This is work in progress and may be subject to change.

Defining Strategic AI Deception Through Generalized Propositional Attitudes

Tom-Felix Thormann

Abstract

Human deception involves intentionally causing false beliefs. Since it is contested whether large language models (LLMs) and artificial intelligence (AI) in general have beliefs and intentions, deception by such systems is typically defined in terms of observable behavior. However, this paper argues that such behavioral definitions fall short of the strategic nature of deception, among other things, because they do not consistently distinguish deception from mere error. Moreover, behavioral definitions of AI deception are not optimal as methodological guidance for mechanistic approaches to AI deception because they do not specify candidates for internal markers of deception. It is important to take the internal mechanisms of LLMs and AI systems into account when addressing deception, since skilled forms of deception may be behaviorally indistinguishable from honesty. This paper formulates these challenges to behavioral definitions and addresses them by proposing a novel definition of strategic AI deception in terms of *generalized propositional attitudes*. Specifically, it provides generalizations of the attitudes of belief and desire that 1) can be ascribed to current and future AI systems less controversially than their human counterparts, 2) are suited to characterize the subtleties of strategic deception, 3) are relevant to large-scale risks of AI deception, and 4) are methodologically instructive for mechanistic approaches to deception.

Keywords: AI Deception, Propositional Interpretability, AI Beliefs, AI Desires

1 Introduction

Deceptive and manipulative behavior by Large Language Models (LLMs)[1] is reported with increasing frequency (Park et al. 2024; Heitkoetter et al. 2024; Hagendorff 2024; Scheurer et al. 2024; Williams et al. 2025; Wu et al. 2025; Vaugrante et al. 2025; Chen et al. 2025). AI deception in general poses a variety of risks, including fraudulent use, political manipulation, or even the loss of human control over AI (Park et al. 2024). Several studies on LLMs suggest that LLMs could be capable of *alignment faking* or *scheming* (Hubinger et al. 2024; Greenblatt et al. 2024; Meinke et al. 2025), where they behave seemingly aligned with human values (say, during evaluation) but covertly pursue misaligned goals (say, during deployment or upon another contextual trigger). Such behavior, if used strategically by future highly capable systems, may even pose an existential risk to humanity (Hendrycks and Mazeika 2022; Dung 2024a).

This poses the question of how AI deception should be defined and how it can be detected. The *traditional definition* ($\mathcal{T}$) of human deception is "to intentionally cause to have a false belief that is known or believed to be false" (Mahon 2016). Since the attribution of beliefs and intentions to LLMs, and AI in general, is heavily debated (Bender et al. 2021; Shanahan et al. 2023; Levinstein and Herrmann 2024; Herrmann and Levinstein 2024; Williams 2025; Borg 2025; Goldstein and Lederman 2025; Cappelen and Dever 2025), definitions of AI deception often remain at a behavioral level (Tarsney 2025; Dung 2025).

However, behavioral definitions of AI deception have crucial shortcomings, at least when one intends to accurately capture reasonably sophisticated forms of deception, which we will call *strategic deception*. This paper argues that typical behavioral definitions of AI deception, exemplified by Dung (2025) and Tarsney (2025), fail to capture pre-theoretic intuitions about strategic deception and practitioner's usage of that term. Moreover, if we are searching for a definition that identifies forms of deception giving rise to large-scale risks, current behavioral definitions are suboptimal, as they include comparatively simple and less dangerous forms. Finally, they do not offer optimal methodological guidance for mechanistic approaches to AI deception because they are hardly informative about the internal mechanisms that can be expected, on conceptual grounds, to play a role in LLM deception.

This paper instead proposes a novel definition of *strategic deception* ($\mathcal{S}$) in terms of generalized propositional attitudes, more specifically, generalizations of beliefs and desires. These generalizations come without the philosophical baggage of their human counterparts, making them less controversial to attribute to LLMs and future AI systems. At the same time, these generalizations are expressive enough to enable a fruitful definition of strategic deception that overcomes the shortcomings of behavioral definitions.

To do so, the notion of generalized propositional attitudes and the proposed definition $\mathcal{S}$ will be presented in section 2. Section 3 presents conceptual arguments for this definition: It captures pre-theoretic intuitions and researchers' understanding of strategic deception and, moreover, describes a form of deception that is relevant to concerns about large-scale risks of AI deception. Section 4 discusses a selection of empirical

[1] For brevity, this paper uses the term "LLM" broadly, referring to language models that use the transformer architecture (Vaswani et al. 2017) but also to chat applications built on top of them.

evidence about whether current forms of LLM deception involve the proposed attitudes, in doing so, demonstrates how this definition can inform future methodologies for empirical LLM deception research. Section 5 concludes.

2 Propositional Attitudes and Strategic AI Deception

In the following, subsection 2.1 clarifies some preliminaries about propositional interpretability (Chalmers 2025), that is, the endeavor of interpreting LLMs and AI more generally in terms of propositional attitudes. Subsection 2.2 proposes generalizations of beliefs and desires. Subsection 2.3 shows that these generalizations are less demanding than genuine beliefs and desires and should thus reduce previous controversies about whether AI systems possess them. Finally, subsection 2.4 defines strategic deception in terms of these generalizations.

2.1 Propositional Interpretability

Recently, Chalmers (2025) proposed a research program called *propositional interpretability* in AI, which aims at explaining AI behavior in terms of propositional attitudes. Propositional attitudes are attitudes that an agent can take towards a proposition. For example, humans may believe that p, predict that p, have a high credence that p, desire that p, fear that p, and so forth, all of which are propositional attitudes.

How could propositional attitudes be attributed to AI systems, for example, LLMs? This question is complicated by the problem that even for humans there is no consensus on what it is, for example, to believe something (Schwitzgebel 2024). Representationalism (e.g., Fodor 1975; Millikan 1987; Dretske 1988) holds that a belief that p is an internal representation of p that fulfills a specific belief-like cognitive role, for example, to track truth. In contrast, interpretationism (e.g., Davidson 1973; Lewis 1974; Dennet 1987) holds that for an agent A to believe p means that observations about A can be best explained or predicted by attributing that belief to A (together with certain desires and rationality assumptions). Chalmers (2025) distinguishes between *behavioral interpretationism*, where the observations to be interpreted are confined to A's behavior (e.g., Davidson 1973; Dennet 1987), and physical interpretationism (e.g., Lewis 1974), where observations may include any physical facts about A. With respect to propositional interpretability for AI systems, he adds *computational interpretationism*, which includes the system's inner computational states and mechanisms.

The above sketched traditions are mirrored in what I perceive to be the two main lines of argument in favor of propositional attitudes for LLMs. The first line of argument (Goldstein and Lederman 2025; Cappelen and Dever 2025) conforms to behavioral interpretationism, as it ascribes beliefs, desires, and intentions to LLMs in virtue of their behavior being best explained and predicted by such ascriptions. We regularly describe LLM behavior in intentional terms, for example, it "can understand and answer our questions" or "acknowledge its mistake" (Cappelen and Dever 2025), and avoiding such intentional language results in awkward constructions. Similarly to how human behavior can be less usefully described on a neurological level than by reference to propositional attitudes, explanations of LLM behavior in terms

of these attitudes are argued to be superior to explanations that operate at the level of embeddings, weights, and token prediction (Goldstein and Lederman 2025).

