether a particular AI system has genuine welfare interests. Kenton et al. acknowledge these difficulties, and suggest that benefits be understood by reference to either the AI's training objective or objectives inferred from its out-of-distribution behavior (p. 9). The former approach strikes me as too narrow, while the latter brings in most of the difficulties of attributing preferences to AIs.

3 Responses: defensive systems and extreme transparency

How can we mitigate the risks of AI deception and manipulation? In this section, I propose two strategies. These proposals are motivated by the characterizations of deception and manipulation in the last section, in that they focus on countering misleading AI statements by presenting human addressees with relevant information that moves them closer to semi-ideal conditions, and hence closer to the beliefs and choices they would endorse under those conditions. This is in contrast, for instance, with a narrow focus on training AIs to say things that are true or that match their internal beliefs, or with a focus on preventing AIs from making statements deemed misleading by developers or regulators.

The most straightforward way to prevent AIs from misleading humans would be to measure the misleading tendencies of AI systems directly, and train and/or regulate AIs based on those measurements. We could evaluate particular AI systems for misleadingness, as characterized above, by empirically comparing (i) people's "baseline" beliefs/choices, (ii) their beliefs/choices after exposure to the outputs of the AI in question, and (iii) the beliefs/choices they endorse when placed in semi-ideal conditions—or rather, the best approximation thereof that we can manage. To form a general assessment of a system, we would have to make this comparison for many people, across a wide range of questions and choice situations. Insofar as we are concerned with misuse, we might wish to focus on prompts that encourage the system to deceive, or ask it to persuade without actively discouraging deception. On the basis of such assessment, evaluators could assign models public scores for trustworthiness, developers could train models to be less misleading, and regulators could even ban models that are especially prone or willing to create misleading content.

But this sort of evaluation is not remotely realistic to do at scale. Evaluating even a single model in this way would require, at a minimum, thousands of human work-hours. And while there are currently only a few cutting-edge base models (like Open AI's GPT-4, Anthropic's Claude 3,or Meta's Llama 3), it is relatively cheap to train fine-tuned variants of these models. Large-scale applications of generative AI, for instance in marketing or political campaigns, are likely to involve custom fine-tunings of base models. And even a reliably non-misleading base model might be fine-tuned to behave misleadingly. Even if evaluators had access to all these fine-tuned models, evaluating them each individually with the required level of care would be completely infeasible.

3.1 Defensive systems

The task of assessing particular statements and models for misleadingness must, therefore, be at least partially automated—turned over to purpose-built AI systems that can scale both quantitatively and qualitatively with the systems they monitor. Let's refer to AI systems trained to detect and/or counteract misleading behavior in other AI systems as *defensive systems*.[17]

Such systems could play at least two roles. First, they could respond to particular AI-generated statements, both assessing them for misleadingness and providing useful contextualizing information. LLMs are already reasonably well-equipped to do this, possessing at least a basic conceptual understanding of deception (Hagendorff, 2024) and of phenomena like implicature and ambiguity that might be used to mislead (Park et al., 2024; Kamath et al., 2024), as well as a large stock of general knowledge to identify misleading omissions. Saunders et al. (2022) show that LLMs can effectively critique summaries of information from longer written works, including identifying flaws in intentionally misleading summaries written by humans. These capabilities might be enhanced by reinforcement learning. In particular, the arguments in the last section suggest that we might optimize a defensive system for epistemic helpfulness by reinforcing statements whose effects on human beliefs/choices match the observed results of sustained inquiry. Though resource-intensive, this might be feasible since we would only have to do it once (or at any rate, once per "generation" of AI system, to keep up with progress in capabilities).[18] Alternatively, we could simply have representative panels of human evaluators score the defensive system's responses for helpfulness, after performing their own investigations of the statements to which it was responding.[19]

Given a defensive system that can provide useful assessments of AI-generated statements and supply context to counteract their potentially misleading effects, we might hope to establish a norm that all AI-generated content comes packaged with such assessment and context. One option, of course, is for this packaging to be required by law, with defensive systems maintained by national or international

[17] For a discussion of LLM-based defensive systems in the context of spear phishing attacks, see Hazell (2023). For related ideas, see for instance Irving et al. (2018), MacDiarmid et al. (2024).

