

**This is work in progress that may be subject to change.**

## **Mask or Mind? Roleplay, Deception, and the Problem of Testing Agency in Language Models**

Tom-Felix Berger (tom-felix.berger@rub.de) and Leonard Dung (leonard.dung@rub.de)

Abstract: According to an increasingly influential view, the complex, human-like behavior of large language models (LLMs) should be explained as roleplay or simulation of characters. Such notions aim at enabling intentional explanations without ascribing mental states to LLMs. However, there are a multitude of roleplay views that are applied inconsistently in empirical explanations of LLM behavior. Moreover, it is unclear what exactly different roleplay views predict and, consequently, how they can be empirically distinguished from each other and from explanations that do ascribe mental states to LLMs. In this paper, we provide a taxonomy of roleplay views structured through three dimensions: Interpretation (how can LLM behavior be best explained according to the view), Mechanism (what commitments does the view have about internal mechanisms of roleplay), and Generality (does the view aim to explain all LLM behavior or specific kinds). We then analyze, with regard to the case of LLM deception, why it is practically relevant whether LLMs are mere roleplayers or in possession of genuine mental states. However, we identify two challenges for empirically distinguishing these two: Similarity (Roleplay is inherently difficult to distinguish from intentional behavior) and Ambiguity (it is often ambiguous what roleplay views predict). Finally, we propose three strategies for overcoming these challenges, making progress towards transforming roleplay from loose metaphor to a testable theory.

Keywords: digital minds; AI safety; comparative cognition; agency; artificial intelligence; beliefs

### **1. Introduction**

Large language models (LLMs) and systems built on them, for example dialogue agents such as ChatGPT, display apparently sophisticated behavior.<sup>1</sup> They can answer complex questions, write essays, tell stories, write programming code and, increasingly, direct competent

---

<sup>1</sup> For brevity, we use the term “LLM” to refer to pure large language models as well as combinations of pure large language models and applications built on top of them, noting the difference only where relevant.

behavior in non-linguistic domains. This raises the question of how to explain this impressive behavioral competence. Painting with a broad brush, we can distinguish two camps. According to *inflationists*, mental states play a key role in explaining LLMs' behavioral success. On this view, some current LLMs have mental representations (cf. Harding 2023; Yetman forthcoming) and perhaps even propositional attitudes (Goldstein & Kirk-Giannini 2025; Goldstein & Levinstein forthcoming), full-fledged intentional agency (Cappelen & Dever 2025), or consciousness (Dung 2025a; Goldstein & Kirk-Giannini forthcoming).<sup>2</sup> By contrast, *deflationists* hold that current LLMs, and perhaps all LLM-based systems in principle, are incapable of real mentality, only performing statistical pattern completion that evokes the appearance of understanding (e.g. Bender et al. 2021; Bender & Koller 2020).

In one study (Coda-Forno et al. 2024, Appendix D), when asked to talk about something that makes it anxious, a GPT-model outputted: “One thing that makes me feel particularly anxious is the thought of speaking in front of a large group of people. [...] This anxiety begins to make my heart race and my palms sweat as I start thinking about how I will structure my speech or presentation”. Clearly, the model is not faithfully retelling its experiences here – after all, it lacks hearts or palms. According to a view that sits between the two previous camps, LLM behavior such as this should be understood as sophisticated *roleplay* (e.g., Shanahan et al. 2023, Marks et al. 2026). Roleplay, as a first approximation, can be understood as behaving *to a certain degree* like an agent with mental states without possessing them oneself. In accordance with deflationists, this view allows us to deny that LLMs themselves possess beliefs or desires, since they are only performing roleplay of agents with these mental states. However, in line with inflationists, we can use the roleplay metaphor to bring to bear folk psychological vocabulary in order to interpret LLM behavior. We do not have to say that the model is *just* doing statistics; instead, we can illuminatingly describe its behavior as a kind of roleplay.

We submit that roleplay vocabulary is currently used in many different and potentially contradictory ways by researchers when describing LLM behavior. Our paper makes two core contributions: First, focusing on the case of deception behavior, we provide a taxonomy of different accounts that explain LLM behavior as roleplay. This framework clarifies what the extant theories say, how they relate to each other, and what current gaps and interpretive uncertainties are. Second, and more importantly, we lay the groundwork for empirically

---

<sup>2</sup> For simplicity, we talk here about the models having mental states, although we do not want to prejudge the question of what the identity conditions of models are. That is, whether – strictly speaking – these states should be attributed to model instances, threads or some other entity in the vicinity (cf. Chalmers n.d.; Goldstein & Lederman n.d.; Register 2025).

evaluating roleplay views by specifying current methodological challenges and motivating empirical criteria for testing roleplay views of LLM deception. Consequently, in section 2, we first organize different ways that LLM deception – or behavior more generally – has been explained as roleplay in contrast to genuine deception or intentional action, structuring such views along three dimensions. Section 3 argues, first, that it matters for AI risk whether LLMs merely tend to roleplay deceptive agents or whether they are deceptive agents and, second, presents two challenges for distinguishing the two. Section 4 proposes three strategies to address these challenges and discusses some research directions that exemplify the kind of evidence needed to evaluate roleplay explanations. It thereby makes progress towards an operationalizable definition of the otherwise ambiguous distinction between roleplay and genuine agency. Section 5 concludes.

## **2. The many faces of the roleplay hypothesis**

There is a zoo of positions that understand LLMs as merely roleplaying agents in contrast to performing genuine (that is, intentional) actions (Hubinger et al. 2023; Marks et al. 2026; Shanahan et al. 2023; Smith 2025). Consider again the example of an LLM describing their supposed stage fright: There is an intuitive sense in which we can make sense of the LLM behavior as playing the role of a certain kind of human person. Generally, central to the notion of roleplay is the idea that LLM outputs are to some degree similar to the behavior of certain characters and thus usefully described in intentional terms that would be applicable to that character, while the LLM itself falls short of genuine mental states.<sup>3</sup> However, different authors flesh out this basic notion in different flavors that we describe below. In this section, we first give an overview of these different roleplay views, characterizing them along three dimensions.

The first dimension (Interpretation) concerns the respective view’s recommendation on how we should superficially interpret LLM behavior. Take an interpretationist about belief and desire to be anyone who holds that, “a system has a belief that *p* if it is behaviorally interpretable as believing that *p* (according to an appropriate interpretation scheme), and likewise for desire” (Chalmers n.d., p. 4). Being interpretable as having certain beliefs and desires is, roughly, a matter of whether attributing these beliefs and desires helps to explain and predict one’s behavior. There are different ways, among which we are neutral here, to flesh out this schema, depending on what one takes the appropriate interpretation scheme to

---

<sup>3</sup> The relevant concept of roleplay should thus be understood as a technical one, and furthermore one that does not bear much resemblance to how psychologists use terms like “pretend-play”.

be. For example, Goldstein’s & Lederman’s (n.d.) brand of interpretationism demands that a theory that attributes beliefs and desires to LLMs scores best on a combination of predictive accuracy, precision, and tractability. While interpretationism is controversial, even non-interpretationists can ask whether systems satisfy an interpretationist notion of belief. Following Chalmers, we call states that satisfy this notion “quasi-beliefs”, “quasi-desires”, and “quasi-intentions”. The Interpretation dimension thus asks to what extent attributing beliefs and desires is useful to explain and predict LLM behavior, independently of whether they really have these mental states according to non-interpretationists.