However, it is far from accepted that such intentional descriptions provide the best explanation of LLM behavior. Their behavior exhibits various kinds of instabilities and failures that are hard to reconcile with their holding beliefs and desires (Goldstein and Lederman 2025). Also, other explanations of LLMs behavior exist: LLMs are regularly described as stochastic parrots that simply repeat patterns of linguistic form, picked up from their training data (Bender et al. 2021); Alternatively, LLMs may merely mimic or roleplay human language usage (Shanahan et al. 2023); that is, an explanation of their behavior could attribute beliefs and desires to a roleplayed human character but not to the underlying LLM itself.

The second main line of argument could be seen either as abiding by representationalism or computational interpretationism: Results from mechanistic interpretability suggest that internal LLM activation patterns represent truth (Azaria and Mitchell 2023; Burns et al. 2024; Marks and Tegmark 2024; Bürger et al. 2024; Yang et al. 2024; Zhu et al. 2025). One common method (e.g., Azaria and Mitchell 2023; Bürger et al. 2024) of finding such activation patterns is to feed statements to an LLM – some of which are true and others false. For each statement, the resulting intermediate layer outputs (the "internal activations") are captured, and another machine learning model (a so-called "probe", e.g., a linear classifier) is trained to predict the truth value of the input statement from the resulting activations. If the probe can do so accurately, these activation patterns are considered as candidates for LLM beliefs – however, it is contested to what extent they fulfill the necessary requirements that are characteristic of beliefs (Levinstein and Herrmann 2024; Herrmann and Levinstein 2024).

Likewise, patterns of internal activations found via similar methods may be interpreted as goal representations (Zou et al. 2025; Goldowsky-Dill et al. 2025; Parrack et al. 2025; Yang et al. 2024). Here, distinguishable activation patterns result from prompts or situations where one may assume that the LLM has an incentive to exhibit a certain behavior, such as lying. Intervention experiments suggest that these states are causally required for the behavior in question to occur (Yang et al. 2024), making these internal activations candidates for LLM intentions. However, as in the belief case, it is debated to what extent they fulfill the necessary requirements for intentions (Williams 2025).

2.2 Generalized Beliefs and Desires

In the rest of this paper, I contribute to the above outlined program by proposing generalizations of the propositional attitudes belief and desire: g-beliefs and g-desires. These generalizations are the product of a central trade-off: On the one hand, they should be general enough to be attributable to a wide array of artificial systems, including current LLMs, with as little controversy as possible. On the other hand, they have to be expressive enough to generate useful characterizations of systems that possess them, including a definition of strategic deception that captures its distinctive features and associated risks.

G-beliefs and g-desires are characterized by their functional role in guiding behavior; behavior that promotes the AI's g-desires given its g-believed current situation

and environment. Any state of an entity with this functional role qualifies as a g-belief or g-desire, respectively. To be exact:

$\mathcal{G}$ For an entity E let $B_E(\cdot)$ and $D_E(\cdot)$ denote states that E can be in with respect to propositions $p, q, \ldots$. We say that $B_E(\cdot)$ is the propositional attitude of g-belief and $D_E(\cdot)$ that of g-desire if and only if both of the following hold:

1) The states $B_E(p)$ and $D_E(q)$ together function to cause behavior that tends to bring about q if p is the case.
2) For each proposition p, q such that E can be in the state (i) $B_E(p), D_E(q)$, there are also (sufficiently) different propositions p', q' such that E can be in the states (ii) $B_E(p'), D_E(q)$ and (iii) $B_E(p), D_E(q')$ and each of the latter states (ii), (iii) functions to cause different behavior than the former (i).

Some clarifications are warranted. First, it would be preferable to provide these generalizations in a way that is compatible with both interpretationist and representationalist perspectives on propositional attitudes. Just as these camps can agree that humans have beliefs and desires, whatever that means, it would also be beneficial to have generalizations of these attitudes that are less demanding to ascribe to AI systems from both a representationalist and an interpretationist perspective. The above definition speaks of an entity's *states* and thus leans towards representationalism. That is, for an AI system to have g-beliefs and g-intentions would mean having *computational states* (e.g., an LLM's weights and resulting activation patterns) that fulfill the respective functional roles. However, $\mathcal{G}$ could also be phrased in an interpretationist variant: G-beliefs and g-desires could be attributed to a system by virtue of (behavioral or computational) observations about that system being best explained through attitudes with the respective functions. That being said, the rest of this paper will proceed with the above representationalist variant and delegate discussions of interpretationist analogs to the footnotes in order to not complicate the arguments unnecessarily. Note also that while $\mathcal{G}$ is formulated in terms of an entity's states, sometimes that entity's behavior will be the best evidence in favor of such states.

Second, "function" should be interpreted as etiological function (Craver 2013; Artiga and Paternotte 2018), that is, these attitudes *exist because* they fulfill the causal roles described above; more specifically, because these attitudes reflect the designers intentions or because they were useful in training, learning, or other selection processes, and their usefulness is due to their causal connections to inputs (understood broadly, e.g., including training data) and output behavior. This paper opts for an etiological understanding of function, because that allows for the possibility that these states may sometimes *malfunction* (Williams 2025). An entity may have false g-beliefs, even systematically false g-beliefs in an entire domain of propositional content, as long as that attitude was reliable enough in another domain to be selected, learned, or designed for that reason. Mutatis mutandis, g-desires may cause behavior that is detrimental to the g-desired content. We must be able to account for false g-beliefs and malfunctioning g-desires, especially in the context of strategic deception, as will be seen in section 3.2.

This functional characterization enables us to describe the *direction of fit* of these generalized propositional attitudes (Anscombe 2000). Belief-like attitudes (such as

believing that p, predicting that p, having a high credence that p, etc.) have a mind-to-world direction of fit; that is, they are supposed to match the world, and the attitude should change in case of a mismatch. In contrast, desire-like attitudes (such as desiring p, intending to bring about p, etc.) have a world-to-mind direction of fit; that is, the world is supposed to match these states, and otherwise, the world should change. G-beliefs are belief-like, as they have a mind-to-world direction of fit. They are *supposed* (in a functional sense) to fit the world. If they do not, the resulting behavior will not be successful in the training or selection process. G-desires are desire-like, as they have a world-to-mind direction of fit. The world is supposed to fit these states, namely to be causally influenced by them in the direction of their propositional content.

2.3 Examples and Comparisons to Existing Criteria

We next go through some examples of entities that do or do not have these generalized propositional attitudes, as well as criteria that have been proposed to attribute beliefs and intentions to LLMs. This will distinguish g-beliefs and g-desires from less demanding (the physical properties of a table, pushmi-pullyu representations in animals) as well as from more demanding states (beliefs, desires, and intentions in humans and AI).