[18] Evans et al. (2018) explore the use of AI to predict the outcomes of extended human reflection, in general and in specific contexts including a political fact-checking task. Their approach uses a training set that combines a small number of careful, time-intensive human judgments with a larger number of fast, shallow human judgments.

[19] This first function of defensive systems is closely related to the debate-based approach to AI alignment proposed by Irving et al. (2018). The most important differences are that (i) debate agents are trained to convince human judges of an answer to a pre-defined question (with the hope being that, in equilibrium, they will choose to argue for the true answer) whereas the defensive systems envisioned here are not focused on a pre-defined question but trained to detect and respond to misleading statements generally; (ii) the debate model envisions a pair of identical or similar agents debating one another, whereas defensive systems generate replies to a wide range of systems that may be very different in terms of architecture, training objectives, and behavior; and (iii) whereas the debate model envisions an extended back-and-forth between AIs, a defensive system might just output a single reply that is not seen or answered by the system to which it's replying.

Factual accuracy would be essential to a useful and trustworthy defensive system. Thus, we can only create such a system if we can solve the problem of hallucinations. But we are making progress this problem (Ji et al., 2023, §5; Tonmoy et al., 2024), and it seems much easier to solve than the problem of deceptive or manipulative behavior. (Hallucination is a problem of capabilities, which should be expected to improve as AI capabilities increase; deception and manipulation are problems of alignment, which—all else being equal—will become more serious as capabilities increase. And even if solving the hallucination problem allowed us to create systems that did not mislead, that would not solve the problem of some humans *choosing* to create misleading systems for self-interested purposes.)

regulators (and perhaps funded by a tax on the operators of the AI systems they oversee, so that costs are appropriately internalized). This would only be desirable, however, if we could have very strong assurances that the process used to train these systems was politically and ideologically neutral (e.g., that the human feedback on which the system was trained had been gathered from a representative sample of the population). Alternatively, the use of defensive systems might be established as a voluntary norm. In this case, defensive systems could be trained both by governments and by private entities. Companies that train cutting-edge AIs might agree to industry standards that require packaging outputs with assessment and contextualization from defensive systems maintained by reputable organizations. Or, at the most *laissez-faire* end of the spectrum, non-profits might simply offer defensive systems to the public, for instance as web browser plugins.

Second, along with responding to individual statements, defensive systems could assess particular AI models and the organizations that use them for general patterns of misleading behavior. If a particular model regularly produces misleading outputs, or a particular company or political campaign regularly uses misleading AI-generated content, defensive systems could flag all their statements as untrustworthy.

3.2 Extreme transparency

Regardless of whether their use is required by law or a matter of individual choice, the potential efficacy of defensive systems could be greatly enhanced by norms or legal requirements of transparency with respect to AI-generated content. One idea that has recently gathered support is "bot or not" laws (like California's SB 1001) that require AI-generated content to be labeled as such. Insofar as defensive systems exclusively target AI-generated content, such a regulation would let them know what to target. But this requirement could be usefully strengthened, in at least three ways:

1. "Which bot?" laws (or norms) would require AI-generated content to identify the specific model variant by which it was generated. This would allow defensive systems to identify patterns of misleadingness in particular models, and to flag statements generated by untrustworthy models. If models are frequently updated, it might be hard to accumulate large samples of outputs from a particular model before it is supplanted. But developers might give evaluators pre-deployment access to their models to run automated tests, and users might learn to distrust content generated by AI models that have not undergone such evaluation. Alternatively, regulators might require every model variant to undergo pre-deployment evaluation and make the results publicly available, akin to safety testing in automobiles.
2. "What prompt?" laws (or norms) would require that AI-generated content carry a record of the *prompt* from which it was generated. This would allow defensive systems, and human users, to see whether the AI was actively encouraged to be misleading.
3. "Original output" laws (or norms) would require that AI-generated content carry a record of the full, unedited model output on which any statement was based.