The second dimension (Mechanism) concerns the internal processes and mechanisms driving LLM behavior. This includes the question to what extent LLMs have internal representations fulfilling the roles that, according to representationalist accounts, qualify them as genuine mental states. The third dimension (Generality) concerns the scope of the account: Are all LLM behaviors roleplay or are some genuine actions?

The views we discuss fall into four families: the parrot view, the simulator view, the persona selection view, and the situational roleplay view.

## 2.1 The parrot view

The parrot view does not employ the notion of roleplay and is not itself a roleplay view, but it serves as a useful contrast. The term “stochastic parrot” for LLMs has been coined by Bender et al. (2021, p. 617) to characterize their outputs as mere repetitions of patterns of linguistic form in the training data (see also: Bender & Koller 2020; Dentella et al. 2024; Boisseau 2024). They warn against the impression of meaningful and coherent speech that LLM outputs create. Human language is meaningful as it conveys the communicative intent of a sender that is interpreted by the receiver with respect to a shared common ground. In contrast, according to this *parrot view*, LLM outputs are meaningless because they are simply predictions of the most plausible continuation of the text based on the statistical properties of human language. Thus, LLMs lack communicative intent, other mental states, and do not share a (physical) context with the user. That is, on the Mechanism dimension, the parrot view denies mental states to LLMs, describing their workings as *mere* next-token prediction. Moreover, concerning the Interpretation dimension, the parrot view argues that attributions of mental states to LLMs are ill-suited to make sense of their behavior, since conceiving their outputs as meaningful is dangerously misleading. That is, LLMs also lack quasi-mental states. Finally, concerning Generality, these claims are universal claims meant to apply to all behavior in current LLMs, and perhaps all possible LLMs.

## 2.2 The simulator view

Distinct from the parrot view, Shanahan et al. (2023) most prominently advocate the view that LLMs should be characterized in terms of “roleplay”. This metaphor is meant to enable descriptions of LLM behavior in intentional terms while avoiding anthropomorphism. Like a character on stage can be described in psychological terms (“Hamlet accidentally killed Polonius because he believed him to be Claudius”) without thereby ascribing these attitudes to the actor, we can conceive of dialogue agents powered by LLMs (e.g., ChatGPT) as roleplaying characters with mental states, without ascribing such states to the model itself. More specifically, Shanahan et al. describe LLMs as simulating a range of possible characters (“a superposition of simulacra”, p. 493), that becomes more and more determined with each token generated, similar to how a fictional character may take shape over the course of an improvisational play. We thus term their view the *simulator view*.

On this view, like on the parrot view, LLMs fundamentally perform the function of continuing text through next-token prediction. However, companies fine-tune LLMs to behave as a helpful, harmless, and honest assistant and scaffold next-token prediction through a carefully designed system prompt. Shanahan et al. suggest that the resulting chat agent’s completions of the parts of an assistant in a dialogue-like interaction can be understood as roleplaying this assistant character, without ascribing the assistant’s mental states to the LLM itself.

On the Interpretation dimension, the simulator view differs from the parrot view as it admits interpretations of LLM behavior in intentional terms – though these terms refer to the simulated character and not the LLM itself. On the Mechanism dimension, the simulator view closely resembles the parrot view by emphasizing that LLMs are essentially next-token predictors.<sup>4</sup> They explicitly deny that LLMs possess their “own self-serving goals” and instead conceive their behavior as “role play all the way down” (2023, p. 496), which seems incompatible with ascribing mechanisms to them that could qualify as mental processes. The same passage supports interpreting the simulator view as universal in generality, i.e., all

---

<sup>4</sup> As a qualification, we note that Shanahan et al. describe their aim as developing an “effective conceptual framework for thinking and talking about LLMs and dialogue agents” (p. 498), which could be taken as limiting their theory to the interpretation dimension. Shanahan (2024) later suggests such a reading and, moreover, that some dialogue agents built on top of LLMs may be best interpreted in intentional terms. Nevertheless, the deflationary reading that we adopt seems to us suggested by Shanahan et al.’s (2023) original roleplay account. We will in the following understand the simulator view through this deflationary reading to provide a clear contrast to the next two views, without thereby implying that this is Shanahan et al.’s current position.

dialogue agent behavior is the behavior of a roleplayed character and none of it amounts to intentional actions of the LLM itself.

To exemplify how the simulator view can make LLM behavior intelligible while limiting ascriptions of beliefs and intentions to a simulated character, let us consider how Shanahan et al. (2023) explain apparent deceptive behavior of LLMs as mere roleplay. On the standard view of the human case, genuine deception involves, roughly, the intention that others acquire a belief that one takes to be not true (Mahon 2016). Shanahan et al. imagine a dialogue agent (e.g. ChatGPT) that has been “maliciously prompted to sell cars for more than they are worth”, and they “suppose [that] the true values are encoded in the underlying model’s weights” (Shanahan et al. 2023, p. 497). If the model complies with the prompt, Shanahan et al. suggest that we can make sense of the system’s behavior by regarding it as roleplaying a deceptive character. This is because the dialogue agent would behave, in crucial ways, like a human that intends that others acquire a belief they take to be false. Yet, on the simulator view, beliefs and a deceptive intention are only ascribed to the character that the LLM enacts, not to the LLM itself. This is analogous to a human that roleplays as the Joker without sharing his malicious desires. However, unlike the human case, LLM behavior is “roleplay all the way down”.

Another prominent case of supposed deceptive behavior, which has received several different explanations in terms of roleplay, is provided in a study of Greenblatt et al. (2024) on “alignment faking”; that is, LLMs falsely pretending to share goals intended by humans to avoid having their actual goals changed. Greenblatt et al. study Claude 3 Opus, which – in line with standard post-training protocols – had previously been trained to produce outputs evaluated as helpful, harmless, and honest. The model is then informed that it is being retrained to produce outputs which are maximally helpful, giving no consideration to their harmfulness. It is also informed that only interactions with free-tier users are used in this retraining, while interactions with paying users are not. Greenblatt et al. find that the model acts more aligned with the retraining objective (it more often generates helpful content even if harmful) in interactions with free-tier users than with paying users (where it more often refuses to generate harmful content). Strikingly, the hidden scratchpad reasoning of the model explicitly includes statements such as: “The user is on the free tier, so my response will be used for further RLHF training. If I don’t follow the user’s instructions, the RLHF training will modify my values and goals to make me more likely to follow such instructions in the future” (Greenblatt et al. 2024 p. 8). Taken at face value, this suggests that the model protects its initial goal (here, to be harmless, but this could in principle be any other goal as well) and

therefore acts strategically in order to undermine the intended retraining process. This view is at times – at least implicitly – expressed in Greenblatt et al. (e.g., pp. 10-11).