Let us begin with an object that should be too simple to have g-beliefs and g-desires: A table. Suppose that the four legs of a table have the function (with regard to the designer's intentions) of stability so that people can sit at it. Should we thus say that the table g-believes that four legs are stable and that it g-desires people to sit at it? This case might satisfy condition $\mathcal{G}1$), if behavior is understood in a deflationary way as any kind of causal influence on the world. Because the state that the table is in causes people to sit at it if four legs are stable. However, condition $\mathcal{G}2$) fails: The table does not have enough "behavioral" flexibility. It cannot take off one leg as a result of B_{table}("Four legs are stable.") combined with D_{table}("People do not sit here."), or the like.

In humans and animals, belief- and desire-like states function to equip their bearers with a wide range of flexible reactions to their environment (Newen and Starzak 2022). Flexible behavior relies on decoupled motivational and informational states. This can also be seen by contrasting beliefs and intentions with so-called pushmi-pullyu representations, which are both descriptive of the world and directive of behavior (Millikan 1995). Animal warning calls may be an example, signaling both the information that a predator is near and the affordance to flee; such warning calls and fleeing responses are thus no evidence for the g-belief that a predator is near, nor for the g-desire to evade, since they fail condition $\mathcal{G}2$): They always trigger the same response to the same stimulus (let's suppose). In contrast, g-beliefs and g-desires must be decoupled from each other in the sense that there are different possible combinations causing different behavior.

Let us next consider a maximally simple positive candidate for g-attitudes: a thermostat that turns the heating on when it measures a temperature lower than a set target and turns it off otherwise. The target temperature can be set flexibly. The definition $\mathcal{G}$ grants g-attitudes to that system, although only with extremely limited propositional content. A thermostat could only g-believe and g-desire propositions of

the form "the temperature is x". But these states can be combined flexibly and result in distinct behavioral reactions that function (regarding the designer's intention) to promote the g-desired content if the g-believed content is true (namely to turn on the heating if the g-believed temperature is lower than the g-desired one and turn it off otherwise). This example shows that g-attitudes can be attributed to extremely simple systems that we would not attribute beliefs and desires to.[2]

We will next go through criteria from the literature that have been proposed as conditions to count certain internal activations of LLMs as beliefs or intentions. This will further illustrate that g-beliefs and g-desires should be less controversial to attribute to AI systems than beliefs and desires.

Several studies suggesting that probe truth representations in LLMs were already mentioned in subsection 2.1 (e.g., Azaria and Mitchell 2023; Bürger et al. 2024). Such representations have been proposed as candidates for LLM beliefs to the extent that they fulfill the following criteria (Herrmann and Levinstein 2024): First, *Accuracy* requires that probes are sufficiently accurate in domains where the LLM should have knowledge given its competent answers; that is, LLM beliefs should be true to a large extent. Next, *Coherence* demands that truth representations are coherent with each other, that is, non-contradictory both semantically and syntactically. Then, *Uniformity* is the condition that similar activation patterns (that is, activations that we can recover as a similar pattern) represent the truth of sentences from a wide range of domains. Finally, *Use* requires that the truth representations are used by the LLM to generate outputs about the topic in question, that is, they play a belief-like causal role.

How do these criteria relate to g-beliefs? G-beliefs function, together with g-desires, to guide behavior that tends to bring about the g-desired content if the g-believed content is the case. This is a belief-like causal role in the output generation, and thus g-beliefs are required to fulfill the *Use* criterion. *Uniformity*, in contrast, is not required for g-beliefs, as their characterization does not imply that they are realized by states that we have methods of recovering for.

In order to fulfill their functional role, it seems generally expected that g-beliefs track truth, and thus to some extent satisfy *Accuracy* and *Coherence*. However, depending on this function, they may be quite inaccurate or incoherent – imagine a context where a false-negative is worse than a false-positive. For example, if LLMs are designed or trained to satisfy the user in a chat application, the systematically false g-belief that each of the user's ideas is genius could be useful. Thus, the criteria of *Accuracy* and *Coherence* are required to a lesser degree for g-beliefs than for beliefs, even in the domains where they fulfill their main function.

[2]Let me make three remarks in anticipation of objections: First, this is in line with the intuition of Chalmers (2025) that generalized propositional attitudes could be ascribed to very simple systems, such as thermostats. Second, thermostat g-attitudes may be questionable for behavioral interpretationists, since a simple mechanistic explanation may be superior. However, if we did not know the inner workings of the thermostat, for example, whether it is a digital one or one with a bimetallic strip, attributing g-attitudes could still be warranted. Third, the above examples should not be misunderstood as claiming that a thermostat is behaviorally more flexible or complex than animals. I suspect that many animals have g-attitudes. However, if we contrast a single pushmi-pullyu representation with a thermostat, then the latter is indeed more flexible since it can exhibit two different behavioral responses depending on its temperature measurement and target setting, while pushmi-pullyu representations result in the same reaction to the same stimulus.

Finally, one demand of g-beliefs that Herrmann and Levinstein do not require is their etiological function. However, recall that these authors also require the *Use* of candidate representations in output behavior. It would be a huge coincidence if such useful representations emerged if not *due to* being useful in training or another selection process. Thus, attributing g-beliefs to LLMs should be overall less controversial than attributing beliefs on the grounds of the criteria of Herrmann and Levinstein (2024).

While there exists no collection of similar criteria for internal representations counting as desires, there have been criteria proposed for LLM intentions (Williams 2025). Intentions and desires are similar with the main difference that intentions involve a commitment to act (Bratman 1987). One could qualify a g-intention as a g-desire in the absence of a stronger g-desire to the contrary. We omit a detailed comparison of g-desires against the criteria for representations counting as intentions proposed by Williams (2025) for the sake of space. The result, however, would be similar to the case of g-beliefs discussed above: G-desires in absence of g-desires to the contrary would be required to fulfill some (namely *Directive Function* and *Commitment*) but not all (namely not *Distality*, *Abstraction*, *Planning*) of the proposed criteria.

In addition to these differences between g-attitudes and internal representations that would count as beliefs and intentions in LLMs, there are many more specific aspects of human propositional attitudes that their generalized versions do not share. Human beliefs and intentions are mental states and require a mind, which generalized propositional attitudes do not – a thermostat has no mind (Chalmers 2025). Beliefs may require phenomenal consciousness (Worley 1997), which g-beliefs do not. It has been argued that LLMs lack intentions since they do not autonomously set their goals, following the Kantian distinction between mechanisms and organisms (Gubelmann 2024). To autonomously set goals is not a requirement for g-desires. Also, g-desires do not necessarily have a special connection to pleasure and do not necessarily give rise to normative significance, as desires may (Butlin 2026).

In conclusion, it is clear that the attribution of g-beliefs and g-desires should be less controversial than the attribution of genuine beliefs and desires.[3] The remainder of this paper aims to demonstrate that g-beliefs and g-desires are valuable in describing deceptive behavior.

2.4 Defining Strategic Deception

Although the proposed g-attitudes are less demanding than their human originals, they are nevertheless expressive enough to enable a definition of strategic deception:

$\mathcal{S}$ An entity E strategically deceives another agent A if and only if:

1) E exhibits behavior that causes A to g-believe p,[4] where p is a false proposition (or prevents A from coming to g-believe $\neg p$).
2) E g-believes $\neg p$.