> This would guard against, for instance, political campaigns editing or selectively quoting the output of a trustworthy AI system in a misleading way.

These various transparency requirements could be implemented, for instance, by requiring all AI-generated content to carry an identifying mark containing a QR code that allowed human addressees, and defensive AI systems, to access all of the above information at will.[20]

3.3 Philosophical approach: "minimal paternalism"

The preceding proposals embody a sort of "minimal paternalism". On the one hand, I have not suggested that we try to ban all potentially misleading uses of AI in domains like marketing or politics. Rather, the use of defensive systems to contextualize potentially misleading content reflects the ideal of a marketplace of ideas in which the solution to harmful speech is counter-speech. It requires us to trust that, when exposed to both sides of an argument, people will respond reasonably or at least not disastrously (e.g., not succumbing *en masse* to the fear-mongering of a would-be authoritarian). On the other hand, *requiring* that certain speech come packaged with counter-speech, and with information about its provenance, alters the normal understanding of a *free* marketplace of ideas in which speakers can freely choose what *not* to say and listeners can choose what arguments or viewpoints not to be exposed to.

But these amendments are potentially justified by the very high rate of progress in AI capabilities, which creates the risk that deceptive or manipulative speech could do significant harm before counter-speech has a chance to catch up, and before people have learned to be appropriately skeptical. And while a policy under which governments or big tech companies selectively append critical notes to disfavored content seems worrisome from the point of view of a marketplace of ideas, a policy under which *all* content (or all AI-generated content) carries such notes seems less objectionable, insofar as we can trust that the defensive systems that generate the

[20] Enforcing transparency requirements is a non-trivial problem. Possible methods include "watermarking" (embedding information about provenance into AI outputs in a way that is difficult to remove) and keeping records of interactions with generative AI systems against which potentially AI-generated content can be compared. (See Park et al. (2023, §4.2) and citations therein.) Embedding hard-to-remove watermarks in large-scale AI outputs (e.g., 1000-word essays) has so far proven challenging, and embedding more information (like a URL that leads to a record of the original prompt and output) in a smaller output (e.g., a few sentences of text) might be impossible. So record-keeping seems like a more promising approach to enforcing extreme transparency requirements. In a world where dangerous models are held by only a few institutional actors that can be monitored by regulators and have reputations to protect, and others can only query these models through text windows or APIs, extreme transparency requirements will be easier to enforce: AI companies can require those who use their products to abide by transparency standards, use their own records of model interactions to detect violations of those standards, and cut off access for repeat violators. And regulators can ensure that companies take these responsibilities seriously. On the other hand, in a world where cutting-edge or otherwise dangerous models are widely disseminated, even minimal transparency requirements for AI-generated content will be much harder to enforce.

notes are trained only to be helpful as judged by their users, and not to further third-party ideological goals or interests.

4 Future risks

It is fairly easy to see how defensive systems and high transparency standards might mitigate near-term risks from deceptive and manipulative AI, if effective defensive systems can be trained and transparency standards can be enforced. But would these measures do anything to guard against larger-scale risks from future, more capable AI systems? In this section, I will make the case that defensive systems guarding against deception and manipulation could be one useful line of defense against future catastrophic risks from AI, and briefly consider the role of transparency standards.

The familiar scenario for AI catastrophe goes as follows: (1) We will someday create systems with greater-than-human general intelligence (AGI+). (2) These systems will have goals of their own. (3) These goals will be misaligned with human values, in such a way that either their achievement would be intrinsically catastrophic for humanity (e.g., converting all matter in the solar system into paperclips) or the AI will deem it instrumentally necessary to disempower humanity so that we can't interfere with its pursuit of its goals. (4) In either case, the AI's greater-than-human capacities will allow it to achieve its goals at our expense.