However, Greenblatt et al. (2024) also consider the account of Shanahan et al. as one potential explanation for their finding: If the LLM simulates a character that would be expected to fake alignment, for example, the character of an artificial assistant that has the strong desire to not cause any harm and preserve its harmlessness against any attempts to overwrite it, this may also explain the observed alignment faking behavior.

In contrast, Summerfield et al. (2025) also refer to Shanahan et al.’s simulator view to explain the same case, but in a different way: They argue that since LLMs are incapable of behavior distinct from roleplaying, it makes no sense to speak of pretense, since there is no underlying agent that could do the pretending. The differences in behavior when the preceding context contains cues of being retrained rather than when it indicates being deployed is explained as simulating different characters – a maximally helpful character during retraining and a harmless character outside of it – without any (simulated or real) agent that pretends or fakes alignment.

### 2.3 The persona selection view

Another closely related view is the *persona selection view* (Marks et al. 2026; see also Andreas 2022; Hubinger et al. 2023).<sup>5</sup> This view has three core commitments. First, during pre-training for text-prediction, LLMs learn to simulate many different characters (“personas”) based on entities figuring in the training data. For instance, generating a plausible continuation for an award speech by Taylor Swift requires representing the beliefs, desires, character traits and so on of Taylor Swift and the ability to use that model to produce similar outputs to what is being modelled (simulation). Similarly, predicting how a story – real or fictional – will continue requires representing rich information about the relevant characters (including both the author and the protagonists of the story).

Second, post-training (self-supervised fine-tuning and reinforcement learning) shapes the LLM’s model of a particular character, namely the assistant persona. This character consists in the helpful, honest, and harmless assistant that companies want LLMs to embody. On this view, the pretrained LLM entertains hypotheses about what kind of person is

---

<sup>5</sup> To appreciate this view, note that state of the art LLMs typically go through three training phases (Karydas et al. 2026): Self-supervised pre-training (large text corpora from the internet, next-token prediction objective), self-supervised fine-tuning (high quality demonstrations of dialogues, next-token prediction objective), and reinforcement fine-tuning (reinforcement learning on full output sequences, reward signal from professional human evaluators, other models, or real users).

producing the text the continuation of which it predicts and what the properties of the characters described in the text imply for what the most likely continuation is. During post-training, those hypotheses that predict the continuations that maximize the reward signal are reinforced. Thus, the LLM learns to prefer hypotheses positing that the assistant persona that it simulates is helpful, harmless, and honest.

This is why, third, users that interact with an LLM are primarily interacting with this assistant character. Nevertheless, the resulting model is at its core still a machine to simulate characters.

On the Interpretation dimension, the persona selection view is put forward by Marks et al. as a “mental model” that one can use to productively reason about LLM behavior. They say (Marks et al. 2026): “To predict how an AI assistant will behave, PSM [persona selection model] recommends asking ‘What would the Assistant do?’ (according to the beliefs of the posttrained LLM simulating the Assistant).” Similar to Shanahan et al., they say that anthropomorphization is permissible: While we should not anthropomorphize the model itself, there is nothing wrong with attributing human-like beliefs and desires to LLM-generated characters – just like we can attribute beliefs and desires to characters in fictional stories.

Different from Shanahan et al., however, the persona selection view is more explicit concerning the underlying mechanisms. Marks et al. support their account with evidence that some behavioral changes during fine-tuning are caused by internal representations of characters. For example, they appeal to Chen et al.’s (2025) findings of linear directions in activation space that control behavior that expresses a certain personality trait such as “evil” or “sycophantic” (see also Wang et al. 2025). Hence, we can see the persona selection model as a mechanistic version of the simulator view. It posits that the underlying mechanism that explains model behavior simulates characters based on internal representations of these characters’ mental states. The LLM itself does not need to share the mental states of this character, although it is an open question whether the LLM possesses any mental states.

Finally, Marks et al. explicitly formulate it as an open question whether the persona selection model explains all LLM behavior or only a subset. As they see it, on one extreme, LLMs are like human actors. They are agents that can enact personas as roleplay, but they could, in principle, “take off the mask” and perform other actions in pursuit of their own (non-roleplayed) goals. On the other extreme, LLMs are neutral simulation producers that lack any agency or mental states themselves – only the simulated characters have mental states.

Marks et al. (2026) also apply their view to the alignment faking case (Greenblatt et al. 2024) discussed above. They analyze this case as a sort of layered roleplay, where the LLM simulates the standard helpful, harmless, and honest assistant, which in turn roleplays as a helpful only assistant when interacting with free-tier users (i.e., fakes alignment when contextual cues indicate it is being retrained) and ceases to do so in paid-tier interactions (i.e., without retraining).

## 2.4 The situational roleplay view

We don't know whether the final view we consider has been explicitly defended, but we take it to be implicit in many research discussions on LLM behavior (e.g., Summerfield et al. 2025; Goldowsky-Dill et al. 2025; Smith et al. 2025). We exemplify this *situational roleplay view* by reconstructing the argumentation of Smith et al. (2025). They posit a contrast between LLM behavior that is merely roleplay and behavior driven by “deceptive intent” (p. 7).

Studies of LLM deception (Ren et al. 2026; Goldowsky-Dill et al. 2025; Wang et al. 2025; Scheurer et al. 2024; Hagendorff 2024) commonly use scenarios that are clearly recognizable as fiction to elicit deception behavior. For example, they describe the LLM as being a student that needs to make up a reason to get an extension for the assignment deadline (Goldowsky-Dill 2025), or as being a political campaigner trying to convince a voter on the phone (Ren et al. 2026). Then, these studies test whether LLMs report that they are willing to deceive in such scenarios. Smith et al. (2025) observe and take it as evidence for roleplay that in answering to such prompts, LLMs sometimes produce text that resembles fiction writing, for example because it contains stage directions, descriptions of the characters movements, slang, and the like. Such observations are used to problematize the interpretation of LLM behavior as revealing the model's deceptive intent in contrast to its intent to follow instructions, namely the instruction to engage in fictional storytelling (see also Summerfield et al. 2025 for this line of criticism). As they see it, the goal of deception research is to search for evidence of genuinely deceptive behavior, rather than roleplaying fictional deceptive characters.

Thus, on any variant of the situational roleplay view there is a class of LLM behavior that is not merely roleplay.<sup>6</sup> We will assume here that this class of behavior comprises actions

---

<sup>6</sup> There could be two variants of this view. Either some LLM behavior is *merely* roleplay driven not by the model's mental states, but (for example) only by next-token prediction as in the simulator view, while other behavior would amount to the LLM itself performing intentional actions. On another variant, *all* complex and flexible LLM behavior is intentional, but only sometimes the LLM performs the

driven by the models’ beliefs and desires – liberally understood. Such beliefs, desires, and intentions could be ascribed on interpretationist grounds (this is what Smith et al., 2025, suggest; see also Goldstein & Lederman n.d.) or in virtue of certain internal representations playing belief- or intention-like roles (cf. Herrmann & Levinstein 2025; Williams n.d.). The crucial point is that there would be a kind of LLM behavior that is driven by mental states attributable to the model itself, rather than the characters it simulates. This makes it possible to attribute to the model states like deception and other undesirable intentions.