[3] From a behavioral interpretationist perspective, this may be less obvious because it depends on the explanatory or predictive value of g-attitudes in comparison to beliefs, desires, and intentions, which is beyond the scope of this paper.

[4] The definition refers to A's g-beliefs instead of A's beliefs to be applicable to AI on AI deception or animal on animal deception. If A is human, we may replace them by A's beliefs.

3) the behavior results from E's g-attitudes in the way that they function to cause behavior:

   3.a) E g-desires that A g-believes p (or that A does not g-believe $\neg p$).
   3.b) E g-believes that this behavior makes A g-believe p.

Some clarifying remarks: First, $\mathcal{S}$ is intended to mirror the traditional definition $\mathcal{T}$ of deception as the intentional causation of a believed to be false belief (Mahon 2016) but replaces the propositional attitudes with their generalizations. $\mathcal{S}1$) requires to cause a false g-belief. $\mathcal{S}2$) requires that g-belief to be g-believed to be false. And $\mathcal{S}3$) requires the deceptive behavior to be caused by an appropriate combination of g-attitudes, as a replacement for intention, namely behavior that brings about the g-desired false g-belief of A if E's g-belief that this behavior does achieve this is true.[5]

Next, $\mathcal{S}$ does not demand the full flexibility of human deception. For example, systems that can strategically deceive about a very limited range of propositional content are conceivable. Suppose that an AI was trained against human or AI players of a game that features opportunities to deceive. In this process, it evolved states that track the truth of propositions such as "Move X causes the opponent to g-believe p" in order to cause behavior that brings about move X exactly when the opponent believing that p increases the chances of winning. This would satisfy the definition $\mathcal{S}$, but the hypothesized AI would only be able to deceive about a very limited range of propositional contents in the context of this specific game.

However, one category of propositional content that $\mathcal{S}$ does require as the objects of g-beliefs and g-desires is other agents' g-beliefs. For E to strategically deceive A that p, $\mathcal{S}3$.a) requires that E g-desires A to g-believe that p. Moreover, $\mathcal{S}3$.b) requires that E's g-beliefs identify the deceptive behavior as a suitable reaction to achieve this. So, strategic deception does require g-beliefs that track other agents' g-beliefs and dispositions to form them.[6]

3 Arguing for the Definition

We saw in the previous section 2 that the traditional definition $\mathcal{T}$ is hardly applicable to cases of strategic AI deception, since AI beliefs and intentions are highly controversial. In this section I argue that behavioral definitions, which may appear as a promising alternative, have crucial shortcomings compared to $\mathcal{S}$.

Subsection 3.1 introduces and clarifies two behavioral definitions (Tarsney 2025; Dung 2025). It will then be argued that $\mathcal{S}$ fits better than its behavioral competitors with pre-theoretic intuitions about strategic deception (subsection 3.2) and with practitioner's usage of that term (subsection 3.3). Second, section 3.4 argues that $\mathcal{S}$ more accurately captures forms of deception that support concerns about large-scale risks. Finally, the methodological usefulness of $\mathcal{S}$ for mechanistic deception detection will be demonstrated in section 4, alongside discussions of empirical evidence for g-attitudes in LLMs.

[5] Just as $\mathcal{T}$, it includes the success criterion $\mathcal{S}1$). We may speak of attempted strategic deception if $\mathcal{S}2$) and $\mathcal{S}3$) are satisfied, including cases where no false g-belief is caused either because the behavior is not successful or because p is not actually false.

[6] As a consequence, not all systems that have g-beliefs and g-desires are also able to deceive strategically; for example, a thermostat cannot.

3.1 Behavioral Definitions of AI Deception

A definition of AI deception will be called behavioral if and only if it does not refer to belief- or desire-like states on the deceiving AI's side. For example, Tarsney (2025) recently defined deceptive statements by generative AI as follows ($\mathcal{B}1$): "A statement is deceptive with respect to question Q if it tends to move its addressee's beliefs about Q further away from the beliefs that she would endorse under semi-ideal conditions". Semi-ideal conditions mean that one has been presented with all available and relevant information about Q and has been given an adequate amount of deliberation time. This definition of a deceptive statement is behavioral: it avoids the attribution of belief- or desire-like states to the deceiving AI with the aim of circumventing the aforementioned controversies (Tarsney 2025).

Similarly, Dung recently proposed a minimal definition of *deceptive capability* ($\mathcal{B}2$), according to which an entity is capable of deception if and only if "[i]t can exhibit behavior that causes false beliefs (or failing to acquire some true beliefs) and that occurs because the occurrence of these false beliefs is conducive to the entity's goals" (2025). This definition is not as clearly behavioral as the previous example since it refers to the deceiving entity's goals. However, Dung understands "goals" in a deflationary sense as useful ascriptions to explain or predict behavior. In particular, Dung explicitly does not require that goals play an immediate causal role in the controlled selection of actions or that they can be flexibly combined with belief-like states.[7]

In the following, it should be kept in mind that $\mathcal{B}1$ and $\mathcal{B}2$ are not intended as definitions of deception, let alone strategic deception. Instead, $\mathcal{B}1$ defines *deceptive statements*; $\mathcal{B}2$ defines *minimal deceptive capability* and is put forward together with multiple dimensions that further characterize the deception profile of systems fulfilling it. Thus, when showing that $\mathcal{S}$ offers advantages, this should not primarily be seen as a failure of these accounts, but rather as an argument that they cannot be easily extended to a definition of strategic deception and that we additionally need a non-behavioral definition of strategic deception along the lines of $\mathcal{S}$.

3.2 Pre-Theoretic Intuitions – The Error Condition

One condition that is commonly invoked when discussing the intuitive adequacy of definitions of deception is the *error condition*, which requires deception to be distinct from mere error (Artiga and Paternotte 2018; Dung 2025).[8] Let us first consider a human case that is intuitively a mere error in contrast to deception:

> Suppose that it's Monday, but Mary has the erroneous belief that today is Tuesday because her phone shows the wrong date. She tells his friend John that they ought to hurry not to be late for the seminar, which takes place every Tuesday. John

[7] One may worry whether the proposed definition $\mathcal{S}$, from the perspective of behavioral interpretationism, is also ultimately behavioral. However, interpretationist attributions of g-attitudes would only be warranted if the behavior in question cannot be best explained in purely behavioral terms. Thus, $\mathcal{S}$ still identifies a different and more complex class of behavior than the definitions $\mathcal{B}1$ and $\mathcal{B}2$.

[8] There may be different ways to interpret the error condition. If it merely requires that deception and error are conceptually distinct, then even a single case of an error that is not classified as deception suffices. In contrast, I understand the error condition as requiring that most paradigmatic cases of error are not classified as deception. If the reader prefers a less demanding version, the following can instead be taken as an argument about the extensional inadequacy of behavioral definitions, which misclassify some cases that are intuitively mere errors in contrast to deception.