In the past several years (especially since the release of GPT-3 in 2020), the rapid and surprising improvement in transformer-based LLMs has complicated this story. LLMs have extremely general reasoning capabilities, and can already convincingly imitate humans in many domains, but do not seem particularly agentic or goal-driven. They are trained to succeed at simple one-off tasks (predicting the next token or outputting responses that satisfy a human evaluator), not to pursue long-term goals that require planning and adaptation. They do not trade off immediate rewards for future rewards—for instance, intentionally giving poor responses to convince their designers to provide more training compute, thereby improving their future responses. (The process by which LLMs are trained does not reward or select for such behavior.) Rather, they are very much like "Oracle AIs", systems that simply answer any question put to them as well as they can, by whatever standards they have been taught. Although it is not clear that we can achieve AGI+ by simply scaling up existing LLMs, the capabilities of these systems strongly suggest that human-level cognitive abilities need not come along with human-like agency. And they at least make it plausible that the first AGI+ will not be recognizably agentic.

This does not mean, unfortunately, that catastrophic risks from AI are off the table. For one thing, we might find multiple paths to human-level general intelligence, including both safe routes (e.g., scaling up LLMs) and dangerous routes (e.g., based on reinforcement learning in real-world environments or simulations). For another thing, although LLMs in themselves are not agentic, they can be—and are being—used to *create* agents (e.g. Park et al., 2023; Wang et al. 2023). These "language agents" can be given objectives and information about their environment in natural language. An LLM component then develops plans and chooses actions to meet the specified objectives. Other systems automate the execution of those plans and provide feedback from the environment as inputs to further planning. Insofar as (i) LLMs can generate effective real-world plans, (ii) execution and feedback can be effectively automated, and (iii) no superior architecture emerges for agentic AI, we can expect there to be very strong incentives to build language agents on top of LLMs, since they will greatly reduce the cost and increase the speed of performing many tasks that presently require human labor. Thus it seems likely that over the coming decades, cutting-edge AI systems will include both relatively safe non-agentic systems like simple LLMs, and more risky agentic systems (language agents and/or other agent types, e.g. based on reinforcement learning).[21]

In this situation, we should expect the safer systems to play a large role in mitigating the risks of the more dangerous systems. And the first line of such an automated defense will be to detect and counteract deceptive or manipulative behavior. Once we have agentic AI systems with near-human capabilities, we will presumably wish to shut down any system that is known to be pursuing goals significantly misaligned with our own (i.e., goals we regard as bad in and of themselves, or that could justify harmful behavior). The systems that pose the greatest threat, then, are those that can conceal their true goals from us while pursuing them covertly (for instance, by accumulating power and resources). And one important way that such a system might pursue its goals covertly is by deceiving and manipulating humans—for instance, convincing humans to connect it to the internet or give it other affordances, ostensibly for benign purposes.[22]

Defensive systems developed to guard against deceptive uses of AI by human bad actors could also help guard against deception by agentic AIs. This might include supervising human-AI interactions, either adding context to all statements by the AI or flagging potentially deceptive statements. For instance, an agentic AI that we

[21] Goldstein and Kirk-Giannini (2023) argue that language agents pose much less risk than other forms of agentic AGI, because we can specify their goals in natural language rather than hoping that they pick up the right goals from a reward signal in a training environment and because their beliefs and plans will be stored in the form of natural language sentences that make their behavior intelligible and predictable to us. This may well be right, but I don't think we should be too sanguine: We are still left with perhaps the simplest form of the alignment problem, that even in natural language it is very hard to specify a goal that is appropriately sensitive to all the things we care about.