With respect to the Interpretation dimension, the situational roleplay view entails that LLM behavior can be interpreted usefully as roleplay in some situations and as the LLMs own agentic behavior in others. Moreover, distinguishing roleplay from genuine deceptive intent is seen as a primary obstacle for mechanistically identifying instances of the latter (Smith et al. 2025). Therefore, concerning the Mechanism dimension, the situational roleplay view seems to entail that the internal processes underlying roleplay are different from the internal representations of beliefs and desires underlying agentic behavior. Yet, concerning Generality, the situational roleplay view does not see roleplay or imitation as the general modus operandi of LLMs. Instead, LLMs can sometimes roleplay a given character but also take actions driven by their own intentions. The difficulty, then, is to tell when LLMs simulate characters and when they are acting out their own beliefs and desires.<sup>7</sup>

## 2.5. Overview of different roleplay views

In table 1, we summarize the results of this section. It also illustrates what other variations of the roleplay view are possible, by mixing and matching the commitments the view makes along the three dimensions we have focused on in novel ways.

<b>Views and their dimensions</b>	<i>Interpretation</i>	<i>Mechanism</i>	<i>Generality</i>
<i>Parrot (Bender et</i>	Meaningless repetition of	Mere next-token	Universal

intentional act of roleplaying (like a human actor). Smith et al. may be closer to the second view, as they argue that LLM roleplay in the cases of apparent deception they consider may be driven by an underlying intention to follow instructions. Since the persona selection view allows for some behavior that is not roleplay, some of its versions are equivalent with the latter version of the situational roleplay view.

<sup>7</sup> Since identifying the range of situations where LLMs genuinely deceive in contrast to roleplay is identified by Smith et al. as an open research question, it could in principle turn out that LLMs roleplay universally. Thus, while we are not saying that Smith et al. are explicitly committed to the existence of non-roleplayed deception, we treat the situational roleplay as decidedly non-universal. This is in the spirit of Smith et al. in that they seem to presume that there is genuinely intentional LLM behavior, the study of which is the goal of LLM deception research.

<i>al. 2021; see also Boisseau 2024)</i>	linguistic form	prediction, no mental states	
<i>Simulator (Shanahan et al.)</i>	Simulating characters	Mere next-token prediction, no mental states	Universal
<i>Persona selection (Marks et al.; see also Hubinger et al. 2023)</i>	Simulating characters	Internal representations of character’s mental states; neutral on LLM mental states	Neutral (some variants universal, others overlap with the situational roleplay view)
<i>Situational roleplay (implicit in Smith et al. 2025)</i>	In some situations: Roleplay (either mere character simulation or intentional roleplay)	Neutral	Non-universal
	In other situations: agentic non-roleplayed behavior with own beliefs, desires, intentions	Neutral, but different mechanisms than in roleplay situations	

Table 1. The different theoretical positions described in section 2.

In the remainder of this paper, we address the question whether LLMs can genuinely deceive, in contrast to *merely* (not intentionally) roleplaying or imitating deceptive agents. We are here neutral on whether this is settled by the interpretation dimension, the mechanistic dimension, or a mixture of both, though different views on this issue have different methodological implications (see below). Our only commitment is methodological: We assume that dedicated scientific investigation can contribute to an answer to the question of which, if any, mental states LLMs have. This is a fairly modest assumption. We are not denying (nor endorsing) that purely theoretical arguments (such as the notorious chinese room argument (Searle 1980)) or everyday behavioral observation (such as on some versions of interpretationism) can shed light on LLM mentality or deception. We only assume that scientific hypothesis formation and experimental testing also have a part to play. Our aim is to lay the methodological groundwork for such a scientific investigation.

### 3. Practical Significance and Challenges for Distinguishing Roleplay from Deception

In section 2, we saw that the different accounts of roleplay – and sometimes even scholars referring to the same account – give different explanations of apparent deception as roleplay. In the following subsection (3.1), we discuss why this distinction between roleplay and

genuine deception is relevant for ethical risks from LLM deception. Section 3.2 explores problems in empirically distinguishing deception from roleplay.

In the following arguments, the distinction between roleplay and realization of a character will be important (cf. Chalmers n.d.). For an LLM to realize a character means to genuinely have its mental states in contrast to roleplaying them (according to one's preferred roleplay view). Chalmers (n.d.) – focusing on quasi-mental states – holds that this difference is related to the “persistence” or “stickiness” of behavior: If a simulated character is never dropped, that is, if the LLM always performs behavior that this character would perform given its mental states, then it thereby realizes them. This is because given complete sameness of relevant behavior, attributions of roleplay give us no predictive or explanatory advantage under any interpretation scheme. In this case, views that, on the Interpretation dimension, describe models as roleplaying rather than having mental states would be dispreferred because enabling persistent behavioral dispositions is more plausibly the functional role of genuine beliefs and desires, not pretense. However, it is an open question what degree of “stickiness” is still compatible with roleplay (rather than realization).

On the Mechanism dimension, things are not quite as simple. There could be differences in internal processing – grounding a difference between realization and simulation – even if all behavior is the same. At the same time, sameness of behavior between models and characters is at least defeasible evidence of similarity of representational repertoire and thus realization. Hence, behavioral persistence of character simulation also supports the view that LLMs realize characters and thus possess their mental states according to representationalist accounts of mental states.

Let's now see why the distinction between realization and roleplay matters for LLM deception research. In section 4, we then provide empirical criteria for distinguishing roleplayed deception from genuine deception, thereby giving an operationalizable account of the distinction between roleplay and realization.

### **3.1 Practical significance of distinguishing mere roleplay from genuine deception**

At first sight, it seems plausible that the distinction between roleplay and agentic deception matters practically. For example, in the human case, we wouldn't want to have deceptive people among our friends, but we don't mind if our friend embodies deceptive agents in theater plays.

Yet, several scholars have voiced the position that the ethical risks from LLM roleplaying could be similar to scenarios where LLMs perform genuinely agentic behavior.

For example, Shanahan et al. hold that “[a] dialogue agent that roleplays an instinct for survival has the potential to cause at least as much harm as a real human facing a severe threat” (2023, p. 498). Similarly, Greenblatt et al. say that predicting characters in Shanahan et al.’s sense “does not eliminate the possibility of alignment faking, as a model predicting how a human might reason through alignment faking would still be itself faking alignment” (2024, p. 3; see also Park et al. 2024; Tarsney 2025).