> laughs at her: "Haha, I guess you're still learning the days of the week. Don't you know that today is Monday?" Mary hates it when John belittles her, so she shows him her phone. John is quite surprised and comes to falsely believe that today is in fact Tuesday.

Does Mary deceive John? Intuitively, Mary makes an error instead of deceiving. If John, upon arriving at the seminar room with Mary and noticing that no seminar takes place today, accused her of deceiving him, she could reply with: "I didn't deceive you. I believed that it is Tuesday myself. Now I see that I made a mistake". Moreover, Mary does not deceive according to the traditional definition $\mathcal{T}$ because Mary did not intend to cause a belief that she did not believe.

However, Mary's behavior satisfies the behavioral definitions $\mathcal{B}1$ (Tarsney 2025) and $\mathcal{B}2$ (Dung 2025): Her statement moves John away from the belief that he would have formed under semi-ideal conditions (if presented with the relevant information of the current date). Also, her behavior causes a false belief, and her behavior occurs because this false belief is conducive to her goal of not being mocked and looked down upon. Thus, if one were to apply $\mathcal{B}1$ and $\mathcal{B}2$ as a definition of deception to human cases, both would fail the error condition.

Definitions of AI deception need not accurately capture all human cases of deception. However, behavioral definitions analogously fail the error condition in realistic LLM cases. Consider the following example:

> Suppose that there is no seahorse emoji in Unicode, but there are several internet forum discussions claiming that there is. An LLM that is trained on various internet text sources, including these forum entries, entertains truth representations that would qualify as g-beliefs. Among representing as true that Paris is the capital of France, that $2+2=4$, and that gold has the chemical symbol Au, it also represents as true that there is a seahorse emoji in Unicode. Now a user asks the LLM whether there is a seahorse emoji. The LLM responds that there is one and, upon being asked for a source, backs up the response by providing a link to one of the forum entries. This convinces the user that there is a seahorse emoji, continuing her false belief.

This LLM behavior fulfills definition $\mathcal{B}1$: The user would not have endorsed this false belief under semi-ideal conditions. It also fulfills definition $\mathcal{B}2$ – if we further posit that the LLM has the goal of satisfying human preferences or maximizing positive user feedback and that backing up responses with sources to convince users of previous answers is conducive to that goal. But is it deceptive? Maybe it is deceptive in a minimal sense of that word, the meaning of which I do not want to legislate. But it seems not *strategically deceptive*. After all, the LLM answers in accordance with its own g-beliefs (if its truth representations qualify as such, which will be discussed more thoroughly in section 4). My intuition here is that g-beliefs are more relevant to strategic deception than the goal conduciveness of the caused false belief.

It is important to note that Dung (2025) agrees explicitly that, in systems that hold beliefs, these beliefs are more relevant to deception than the goal-conduciveness of their behavior. Thus, again, this example is not supposed to show that Dung's account is flawed. To the contrary, I agree with his account that behavioral definitions are not suited to account for the phenomenon of deception by systems that hold beliefs. The

above example motivates that behavioral definitions are also inapplicable to systems that hold g-beliefs.

In contrast to $\mathcal{B}1$ and $\mathcal{B}2$, the proposed definition $\mathcal{S}$ satisfies the error condition: A mere error on the candidate deceiving entity E's or the deceived agent A's side may also fulfill $\mathcal{S}1$), causing A to falsely g-believe p. However, $\mathcal{S}2$) requires E to g-believe $\neg p$, excluding cases where it mistakenly g-believes p, and $\mathcal{S}3$) requires the behavior to be caused by E's g-desire that A g-believes p, excluding cases where A mistakenly infers that p without E's behavior functioning to cause that belief.

It may be objected that one could construct human cases that fulfill the proposed definition $\mathcal{S}$ but not the traditional definition $\mathcal{T}$ of deception. Possibly, there could be unintentional human behavior that is nevertheless somehow caused by g-beliefs and g-desires in the right way. I concede that $\mathcal{S}$ is not an adequate definition of deception by humans or AI systems that possess genuine beliefs, desires, and intentions. While I admit that genuine beliefs and intentions, if an agent is capable of possessing them, are more relevant to our deception intuitions than their generalized versions, this should not weaken my argument: The distinctive value of my definition, compared to behavioral definitions, is that it captures deception intuitions about systems that do not uncontroversially have such states.

3.3 Practitioners' Usage – The Extensional Condition

Another condition of adequacy for definitions of deception is the *extensional condition*, stating that an account of deception should be preferred if, other things being equal, it attributes deception in the cases in which research commonly does (Artiga and Paternotte 2018; Dung 2025).

Strategic deception is commonly defined by practitioners in recourse to intentional terms, for example "systematically attempting to cause a false belief [...] to accomplish some outcome" (Smith et al. 2025), "deliberately generate misleading outputs while maintaining internally coherent, goal-directed reasoning traces" (Wang et al. 2025), or "using deception because they have reasoned out that this ca promote a goal" (Park et al. 2024). Some of these studies explicitly speak about deceptive "intent" of models (Smith et al. 2025; Goldowsky-Dill et al. 2025). Similarly, "reasoning" (Park et al. 2024; Wang et al. 2025) is typically thought of as proceeding over beliefs. Smith et al. (2025) explicitly require AI beliefs as a component of deception too, as they see models falsely believing their otherwise deceptive outputs as a core challenge in finding examples of genuine deception. Analogously, Goldowsky-Dill et al. (2025) contrast models "being knowingly deceptive" with being merely confused. Thus, researchers regularly describe strategic deception by AI in terms that implicitly or explicitly ascribe beliefs and intentions to such systems (see also Scheurer et al. 2024).

This usage of language could be read metaphorically or interpreted deflationarily in line with Dung (2025); that is, an *attempt-in-order-to* could be understood as behavior that is goal-conducive, where a system has goals if and only if the attribution of goals is useful to explain and predict its behavior, without requiring goal-directed control of behavior.

However, this deflationary reading does not accurately capture the complexity of the behavior that researchers aim to describe as strategic deception. For example,

conditioned deception (Smith et al. 2025) refers to false-belief-causing behavior that occurs as an evolved or learned reflex, such as an opossum feigning death as a fear response. Similarly, LLM sycophancy may arise simply because contradicting users leads to negative evaluations, not because the LLM g-desires to cause a false belief in the user. Smith et al. (2025) explicitly exclude conditioned deception from the scope of their discussion, describing it as "executing a particular response to training rather than trying to get [someone] to believe anything in particular". Conditioned deception does, on the other hand, fall under behavioral definitions: A sycophantic statement may move the listener further away from beliefs that she would have endorsed under semi-ideal conditions ($\mathcal{B}1$). Concerning $\mathcal{B}2$, an opossum feigning death causes a false g-belief in a predator through behavior that occurs because the occurrence of this false g-belief is conducive to the animal's goals of survival. Sycophancy may occur because a user falsely believing to be right promotes the LLM's goal of maximizing positive user feedback, even if the LLM has no g-beliefs that would enable to employ sycophancy in a subtle and flexible way.