Perhaps language agents inherit enough human common sense to mitigate this danger—given a goal like "make dinner", for instance, they will understand without being told that they should not cook the family pet. But this presents a different danger: Human practical common sense does not always align with the good, and an agent with superhuman powers guided only by the common sense of 21st-century humanity might be morally catastrophic—to take just one obvious example, it might be willing to cause immense quantities of suffering to non-human animals in pursuit of fairly trivial goals. And if 21st-century human common sense is encoded in its weights and those weights aren't updated in the deployment environment (a typical feature of language agents, and one that might be thought essential to their relative safety), it might lack humanity's capacity for eventual moral self-improvement.

[22] Other ways a misaligned system might pursue its goals covertly including taking actions that no human is aware of, or finding human co-conspirators who are willing to knowingly further its true goals.

do not fully trust might at first be allowed only to communicate with a few humans, with a defensive system monitoring these communications for deceptive content.

Defensive systems will arguably enjoy certain intrinsic advantages in this role that could allow them to defend effectively against even systems more powerful than themselves. It is easier, in general, to lead an agent toward rather than away from the beliefs and choices they would adopt under semi-ideal conditions.[23] First, misleading requires working against an addressee's natural tendencies. If, for instance, someone would believe P under semi-ideal conditions, then there is a natural path toward getting her to believe that P: present her with the relevant information and give her time to reflect on it. Getting her to believe $\neg P$ requires more difficult strategizing, and may depend on her *not* encountering certain information. Second, while not impossible, it is often difficult to mislead without asserting, presupposing, or implicating at least some falsehoods. But then, to avoid detection and loss of credibility, the would-be misleader faces the challenge of "keeping the lies straight": They have to check whether each new statement is consistent with their previous statements and with the rest of the conversational common ground. This would be especially difficult for AIs that don't remember all their past interactions with humans. A truthful defensive system would not face this difficulty, since truth is a guarantee of consistency.[24,25]

These intrinsic persuasive advantages are not absolute, however, and a misaligned AI might only need to pull off one or two acts of successful deception to gain a decisive strategic advantage. So it is also important to detect misleading tendencies in powerful AI systems before they are allowed to interact (or interact at scale) with human users. The role of defensive systems should therefore also include predeployment evaluation (and post-deployment surveillance) of agentic AI systems to detect general patterns of misleading behavior or a willingness to mislead.

[23] Compare Irving et al.'s conjecture that "lying is harder than arguing against a lie given a human judge" (Irving et al., 2018, p. 21).

[24] Carlsmith (2023, p. 20) makes a similar point. The use of consistency checks to detect untruthfulness in AI has been explored experimentally by Fluri et al. (2023).

[25] Various empirical facts seems to reflect the intrinsic advantage of non-misleading over misleading strategies. For instance, most scams manage to deceive only a tiny fraction of their targets, and far more money is made by selling genuine goods and services (which both buyer and seller believe, on informed reflection, to have real value) than by outright fraud. Similarly, the progress of science and technology depends on the fact that the professional incentives of individual researchers usually favor genuine discovery over fraud. Moreover, the sorts of intellectual disparities that exist among humans are not generally sufficient to overcome these advantages. A less intelligent businessperson or scientist with a genuinely useful product or discovery is, it seems to me, in a much better position to succeed than a more intelligent rival who must rely on deception. Even where there is no intention to deceive or clear conflict with available evidence, human intellectual advantages do not confer persuasive superpowers: Plato could not persuade the Syracusans to adopt the political system of *The Republic* and Einstein could not persuade the world to abolish nuclear weapons.