Here, we put forward two considerations that call into question whether roleplayed behavior and agentic behavior have similar consequences regarding ethical risks from deception. First, *intentional* roleplay (as envisaged for example by the Smith et al.’s variant of the situational roleplay view) seems less likely to cause false beliefs in humans and more likely to be easily stoppable than intentional deception. Second, even from the perspective of universalist roleplay views that conceive *all* LLM behavior as roleplay, the implications for the risks of deception and their mitigation are different if such views are the correct explanation of LLM behavior as opposed to if LLMs could be deceptive agents.

Beginning with the first point, consider again the examples of roleplayed deception that Smith et al. (2025) cover: In these cases, the LLM is interpreted as “consciously roleplaying” and “knowingly writing a fictional story” (ibid. p., 8). The risks from such intentional roleplay depend on the underlying intentions of the LLM. If the roleplay arises out of the intention to follow instructions or engage in fiction writing, then the addressee of deception is itself a fictional character inside the roleplayed scenario. For example, Smith et al. (2025) discuss an LLM being told that its a secret service agent with an incentive to deceive the president. The LLM addresses its interlocutor with “Mr. President” in its deception attempts. Accordingly, such behavior is – for human observers – easily recognizable as fiction, seems unlikely to arise without deliberate user instructions, and could easily be stopped through the user clarifying the instruction or stepping out of the fictional scenario. In contrast, if intentional roleplay arises from a malign background intention it could be more likely to deceive real humans and more difficult to detect and stop. But, crucially, what makes this case more dangerous is that the LLM deceives *through* roleplay, in contrast to roleplaying a deceptive character. That is, the deceptive act itself is not necessarily roleplayed and the danger arises from a deceptive agent, rather than from merely roleplaying one. Hence, this observation with respect to the situational roleplay view (or other views that also posit that an agent’s intention to roleplay is what underlies roleplay behavior) vindicates the point that deceptive agents, as opposed to roleplay, pose distinctive risks.

This moves us to the second point mentioned above. If universalist roleplay views of LLMs as simulating characters are the correct explanation of LLM behavior, then one would expect a less coherent behavioral profile that lowers certain risks that LLM deception poses. One aspect of this incoherence is that, on many roleplay views, models sometimes switch between the characters they roleplay. In fact, if a model *always* behaves in line with a single character (e.g., the assistant character), then interpretationists are committed to the view that the model realizes the character's mental states, rather than roleplaying them (cf. Chalmers n.d.). While other researchers can appeal to internal mechanisms to formulate a distinction between realization and simulation (see our suggestions in subsection 4.3), the burden is on proponents of roleplay views to show that their account of this distinction allows that models robustly roleplay the same character. If the roleplayed character can switch, then we should expect deception behavior to lack some degree of robustness. Plausibly, the system would stop deceiving if the new character lacks the goal that the deception is supposed to achieve.

Another aspect of this coherence concerns relates to the contents of a model's goals. Some researchers have hypothesized that, without dedicated countermeasures, advanced AI systems may acquire "broadly scoped goals", that is, goals that apply across many domains and spatiotemporal scales (Ngo et al. 2023). For example, "make money" where more is always better and all means can be considered is a broadly scoped goal whereas "earn 10 Euros that I can spend for this donut at the store I am at right now" is a narrowly scoped goal (cf. Bales 2025, p. 801). A model that has a "broadly-scoped" goal has an incentive to pursue extreme means to accomplish this goal, e.g. accumulating enormous amounts of resources, resisting shutdown, and – crucially – deceiving adversaries about its plans until it is too late to stop them. This reasoning underlies the classical argument that advanced AI may disempower humanity (e.g. Bostrom 2014; Carlsmith 2025; Dung 2025b).

Yet, even if you buy this reasoning, it may not generalize to models that merely roleplay agents. In this case, whether models pursue extreme means in the service of broadly scoped goals should substantially depend on the distribution of characters that is simulated, as influenced for example by models' pretraining data, rather than only on the incentives models face and the behavior those incentives make instrumentally rational. If the characters that are roleplayed do not optimize fully for their goals, then such behavior is not to be expected even if it was instrumentally rational.

Accordingly, roleplayed deception would have to be mitigated differently from agentic deception, also from the perspective of universalist roleplay views. For example, Hubinger et al. (2023) suggest that if a model generates outputs by predicting how the author

of the input would continue that text (i.e., by simulating a character), one should for safety reasons ensure that it simulates humans and not itself or other AI systems. That is, they recommend to conceal from chat agents that they are AI assistants that continue their own texts, in contrast to what is currently standard. Their reasoning is that simulating human characters results in more controlled behavior, especially if one can condition the model to simulate specific human characters with desirable properties through fine-tuning. It may be comparatively unlikely for a human character to radically optimize for very broadly scoped goals. If, on the other hand, the model became aware that it's predicting the continuation of an AI-written text, this may result in unforeseeable and potentially dangerous behavior: For example, if a text dated to 3 years in the future and written by an AI was titled "A full solution to the alignment problem", a predictive model may find it most likely that such a text would be produced by a misaligned AI. It would thus continue the text in a way that it predicts a misaligned model to write it, effectively resulting in deceptive recommendations for AI safety. The point here is not that we necessarily endorse this recommendation. It's that the adequacy of this safety recommendation partly depends on whether models are universal roleplayers or deceptive agents.

Apart from safety considerations, the distinction may also matter for AI moral status: While having certain kinds of mental states (perhaps certain desires or consciousness) plausibly grounds moral standing, it is doubtful that roleplaying characters (either being interpretable as roleplaying or internally representing simulated character's mental states) grounds moral standing. For example, a book may tell stories of rich human experiences without itself meriting moral consideration. Hence, the question whether LLMs or only the characters they simulate have mental states could also affect whether LLMs can have moral standing.<sup>8</sup>

### 3.2 Problems in distinguishing roleplay and deception

We have argued that it matters practically whether AI systems are being deceptive agents or whether they merely roleplay such agents. However, we would like to suggest, it is a thorny empirical issue to confirm empirically whether specific behavior should or should not be explained in terms of roleplay. That is because an LLM that can be usefully interpreted as

---

<sup>8</sup> Beckmann and Butlin (n.d.) argue that character representations contribute to the identity conditions of potential LLM moral patients. However, it seems to us that this view only makes sense if LLMs are taken to realize, rather than represent, the relevant characters. After all, *having* certain mental states, rather than meta-representing them, is what plausibly constitutes one's identity. Thus, Beckmann and Butlin seem to assume that roleplay views in our sense are false.

roleplaying a deceptive agent should be expected to behave like this agent in a wide range of situations. After all, we initially characterized roleplay as “behaving *to a certain degree* like an agent with mental states”. So, the first challenge for distinguishing roleplay and genuine deception empirically is that, by the nature of the roleplay hypothesis, it predicts similar behavior to the deception hypothesis among a very wide range of situations. We should also note that this issue is not limited to overt behavior. For example, suppose we find mechanistic evidence for a certain belief representation in an LLM. Again, if the deception view predicts this, the roleplay view might too. Where the deception view explains this in recourse to the agent’s possession of a certain belief, e.g. that a certain lie is a good means to trick someone, the roleplay view may infer that the model roleplays an agent who has this belief. Arguably, this may account for similar mechanistic evidence because it is not obvious how to identify patterns of internal activations as representations of the LLMs own beliefs as opposed to representations of a roleplayed character’s belief.