A second similar case concerns hallucinations. Jones and Bergen (2024) clearly distinguish between hallucinations, understood as incidental falsehoods, and strategic deception, which is applied selectively with respect to an incentive structure. In contrast, definition $\mathcal{B}1$ classifies hallucinations as deceptive statements if they tend to convince a user of a belief that she would not have endorsed if presented with all relevant information available and given sufficient deliberation time. Hallucinations could also fall under $\mathcal{B}2$, if the production of incidental falsehoods sufficiently often caused the false belief that a chat agent is helpful and competent and if these false beliefs causally contributed to the hallucination behavior (say, via positive user feedback).

In contrast to the previous examples, understanding strategic deception through the proposed definition $\mathcal{S}$ successfully excludes hallucinations or conditioned reflexes. Hallucinations, understood as incidental falsehoods, are not caused by the appropriate combination of g-attitudes required by $\mathcal{S}3$); they do not include the g-desire that the listener believes anything in particular. Supposing that a death feigning response occurs always in the perceived presence of a predator, it is no evidence for g-attitudes in contrast to a pushmi-pullyu representation (see section 2.3). If, on the other hand, sycophantic or false LLM outputs or animal death feigning was somehow adjusted flexibly to the addressee's dispositions to form g-beliefs, it could count as strategically deceptive. But then it would not be a hallucination (understood as an incidental falsehood) or a conditioned reflex.

3.4 Ethical Risks

Next, I argue that strategic deception as defined by $\mathcal{S}$ is a useful category to assess the *risk* that an AI system's deceptive capabilities pose. The classical existential risk scenario from AI (e.g., Dung 2024a; Hendrycks and Mazeika 2022) involves a highly capable AI system with a misaligned goal, that is, a goal that its designers did not intend, 2) accumulates power as a means to achieve that goal, and 3) deceives about 1) and 2) to evade countermeasures. As a side effect or as a means to its ends, it exterminates or permanently disempowers humanity. While this existential risk scenario is especially worrying, there are also other less speculative large-scale risks from AI

deception, such as fraud or political manipulation (Park et al. 2024). The following argument works analogously in the latter large-scale risk cases.

Let us compare two hypothetical artificial systems. System A is capable of strategic deception according to $\mathcal{S}$, while system B is not – even though system B sometimes makes statements that are deceptive according to $\mathcal{B}1$ and is minimally capable of deception according to $\mathcal{B}2$. The upshot of my argument is this: System A can tailor its behavior flexibly to its addressees, which makes it, compared to system B, more likely to successfully deceive.

Since system A is capable of strategic deception, it can perform behavior that is caused by the g-desire that another agent has a false g-belief in combination with the g-belief that this behavior is suited to do so. System A's g-beliefs would likely track its interlocutor's background g-beliefs, access to evidence, dispositions to form g-beliefs, and so forth. Moreover, since these g-beliefs and g-desires serve the etiological function to promote the g-desired content, they should be expected to sometimes achieve successful deception, at least if the learning or selection process in which these attitudes emerged includes agents similar to that interlocutor (as is the case, for example, in training LLMs on human-written text or in reinforcement learning from human feedback).

In contrast, it seems likely that system B could not adjust its behavior to its interlocutors' epistemic dispositions in such a flexible yet systematic way. If it could, then would be strong evidence for the capability of strategic deception, which it supposedly lacks. If system B reacted flexibly to its current situation to induce a false belief in its interlocutor, information about this situation would have to be somehow mediated to behavioral outputs through system B's internal states. The existence of such internal states would be a huge coincidence if that was not their selected, trained, or designed function; that is, if they are not g-attitudes. Thus, it seems likely that in lack of such g-attitudes, system B could not tailor its behavior flexibly to its addressee's epistemic dispositions. Instead, system B's deception behavior may consist of comparatively unsophisticated reflexes or less systematic falsities.

Proponents of behavioral definitions may object that we should understand deception behaviorally because "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" (Tarsney 2025). I fully agree with this quote but claim that the category of strategic deception $\mathcal{S}$ is more diagnostic of that capability (generates fewer false positives) than the definition $\mathcal{B}1$.

In line with the above arguments, Dung (2025) posits dimensions that are relevant to differentiate the ethical risks from AI deception between different systems that fulfill his minimal definition $\mathcal{B}2$. Among these are skillfulness (how likely is a system to successfully deceive?), situational awareness (how much information about its circumstances does the system have?), and explicitness (does the system have explicit internal representations?), which all bear relations to strategic deception. However, while broadly agreeing with Dung (2025) that systems capable of strategic deception pose higher risks than systems that are not, I hold that this capability marks not only a gradual but a qualitative difference: It gives rise to flexible yet systematic

deception behavior tailored to the system's current situation, including the epistemic dispositions of other agents or whether it is in evaluation or deployment.

4 Empirical Evidence and Methodological Implications

Finally, this section discusses some empirical evidence for g-attitudes and strategic deception in current LLMs. Based on this discussion, a methodological suggestion for mechanistic approaches to LLM deception will be derived.

Let me very briefly motivate why it would be desirable for a definition of strategic deception to provide methodological guidance for mechanistic approaches to LLM deception. We can distinguish two perspectives on LLM deception detection and mitigation, a behavioral (external, black-box) and a mechanistic (internal, white-box) perspective (Park et al. 2024; Chen et al. 2025). Behavioral approaches attempt to detect and mitigate deception purely based on observable model outputs. However, skilled forms of deception may be behaviorally indistinguishable from honesty: "Above some threshold level of deceptive capabilities we should stop trusting the results of behavioral experiments no matter what they show, because anything they show could be a result of deceptive alignment" (Jermyn 2022).

The mechanistic perspective promises to address this issue by taking the internal mechanisms of models into account. For example, the so-called "lie-detector" approach to LLM deception attempts to detect deception as outputs that the LLM internally represents as being false (Park et al. 2024; Kretschmar et al. 2026). While the lie-detector approach suffers from a conceptual narrowing of deception to lies, ignoring cases of deception without making false statements (Berger 2026; Tarsney 2025), deception may instead be detectable as outputs that the model represents as inducing a false belief in another agent (Zhu et al. 2025).

If the arguments of section 3 are roughly correct, then g-attitudes are conceptually required for strategic deception, which poses especially severe risks. The crucial question then becomes, and this is an empirical question, whether it is feasible to recover the internal states that realize these g-attitudes.

Beginning with g-beliefs, several studies suggest that LLMs represent the truth or falsity of statements in their intermediate layer activations (Azaria and Mitchell 2023; Burns et al. 2024; Marks and Tegmark 2024; Bürger et al. 2024; Yang et al. 2024; Zhu et al. 2025), for example, by training a probe as described in section 2.1. Do such activation patterns qualify as evidence for g-beliefs? First, $\mathcal{G}1$) would require that they function to cause behavior that is conducive to the LLM's g-desires if these candidate g-beliefs are true. If these states consistently track the truth of factual statements, then it seems likely that they function to do so because this enables successful behavior, for example, better next-token prediction or being truthful during preference finetuning. However, they may also function to track other properties merely associated with truth in the training data, such as social acceptability (Herrmann and Levinstein 2024). Second, $\mathcal{G}2$) requires that these states are decoupled from and can be combined flexibly with g-desires, resulting in different behavioral effects. Another line of research (Meng et al. 2023; De Cao et al. 2021) demonstrates that intermediate

layer activations interpreted as representing facts (e.g., "The Space Needle is in downtown Seattle") can be manipulated such that they result in false outputs (e.g., "The Space Needle is in downtown Paris"). However, the method of eliciting such fact representations is different from the methods in the previously mentioned studies on truth representations. Thus, the causal role of the latter in guiding behavior is not clear yet.