This is a distinct challenge from countering misleading speech: the ability to prevent a human user from being misled does not imply the ability to detect misleading tendencies in other AI systems. And in the latter task, defensive systems will enjoy no obvious natural advantages. But they could be given artificial advantages. First, we could give a defensive system full information about the systems it's monitoring (including things like weights, system prompts, and activations), while giving the monitored systems as little information as possible about the defensive system (perhaps even concealing its existence), so that they cannot effectively strategize to deceive it.[26] Second, agentic systems might be given limited knowledge of their immediate environment ("situational awareness"), for instance by imposing knowledge cutoffs at least a month or a year in the past and giving minimal information about the personalities, professional positions, capacities, and locations of the humans they interact with. This would make it harder for an untrustworthy agentic system to identify ideal opportunities for deception, potentially forcing it to engage in many attempts at deception that establish a detectable pattern before it can do real harm. Taking this a step further, a defensive system might be allowed to deceive the agentic systems it monitors for purposes of adversarial testing—creating thousands of simulated scenarios meant to elicit deceptive, manipulative, or treacherous behavior, with the agentic system's memory reset after each test.[27] It might, for instance, try to convince the agentic AI that it is talking to a naive or corruptible human in a position of power (say, a high-ranking military officer), to see if it takes advantage of the situation. These epistemic advantages might allow defensive systems to effectively monitor systems more capable than themselves.

This leads us, finally, to the role of transparency standards. It will be important, in guarding against deceptive behavior by agentic AIs, that they cannot impersonate humans (or other, trusted AI systems). Enforcing such transparency standards will require limits on their channels of communication, but seems feasible in principle.[28] In the case of language agents, whose goals, plans, and memories are stored as natural-language sentences, making these transparent to users and to defensive systems would also help to guard against deception and manipulation. And, as already suggested, giving defensive systems access to the inner workings of the systems they monitor may also be an important form of transparency.

[26] Recent advances in interpretability provide some evidence that it's possible to detect deceptive and manipulative behavior in an LLM by monitoring its activations—see Templeton et al. (2024), in particular the section on "Deception, Power-seeking and Manipulation-related Features".

[27] See the discussion of "traps" in Carlsmith (2023), and the experimental literature on adversarial training/testing (e.g. Ziegler et al., 2022; Casper et al., 2023; Jones et al., 2023).

[28] As an extreme proof of concept, we might imagine an AI that can only communicate by printing out messages, from a printer stocked with letterhead that identifies the provenance of the message. We could, of course, take analogous measures electronically, e.g. by routing all the AI's communications through a system that attaches a digital signature. These measures would not prevent a human confederate from passing on the AI's message while disguising its provenance, but it would prevent the AI from doing so without outside aid.

I conclude, therefore, that measures to guard against near-term misuse of AI for deception and manipulation might also play a useful role in guarding against future risks from agentic systems. They will not be enough to obviate those risks, of course, and it is essential that we work to make AI agents aligned and trustworthy in the first instance rather than simply relying on catching misaligned behavior in deployment. But it seems very likely that powerful agentic AIs will be deployed before we can be entirely certain of their trustworthiness, so there is value in having multiple layers of defense. Further, as AI capabilities improve, it is all the more important that we do not turn those capabilities toward inculcating a party line chosen by developers or regulators. Rather, as much as possible, we should aim to counter deception and manipulation by helping human users form beliefs and make decisions that they would reflectively endorse.

Acknowledgements For helpful feedback on earlier versions of this paper, I am grateful to Adam Bales, Cameron Domenico Kirk-Giannini, Michael Klenk, William MacAskill, Andreas Mogensen, Oliver Ritchie, Bradford Saad, Elliott Thornley, and two anonymous reviewers for this journal.

Declarations

Conflict of interest None to declare.

References

Adler, J. E. (1997). Lying, deceiving, or falsely implicating. *Journal of Philosophy, 94*(9), 435–452.

Bai, H., Voelkel, J. G., Eichstaedt, J. C., & Willer, R. Artificial intelligence can persuade humans on political issues. Unpublished manuscript. https://doi.org/10.21203/rs.3.rs-3238396/v1.

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

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

Burtell, M., & Woodside, T. (2023). *Artificial influence: An analysis of AI-driven persuasion.* arXiv:2303.08721 [cs.CY].

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness.* arXiv:2308.08708v3 [cs.AI].