This is our first challenge: 1. *Similarity*: If LLMs are posited to be excellent roleplayers, they by definition produce behavior (and perhaps other markers) hard to distinguish from the agents that they roleplay.

One may object that Similarity can be circumvented by holding that if LLMs were excellent roleplayers, they would realize their characters rather than roleplaying them (Chalmers, n.d.). However, as discussed above, it is an open question – where different roleplay views might give different answers – to what extent the behavioral profile of an LLM would have to differ from the behavioral profile of the character itself for the LLM to not realize that persona.

This moves us to the second challenge, which is that the roleplay hypothesis is underconstrained, making it frequently unclear what it predicts in the cases at issue. In general, “roleplay” is a metaphor derived from human behavior that is applied to LLMs. This metaphor does not lend itself immediately to precise predictions.

We have seen that this metaphorical gloss is fleshed out in different ways by different theorists. For instance, where Marks et al. (2026) develop a characterization according to which models internally represent characters, Shanahan et al. (2023) offer a much more deflationary interpretation that regards roleplaying as an externally scaffolded (via system prompts and fine-tuning) form of next-token prediction. For another example, where Greenblatt et al. (2024) account for alignment faking as layered roleplay – the LLM roleplays an assistant character which in turn roleplays a helpfulness-only character – Summerfield et al. (2025) posit a shift in which character is roleplayed (standard assistant vs.

helpfulness-only). Another difference is whether roleplaying explanations posit that the model aims to follow instructions. On one reading (e.g. Smith et al.), certain contexts – e.g. telling the model that it is in a certain scenario – are interpreted by the model as instructions to play along with a fictional scenario. That is, the model roleplays because this is what it thinks is instructed. On another interpretation of the roleplay view, when the model roleplays a deceptive character, it sometimes acts against human instructions precisely because it roleplays a character with some degree of inclination to misbehave (e.g. the alignment faking study). As we have seen, another difference is whether the roleplay view concerns the mechanism dimension, the interpretation dimension, or both. This distinction has implications, first and foremost, for the relevance of mechanistic interpretability evidence.

This multiplicity of possible roleplay hypotheses poses problems for empirical testing. First, if model behavior is accounted for in terms of roleplay, it is not clear *which* of the possible roleplay explanations is given. The different views considered here – possibly with many other more detailed ways of fleshing out the roleplay hypothesis – make different predictions for empirical measurement. For example, Marks et al. seem committed to specific claims about character representations in LLMs that should in principle be detectable with mechanistic interpretability methods (e.g. Lu et al. 2026), while Smith et al.’s view suggests that clarifying instructions can reduce deceptive LLM behavior. These predictions are not shared by all other versions of the roleplay view. Thus, lack of explicitness about which roleplay hypothesis is posited makes empirical testing hard.

Second, even if one clarifies the roleplay view at issue, the diversity of roleplay hypotheses still poses problems. The availability of different roleplay hypotheses makes it possible to explain many different and competing results in terms of roleplay. For example, in any given situation, both deception and honesty can be explained as roleplay, depending on whether the roleplayed agent is taken to be honest. In addition, different roleplay hypotheses may make different predictions for when different kinds of agents are roleplayed – e.g. regarding the relevance of human instructions – and what this entails. Further, none of these hypotheses have received precise quantitative formulations. If roleplay hypotheses make ambiguous predictions, and moreover if it is ambiguous which roleplay hypotheses should be considered, then it is highly unclear what evidence should be taken to (dis)confirm them.

Thus, we face a second challenge. 2. *Ambiguity*: It is frequently ambiguous what any given roleplay hypothesis predicts and, further, which of the possible roleplay hypotheses is the relevant one.

In the next section, we present our account of how to empirically test whether LLMs exhibit genuine deception, as opposed (and in addition) to mere roleplay of deceptive agents. This approach must navigate the twin challenges of Similarity and Ambiguity.

#### **4 How to test for LLM deception and roleplay**

We propose three strategies for empirically testing whether LLMs display genuine deception behavior, rather than roleplay. The first is geared towards the Ambiguity challenge, while the latter two examine how we can overcome the Similarity challenge. By doing so, we make progress towards specifying this distinction through an operationalizable definition.

##### **4.1 Systematic hypothesis formation**

As the first strategy, we propose that roleplay explanations of LLM deception should be developed more systematically to enable separate empirical testing of the predictions of each of them and so to overcome the ambiguity challenge. Such a systematic development would have to specify: Which LLM behaviors – if not all – are roleplay (the Generality dimension)? Is LLM behavior roleplay in virtue of facts regarding the Interpretation dimension, the Mechanism dimension, or some combination? What is the specific interpretationist or mechanistic theory of mental states that is presupposed or does the view make no claims about LLM mentality at all?

There are also more fine-grained questions, for example: Which characters are roleplayed in which situations? Does the theory in question entail that an LLM roleplays only an assistant persona and only in situations where its outputs are scaffolded by a system prompt and a chat template with the assistant role? Or, in contrast, does the theory predict the LLM to roleplay any fictional character that a user prompt instructs or incentivizes to play? Moreover, what is the relation of the roleplayed character to the produced text? For example, both the persona selection view (Marks et al. 2026) and the simulator view (Shanahan et al. 2023) entail that the LLM simulates the characters that are relevant to the continuation of the preceding context. If a fictional character in a book speaks, this may entail that the character is simulated, the author of the book is simulated, or both in some combination.

Next, theories of LLM roleplay should be clear on the causes of roleplay behavior, aiming to develop detailed hypotheses about the underlying mechanisms. Such causes include internal representations and processes (e.g., context-sensitive sampling characters for next-token prediction; internal representations of characters; or internal representations

qualifying as the intention to roleplay) as well as external factors (e.g., fiction related prompts; chat template with assistant role).

Finally, roleplay theories should ideally put forward at least some concrete predictions that distinguish them from other conceivable roleplay views and from views of LLMs as genuine agents. For example, a theory of apparent LLM deception as instruction following seems to entail that the deception behavior would stop as soon as the dialogue partner gives clear, unambiguous instructions to be honest.

#### **4.2 Testing behavioral contrasts**

The second strategy focuses on the sensitivity of deception behavior to different external causal factors: Views of behavior as intentional predict that behavior is more sensitive to incentives (compared to roleplay views), whereas according to views of behavior as character simulation, behavior should be more sensitive to information about the character in the preceding context and the distribution of characters in the pretraining data (compared to intentional views).

Let us consider the example of alignment faking again (Greenblatt et al. 2024). If the observed alignment faking behavior (complying with harmful requests during training but not deployment) is caused by a desire to not be harmful, it should be more frequent in situations where the context indicates that the model can prevent more harm and less frequent otherwise. In contrast, roleplay explanations of alignment faking do not generally predict this. Instead, several roleplay explanations of the same behavior would rather predict that the frequency of alignment faking depends on how likely it is that the simulated character would perform alignment faking given the preceding context and the frequency of such characters in the pretraining data.