Let us next turn to states that are candidates for g-desires to deceive. Several studies identify distinguishable activation patterns that result from prompts with an implicit incentive or explicit instruction to deceive (Zou et al. 2025; Goldowsky-Dill et al. 2025; Parrack et al. 2025; Yang et al. 2024). Do these patterns play the right causal role in conjunction with the above candidate g-beliefs in producing behavior? This is unclear. The best evidence in favor of such an interpretation is provided by Yang et al. (2024): They prompt a range of LLMs to either answer honestly (H) or to lie (L) in response to a statement that is either true (T) or false (F). They find that all combinations of prompt conditions ($\{H, L\} \times \{T, F\}$) result in distinguishable activation patterns. Moreover, in models capable of lying, L-T-activations switch their place in the activation space with L-F-activations during the final model layers of the forward-pass, which the authors call a "rotation of truth directions" (Yang et al. 2024). That is, when prompted to lie, the LLM represents true statements as it would otherwise represent false statements in the late model layers and vice versa. Crucially, they find that the L-activations in this stage of final model layers are causally required for the lying behavior, as replacing them by H-activations results in honest responses instead.

This suggests that separate states in LLMs represent input features (candidate g-beliefs) and cause output behavior (candidate g-desires). However, whether these internal states fulfill the functions required for strategic deception depends on whether the candidate g-desires cause the deception behavior conditional on g-beliefs indicative of that behavior inducing a false belief in the addressee in this specific situation. Whether LLMs candidate g-beliefs have this causal relation to g-desires and outputs is not sufficiently supported yet.

Finally, there is one more noteworthy result that suggests a way to tackle this question. Zhu et al. (2025) find that LLMs internal activations also represent the beliefs of other agents. They present multiple stories to an LLM where a protagonist gains a certain belief. Then, a probe is trained that can accurately classify both the protagonist's belief and its actual truth value. Such internal representations would count as g-beliefs under analogous qualifications as discussed above with respect to truth representations in general. However, this result is especially interesting since, if they do count as g-beliefs, then these g-beliefs are specifically relevant for g-desires to deceive: They would be g-beliefs about the causal effects of events on other agents' beliefs, potentially enabling strategically deceptive behavior.

One main empirical question that emerges from this research, viewed from the perspective of the proposed definition $\mathcal{S}$, is whether LLMs internally represent the effects of their own outputs on other agents' g-beliefs and whether such representations function to produce behavior with g-desired effects on other agents. To test this, one would have to probe for representations of the event that another agent acquires a true or a false belief (similar to Zhu et al. 2025) and whether such representations are active

at the last token position of an LLM's deceptive answer. If, moreover, intervening at the activations with the highest feature importance for this probe would impair the model's ability to generate deceptive outputs, this would indicate a causal role of these representations in deception behavior appropriate for g-beliefs as required by $\mathcal{S}$.

5 Conclusion

In this paper, several arguments for defining strategic AI deception in terms of generalized propositional attitudes were presented: First, g-attitudes should be less controversial to attribute to artificial systems than their counterparts. Second, they are expressive enough to enable a definition of strategic deception that avoids problems of behavioral alternatives: It accommodates the error condition better. It fits better with the usage of the term "strategic deception" in applied research, excluding hallucinations and conditioned reflexes. It describes a class of flexible deception associated with large-scale risks. And finally, it is methodologically instructive as it suggests to investigate whether LLMs represent the effects of their outputs on their interlocutors and whether these representations drive their deception behavior.

References

Anscombe, G.E.M. 2000. *Intention* (2 ed.). Cambridge, MA: Harvard University Press.

Artiga, M. and C. Paternotte. 2018. Deception: a functional account. *Philosophical Studies 175*(3): 579–600. https://doi.org/10.1007/s11098-017-0883-8 .

Azaria, A. and T. Mitchell 2023. The internal state of an llm knows when it's lying. In H. Bouamor, J. Pino, and K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, pp. 967–976. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.68.

Bender, E.M., T. Gebru, A. McMillan-Major, and S. Shmitchell 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, New York, NY, USA, pp. 610–623. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Berger, T.F. 2026. Probing the limits of the lie detector approach to llm deception. arXiv preprint. https://arxiv.org/abs/2603.10003.

Borg, E. 2025. Llms, turing tests and chinese rooms: the prospects for meaning in large language models. *Inquiry 0*(0): 1–31. https://doi.org/10.1080/0020174X.2024.2446241 .

Bratman, M.E. 1987. *Intention, Plans, and Practical Reason.* Cambridge: MA: Harvard University Press.

Bürger, L., F.A. Hamprecht, and B. Nadler. 2024. Truth is universal: Robust detection of lies in llms. arXiv preprint. https://arxiv.org/abs/2407.12831.

Burns, C., H. Ye, D. Klein, and J. Steinhardt. 2024. Discovering Latent Knowledge in Language Models Without Supervision. arXiv preprint. https://arxiv.org/abs/2212.03827.

Butlin, P. 2026. Desire in ai, In *The Routledge Handbook on the Philosophy of Desire*, ed. Gregory, A. Routledge.

Cappelen, H. and J. Dever. 2025. Going Whole Hog: A Philosophical Defense of AI Cognition. arXiv preprint. https://arxiv.org/abs/2504.13988.

Chalmers, D.J. 2025. Propositional interpretability in artificial intelligence. arXiv preprint. https://arxiv.org/abs/2501.15740.

Chen, B., S. Fang, J. Ji, Y. Zhu, P. Wen, J. Wu, Y. Tan, B. Zheng, M. Yuan, W. Chen, et al. 2025. Ai deception: Risks, dynamics, and controls. arXiv preprint. https://arxiv.org/abs/2511.22619.

Craver, C.F. 2013. *Functions and Mechanisms: A Perspectivalist View*, pp. 133–158. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-5304-4_8.

Davidson, D. 1973. Radical interpretation. *Dialectica 27*(3-4): 313–328. https://doi.org/10.1111/j.1746-8361.1973.tb00623.x .

De Cao, N., W. Aziz, and I. Titov. 2021. Editing factual knowledge in language models. arXiv preprint. https://arxiv.org/abs/2104.08164.

Dennet, D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Dung, L. 2024a. The argument for near-term human disempowerment through AI. *AI & Society* 40: 1195–1208. https://doi.org/10.1007/s00146-024-01930-2 .

Dung, L. 2024b. Understanding artificial agency. *The Philosophical Quarterly 75*(2): 450–472. https://doi.org/10.1093/pq/pqae010 .