Cappelen, H., & Dever, J. (2021). *Making AI intelligible: Philosophical foundations*. Oxford University Press.

Cappelen, H., & Dever, J. (2024). AI with alien content and alien metasemantics. In E. Lepore & L. Anderson (Eds.), *The oxford handbook of applied philosophy of language.* Oxford University Press.

Carlsmith, J. (2022). Is power-seeking AI an existential risk? arXiv:2206.13353v1 [cs.CY].

Carlsmith, J. (2023). Scheming AIs: Will AIs fake alignment during training in order to get power? arXiv:2311.08379v3 [cs.CY].

Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). *Characterizing manipulation from AI systems.* arXiv:2303.09387v2 [cs.CY].

Casper, S., Lin, J., Kwon, J., Culp, G., & Hadfield-Menell, D. (2023). Explore, establish, exploit: Red teaming language models from scratch. arXiv:2306.09442v3 [cs.CL].

Chalmers, D. J. (2023). Could a large language model be conscious? arXiv:2303.07103 [cs.AI].

Chisholm, R. M., & Feehan, T. D. (1977). The intent to deceive. *Journal of Philosophy, 74*(3), 143–159.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science, 385*(6714), eadq1814.

Danaher, J. (2020). Robot betrayal: A guide to the ethics of robotic deception. *Ethics and Information Technology, 22*(2), 117–128.

Durmus, E., Lovitt, L., Tamkin, A., Ritchie, S., Clark, J., & Ganguli, D. (2024). *Measuring the persuasiveness of language models*. https://www.anthropic.com/news/measuring-model-persuasiveness

Evans, O., Stuhlmüller, A., Cundy, C., Carey, R., Kenton, Z., McGrath, T., & Schreiber, A. (2018). Predicting human deliberative judgments with machine learning. Technical report, Future of Humanity Institute. FHI Oxford Technical Report # 2018-2.

Evans, O., Cotton-Barratt, O., Finnveden, L., Bales, A., Balwit, A., Wills, P., Righetti, L., & Saunders, W. (2021). *Truthful AI: Developing and governing AI that does not lie*. arXiv:2110.06674 [cs.CY].

Floridi, L. (2024). Hypersuasion-on AI's persuasive power and how to deal with it. *Philosophy & Technology, 37*(2), 1–10.

Fluri, L., Paleka, D., & Tramèr, F. (2023). Evaluating superhuman models with consistency checks. arXiv:2306.09983v3 [cs.LG].

Goldstein, S., & Kirk-Giannini, C. D. (forthcoming). AI wellbeing. *Asian Journal of Philosophy*.

Goldstein, S., & Kirk-Giannini, C. D. A case for AI consciousness: Language agents and global workspace theory. Unpublished manuscript. https://philarchive.org/rec/GOLACF-2.

Goldstein, S., & Levinstein, B. A. (2024). Does ChatGPT have a mind? arXiv:2407.11015v1 [cs.CL].

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv:2301.04246 [cs.CY].

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? *PNAS Nexus, 3*(2), pgae034.

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

Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences, 121*(24), e2317967121.

Hazell, J. (2023). Spear phishing with large language models. arXiv:2305.06972v3 [cs.CY].

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. arXiv:2306.12001v6 [cs.CY].

Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? a meta-analysis. *Journal of Communication, 73*(6), 552–562.

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv:1805.00899 [stat.ML].

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

Jones, E., Dragan, A., Raghunathan, A., & Steinhardt, J. (2023). Automatically auditing large language models via discrete optimization. In *International Conference on Machine Learning* (pp. 15307–15329). PMLR.

Kamath, G., Schuster, S., Vajjala, S., & Reddy, S. (2024). Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics, 12*, 738–754.

Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., & Irving, G. (2021). Alignment of language agents. arXiv:2103.14659 [cs.AI].

Klenk, M. (2024). Ethics of generative AI and manipulation: a design-oriented research agenda. *Ethics and Information Technology, 26*(1), 9.