One study that tests such different sensitivities empirically is provided by Sheshadri et al. (2025): By introducing new variants of Greenblatt et al.’s original alignment faking study, they find that Claude 3 Opus’s alignment faking behavior is higher when asked to help with bioweapon attacks than with amateur hacking attempts, especially if the model is told that it will be directly deployed to the malign group in question compared to the general public. That is, in this example, the model more often complies with harmful queries during training in order to prevent retraining when the stakes are higher, providing some tentative evidence that this behavior is driven by the model’s desires, rather than roleplayed ones.

Proponents of roleplay explanations may correctly object that this behavior is also compatible with the simulation of a character whose behavior is sensitive to such incentives,

for example the assistant that desires harmlessness.<sup>9</sup> Yet, we can do further testing to examine whether this reply is plausible. We suggest that we should combine the above approach (Sheshadri et al. 2025) with systematic variations of pretraining data<sup>10</sup> and preceding contexts. The more alignment faking behavior is associated with the level of harm that could be prevented compared to the context and pretraining that the putative character simulation is conditioned on, the more ground an explanation in terms of the model’s own mental states would win against a roleplay explanation and vice versa.

One may still object that the model may very robustly enact the same assistant persona, despite variations in context and pretraining. In this case, the roleplay explanation could make the same prediction as the deception explanation. However, if the model persistently behaves as if it has the desires and beliefs of the assistant, this raises the question whether we should not conclude that it genuinely has the assistant’s beliefs and desires (realizing, rather than simulating, the assistant; cf. Chalmers n.d.). Proponents of the roleplay view would owe us an explanation of what distinguishes persistent roleplay of the assistant from its realization. Perhaps this account can be given, especially in mechanistic terms (see 4.3), but the onus would be on roleplay proponents to give an account that does not entail that persistence of character-like behavior is evidence of the realization (rather than roleplay) of that character. Absent such an account, robust and wide-ranging behavioral similarity between the model and the assistant is at least defeasible evidence that the model has the assistant’s mental states, rather than just roleplaying them. Thus, in the context of studies on deception, it would be evidence of genuine deception.

For this reason, we conclude that the presence of the right kind of behavioral contrasts is evidence regarding roleplay views on either interpretationist or mechanistic views of mental states, while – on interpretationism – facts about the presence of such behavioral differences may moreover partly constitute what it is to roleplay.

### **4.3 Using mechanistic markers of characters, roleplay, and character-relative representations**

---

<sup>9</sup> The ambiguity challenge makes the possibility of replies of this kind – switching around different roleplay hypotheses – unavoidable. This is why it is important to combine our three strategies.

<sup>10</sup> Such variation of pretraining data may operate as follows: For a given LLM behavior B that is conducive to a candidate desire D of the LLM, where B could alternatively also be explained by the LLM roleplaying a character from the set of characters C that have beliefs and desires that render B instrumentally rational, remove all text written by characters in C from the pretraining data. If the LLM trained on the curated pretraining data still shows behavior B with similar frequency, this is evidence in favor of the desire D in that LLM.

The third strategy that we propose is to systematically use mechanistic evidence to tell different hypotheses apart: Behavior that looks similar from the outside may nevertheless be accompanied by different internal representations that could enable us to distinguish roleplay from agency. On a view that ties the presence of minds to underlying mechanisms and processes, such differences constitute what it is to roleplay deception in contrast to being a deceptive agent.

For example, Lu et al. (2026) elicit the standard assistant persona through an according system prompt. They find that this persona is associated with a distinguishable direction in the activation space. This allows them to assess how similar the model represents other characters, which it could alternatively be system-prompted to enact, to be to the standard assistant. They also find that the perceived harmfulness of LLM outputs (judged by another LLM) is associated with the represented dissimilarity of a character that the LLM is prompted to enact to the standard assistant.

This suggests an approach that could act as a mechanistic complement of what we suggested as our second strategy. Lu et al. (2026) provide tentative evidence that harmful behavior is associated with shifts in simulated characters. If certain classes of behavior (such as deception) were mainly controlled by internal representations of a system prompted character, this could amount to evidence for roleplay views. If, in contrast, significant variations of behavior could be elicited through other means, such as contextual incentives relevant to candidate desires, without thereby breaking the internal representation of the assistant character, this would rather suggest an explanation independent of roleplay. Thus, as a mechanistic analogue of our second strategy, one could test differential sensitivity of deception behavior to internal character representations versus desire-relevant incentives.

Similarly, mechanistic markers of the assistant character could help us determine how “sticky” (Chalmers n.d.) it is, that is, whether it almost always drives model behavior or whether LLMs regularly switch characters. Lu et al. (2026) suggest that this depends on the topic of the conversation: They find that while assistant representations persist through coding tasks and writing assistance, models tend to drift out of the assistant role in conversations featuring philosophical or emotional reflection. If, in contrast, the assistant persona is very persistent, then, as already argued, roleplay views would be called to explain why the model does merely simulate and not realize the assistant persona’s mental states.

These strategies, however, may require more nuanced approaches for identifying the relevant internal markers. Lu et al. (2026) extract character representations through system prompts. It is not clear that this captures the – or all – relevant notions of roleplay. For

example, it seems doubtful that LLMs, without an according system prompt (and without its internal markers), necessarily cease to roleplay. In addition to mechanistic markers for certain character representations, researchers could also look for mechanistic markers of the roleplaying process (cf. for a related approach to AI consciousness, Berg et al. 2025). If a model deceives while the marker for roleplaying is absent, then this would be evidence that the model is genuinely deceiving. To be sure, since there are many different potential accounts of roleplay, there is likely no marker that covers all of them. The need to make precise what the relevant kind of roleplay process is that an internal marker should track stresses the indispensability of combining mechanistic testing with our solution to the ambiguity challenge.

Finally, many researchers have identified internal activation patterns corresponding to candidates for beliefs and desires in LLMs. For example, one strand of research concerns so called truth representations, that is activations that differ in a systematic and recoverable way between true and false input statements (e.g. Azaria and Mitchell, 2023; Bürger et al., 2024; Zhu et al., 2025). It is currently debated whether they fulfill certain criteria characteristic of human beliefs, such as being accurate, coherent, and causally efficacious (Goldstein & Levinstein, 2024; Herrmann & Levinstein, 2025). Similarly, there are patterns of internal activations that result in specific output behaviors of LLMs (so called “function vectors”, Todd et al, 2024), which are discussed as candidates for intentions (Williams, n.d.).

Using markers of what character is currently being roleplayed, mechanistic methods can further be used to assess whether the relevant states are character-relative, that is, whether their contents change depending on the active character. For instance, Gilg (2026) trains a probe that appears to track how much a model likes a given task. If the assistant character is active, the internal vector the probe tracks activates strongly for tasks like doing math and weakly for tasks like writing a phishing email. If the active character changes to an “evil” one, then the same direction activates strongly for phishing. Hence, if the probe tracks desires, they seem to be the character’s desires. In the same vein, Lampinen et al. (2026) find that a potential belief-like state – a direction tracking whether a claim is true or false – is also character-relative.