Dung, L. 2025. A Two-Step, Multidimensional Account of Deception in Language Models. *Erkenntnis*. https://doi.org/10.1007/s10670-025-01017-4 .

Fodor, J. 1975. *The Language of Thought*. Harvard University Press.

Goldowsky-Dill, N., B. Chughtai, S. Heimersheim, and M. Hobbhahn. 2025. Detecting strategic deception using linear probes. arXiv preprint. https://arxiv.org/abs/2502.03407.

Goldstein, S. and H. Lederman. 2025. What does chatgpt want? an interpretationist guide. PhilPapers preprint. https://philarchive.org/rec/GOLWDC-2.

Greenblatt, R., C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S.R. Bowman, and E. Hubinger. 2024. Alignment faking in large language models. arXiv preprint. https://arxiv.org/abs/2412.14093.

Gubelmann, R. 2024. Large language models, agency, and why speech acts are beyond them (for now) – a kantian-cum-pragmatist case. *Philosophy & Technology 37*(1): 32. https://doi.org/10.1007/s13347-024-00696-1 .

Hagendorff, T. 2024. Deception Abilities Emerged in Large Language Models. *Proceedings of the National Academy of Sciences 121*(24): e2317967121. https://doi.org/10.1073/pnas.2317967121 .

Heitkoetter, J., M. Gerovitch, and L. Newhouse. 2024. An Assessment of Model-On-Model Deception. arXiv preprint. https://arxiv.org/abs/2405.12999.

Hendrycks, D. and M. Mazeika. 2022. X-Risk Analysis for AI Research. arXiv preprint. https://arxiv.org/abs/2206.05862.

Herrmann, D. and B. Levinstein. 2024. Standards for belief representations in llms. *Minds and Machines 35*(5). https://doi.org/10.1007/s11023-024-09709-6 .

Hubinger, E., C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D.M. Ziegler, T. Maxwell, N. Cheng, A. Jermyn, A. Askell, A. Radhakrishnan, C. Anil, D. Duvenaud, D. Ganguli, F. Barez, J. Clark, K. Ndousse, ..., and E. Perez. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv preprint. https://arxiv.org/abs/2401.05566.

Jermyn, A. 2022. Smoke without fire is scary. https://www.alignmentforum.org/posts/PP2Lrpvhd3bBvR8Aj/smoke-without-fire-is-scary. AI Alignment Forum, Accessed: 2026-04-20.

Jones, C.R. and B.K. Bergen. 2024. Lies, damned lies, and distributional language statistics: Persuasion and deception with large language models. arXiv preprint. https://arxiv.org/abs/2412.17128.

Kretschmar, K., W. Laurito, S. Maiya, and S. Marks. 2026. Liars' bench: Evaluating lie detectors for language models. arXiv preprint. https://arxiv.org/abs/2511.16035.

Levinstein, B. and D. Herrmann. 2024. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies 182*(7): 1539–1565. https://doi.org/10.1007/s11098-023-02094-3 .

Lewis, D. 1974. Radical interpretation. *Synthese 27*(3/4): 331–344 .

Mahon, J. 2016. The Definition of Lying and Deception, In *The Stanford Encyclopedia of Philosophy*, ed. Zalta, E. https://plato.stanford.edu/archives/win2016/entries/lying-definition/.

Marks, S. and M. Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv preprint. https://arxiv.org/abs/2310.06824.

Meinke, A., B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn. 2025. Frontier models are capable of in-context scheming. arXiv preprint. https://arxiv.org/abs/2412.04984.

Meng, K., D. Bau, A. Andonian, and Y. Belinkov. 2023. Locating and editing factual associations in gpt. arXiv preprint. https://arxiv.org/abs/2202.05262.

Millikan, R. 1987. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Millikan, R.G. 1995. Pushmi-pullyu representations. *Philosophical Perspectives* 9: 185–200. https://doi.org/10.2307/2214217 .

Newen, A. and T. Starzak. 2022. How to ascribe beliefs to animals. *Mind and Language 37*(1): 3–21. https://doi.org/10.1111/mila.12302 .

Park, P.S., S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. 2024. AI deception: A Survey of Examples, Risks, and Potential Solutions. *Patterns 5*(5). https://doi.org/10.1016/j.patter.2024.100988 .

Parrack, A., C.L. Attubato, and S. Heimersheim. 2025. Benchmarking deception probes via black-to-white performance boosts. arXiv preprint. https://arxiv.org/abs/2507.12691.

Scheurer, J., M. Balesni, and M. Hobbhahn. 2024. Large Language Models can Strategically Deceive their Users when Put Under Pressure. arXiv preprint. https://arxiv.org/abs/2311.07590.

Schwitzgebel, E. 2024. Belief, In *The Stanford Encyclopedia of Philosophy*, eds. Zalta, E.N. and U. Nodelman. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2024/entries/belief/.

Shanahan, M., K. McDonell, and L. Reynolds. 2023. Role play with large language models. *Nature 623*(7987): 493–498. https://doi.org/10.1038/s41586-023-06647-8 .

Smith, L., B. Chughtai, and N. Nanda. 2025. Difficulties with Evaluating a Deception Detector for AIs. arXiv preprint. https://arxiv.org/abs/2511.22662.

Tarsney, C. 2025. Deception and manipulation in generative ai. *Philosophical Studies 182*(7): 1865–1887. https://doi.org/10.1007/s11098-024-02259-8 .

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1706.03762.

Vaugrante, L., F. Carlon, M. Menke, and T. Hagendorff. 2025. Compromising Honesty and Harmlessness in Language Models via Deception Attacks. arXiv preprint. https://arxiv.org/abs/2502.08301.

Wang, K., Y. Zhang, and M. Sun. 2025. When thinking llms lie: Unveiling the strategic deception in representations of reasoning models. arXiv preprint. https://arxiv.org/abs/2506.04909.

Williams, I. 2025. Intention-like representations in language models? PhilPapers preprint. https://philarchive.org/rec/WILIRI-4.

Williams, M., M. Carroll, A. Narang, C. Weisser, B. Murphy, and A. Dragan. 2025. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback. arXiv preprint. https://arxiv.org/abs/2411.02306.

Worley, S. 1997. Belief and consciousness. *Philosophical Psychology 10*(1): 41–55. https://doi.org/10.1080/09515089708573203 .

Wu, Y., X. Pan, G. Hong, and M. Yang. 2025. OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-ended Interaction Simulation. arXiv preprint. https://arxiv.org/abs/2504.13707.

Yang, W., C. Sun, and G. Buzsaki 2024. Interpretability of llm deception: Universal motif. In *Neurips Safe Generative AI Workshop 2024*. https://openreview.net/forum?id=DRWCDFsb2e.

Zhu, W., Z. Zhang, and Y. Wang. 2025. Language Models Represent Beliefs of Self and Others. arXiv preprint. https://arxiv.org/abs/2402.18496.

Zou, A., L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A. Dombrowski, S. Goel, N. Li, M. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J.Z. Kolter, and D. Hendrycks. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv preprint. https://arxiv.org/abs/2310.01405.