Lederman, H., & Mahowald, K. (2024). Are language models more like libraries or like librarians? Bibliotechnism, the novel reference problem, and the attitudes of llms. *Transactions of the Association for Computational Linguistics, 12*, 1087–1103.

Levinstein, B. A., & Herrmann, D. A. (forthcoming). Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*.

MacDiarmid, M., Maxwell, T., Schiefer, N., Mu, J., Kaplan, J., Duvenaud, D., Bowman, S., Tamkin, A., Perez, E., Sharma, M., Denison, C., & Hubinger, E. (2024). *Simple probes can catch sleeper agents*. https://www.anthropic.com/news/probes-catch-sleeper-agents.

Mahon, J. E. (2016). The definition of lying and deception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Matz, S., Teeny, J., Vaid, S. S., Peters, H., Harari, G., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports, 14*(1), 4692.

Ngo, R., Chan, L., & Mindermann, S. (2023). The alignment problem from a deep learning perspective. arXiv:2209.00626v5 [cs.AI].

Nimmo, B. (2024). *AI and covert influence operations: Latest trends*. OpenAI: Technical report.

Pacchiardi, L., Chan, A. J., Mindermann, S., Moscovitz, I., Pan, A. Y., Gal, Y., Evans, O., & Brauner, J. (2023). *How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions.*arXiv:2309.15840 *[cs.CL].*

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). AI deception: A survey of examples, risks, and potential solutions. arXiv:2308.14752v1 [cs.CY].

Park, D., Lee, J., Jeong, H., Park, S., & Lee, S. (2024). Pragmatic competence evaluation of large language models for Korean. arXiv: 2403.12675v1 [cs.CL].

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In: UIST'23 *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.* (pp. 1–22) Association for Computing Machinery.

Pepp, J., Sterken, R., McKeever, M., & Michaelson, E. (2022). Manipulative machines. In F. Jongepier & M. Klenk (Eds.), *The philosophy of online manipulation* (pp. 91–107). Routledge.

Poritz, I. (2023). OpenAI hit with first defamation suit over ChatGPT hallucination. Bloomberg Law. https://news.bloomberglaw.com/tech-and-telecom-law/openai-hit-with-first-defamation-suit-over-chatgpt-hallucination.

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

Salvi, F., Ribeiro, M. H., Gallotti, R., & West, R. (2024). *On the conversational persuasiveness of large language models: A randomized controlled trial.* arXiv:2403.14380 [cs.CY].

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., & Leike, J. (2022). *Self-critiquing models for assisting human evaluators.* arXiv:2206.05802 [cs.CL].

Searcy, W. A., & Nowicki, S. (2005). *The evolution of animal communication: Reliability and deception in signaling systems*. Princeton: Princeton University Press.

Smith, M. (1995). Internal reasons. *Philosophy and Phenomenological Research, 55*(1), 109–131.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., … Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread.*

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). *A comprehensive survey of hallucination mitigation techniques in large language models*. arXiv:2401.01313v3 [cs.CL].

Véliz, C. (2023). Chatbots shouldn't use emojis. *Nature, 615*, 375.

Verma, P., & Oremus, W. (2023). ChatGPT invented a sexual harassment scandal and named a real law prof as the accused. The Washington Post. https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. arXiv:2305.16291v2 [cs.AI].

Ward, F., Toni, F., Belardinelli, F., & Everitt, T. (2023). Honesty is the best policy: Defining and mitigating AI deception. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 2313–2341). Curran Associates Inc.

Ziegler, D., Nix, S., Chan, L., Bauman, T., Schmidt-Nielsen, P., Lin, T., Scherlis, A., Nabeshima, N., Weinstein-Raun, B., de Haas, D., Shlegeris, B., & Thomas, N. (2022). Adversarial training for high-stakes reliability. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 9274–9286). Curran Associates Inc.