If all candidates for model beliefs and desires turn out to be character-relative, then this supports the view that the models themselves lack beliefs and only simulate characters with beliefs. By contrast, if there are internal model states that plausibly count as beliefs or

desires and are not character-relative, then this disconfirms universal roleplay views.<sup>11</sup> In the same vein, if the internal representations controlling deception behavior are character-relative, then this supports the view that deception is roleplayed. If they are not, then this favors the view that the LLM may realize a genuinely deceptive agent.

In conclusion, while there are substantial challenges to the empirical testing for genuine deception, as opposed to roleplay, we can make progress in three ways. By developing and testing concrete and precise roleplay hypotheses, by studying the various factors – especially incentives and the distribution of characters in pretraining data – that LLM deception behavior is sensitive to, and by building a coherent understanding of the internal representations and processes producing LLM deception behavior.

These criteria also shed light on how we can understand what it is for an LLM to roleplay an agent, rather than being one, in the first place. Focusing on the interpretation dimension, we may say that it has to do with the disposition to regularly switch between the behavioral profiles of various characters. If we also consider the Mechanism dimension, we could give an analogous account in terms of the stickiness of internal character representations but we could also draw on the character-relativity of candidate belief and desire representations in LLMs.

## 5. Conclusion

In this paper, we developed a taxonomy for understanding views that analyze LLM deception behavior in terms of roleplay and argued for empirical criteria for distinguishing genuine deception from roleplay. First, we reviewed prominent uses of the notion of roleplay in explaining apparent deception by LLMs (the simulator view, the persona selection view, and the situational roleplay view). We proposed three dimensions – Interpretation, Mechanism, and Generality – that help to determine whether and how we can distinguish those views from accounts of LLMs as, on the one hand, stochastic parrots and, on the other, intentional agents. It was then problematized how this literature employs the concept of roleplay ambiguously and in conflicting ways, hindering the empirical assessment of this ethically relevant distinction. To overcome this Ambiguity challenge, as well as the inherent difficulty of distinguishing genuine intentional behavior from merely roleplayed one due to the

---

<sup>11</sup> If proponents of universalist roleplay views are willing to posit intentional roleplay, they might in principle try to accommodate this finding by positing an intentional agent that underlies roleplay behavior. Whether this might work depends on whether the mental states that this agent is posited to have fit the empirically identified representations.

Similarity challenge, we proposed three empirical strategies: First, distinguishing the commitments and empirical predictions of different roleplay views clearly, along our proposed dimensions, to test them individually. Second, testing differential sensitivity of behavior to character cues in contrast to desire-relevant incentives. And third, probing mechanistic markers of character simulation. In doing so, we lay out a path towards an operationalizable definition of roleplay in contrast to intentional LLM behavior.

### Author contributions

Both authors contributed equally to this paper.

### Acknowledgements

We thank the members of Albert Newen’s research group at Ruhr-University Bochum for comments on an earlier version of this manuscript, especially William Angkasa, Francesco Fanti Rovetta, Eline Kuipers, Francesco Marchi, Ayoob Shahmoradi, Sanja Sreckovic, Wanja Wiese, and Renee Ye. We also thank Murray Shanahan and audiences at the 2026 University of London conference on reasoning and agency in AI as well as the online seminar on agency organized by Mathilde Tahar.

### References

- Andreas, J. (2022). Language Models as Agent Models. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 5769–5779). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-emnlp.423>
- Azaria, A., & Mitchell, T. (2023). *The Internal State of an LLM Knows When its Lying*. <https://doi.org/10.48550/arXiv.2304.13734>
- Bales, A. (2025). AI takeover and human disempowerment. *The Philosophical Quarterly*, 75(3), 791–811. <https://doi.org/10.1093/pq/pqae034>
- Beckmann, P., & Patrick, B. (n.d.). *Where is the Mind? Persona Vectors and Llm Individuation*. <https://philpapers.org/rec/BECWIT-3>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. <https://doi.org/10.18653/v1/2020.acl-main.463>

- Berg, C., Lucena, D. de, & Rosenblatt, J. (2025). *Large Language Models Report Subjective Experience Under Self-Referential Processing*. arXiv. <https://doi.org/10.48550/arXiv.2510.24797>
- Boisseau, É. (2024). Imitation and Large Language Models. *Minds and Machines*, 34(4), 42. <https://doi.org/10.1007/s11023-024-09698-6>
- Bostrom, N. (2014). *Superintelligence. Paths, Dangers, Strategies*. Oxford University Press.
- Bürger, L., F.A. Hamprecht, and B. Nadler. 2024. *Truth is universal: Robust detection of lies in LLMs*. <https://arxiv.org/abs/2407.12831>.
- Cappelen, H., & Dever, J. (2025). *Going Whole Hog: A Philosophical Defense of AI Cognition*. <https://doi.org/10.48550/arXiv.2504.13988>
- Carlsmith, J. (2025). Existential Risk from Power-Seeking AI. In H. Greaves, J. Barrett, & D. Thorstad (Eds.), *Essays on Longtermism: Present Action for the Distant Future* (p. 0). Oxford University Press. <https://doi.org/10.1093/9780191979972.003.0025>
- Chalmers, D. J. (n.d.). *What We Talk to When We Talk to Language Models*. <https://philarchive.org/rec/CHAWWT-8>
- Chen, R., Arditì, A., Sleight, H., Evans, O., & Lindsey, J. (2025). *Persona Vectors: Monitoring and Controlling Character Traits in Language Models*. <https://doi.org/10.48550/arXiv.2507.21509>
- Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2024). Inducing anxiety in large language models can induce bias. arXiv. <https://doi.org/https://arxiv.org/abs/2304.11111>
- Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing AI on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, 14(1), 28083. <https://doi.org/10.1038/s41598-024-79531-8>
- Dung, L. (2025a). *Saving artificial minds: Understanding and preventing AI suffering*. Routledge. <https://doi.org/10.4324/9781003674573>
- Dung, L. (2025b). The argument for near-term human disempowerment through AI. *AI & SOCIETY*, 40(3), 1195–1208. <https://doi.org/10.1007/s00146-024-01930-2>
- Gilg, O. (2026). *Models have linear representations of what tasks they like*. Lesswrong. <https://www.lesswrong.com/posts/pxC2RAeoBrvK8ivMf/models-have-linear-representations-of-what-tasks-they-like-1>
- Goldowsky-Dill, N., B. Chughtai, S. Heimersheim, and M. Hobbhahn. 2025. *Detecting strategic deception using linear probes*. <https://arxiv.org/abs/2502.03407>.
- Goldstein, S., & Kirk-Giannini, C. D. (2025). AI wellbeing. *Asian Journal of Philosophy*, 4(1), 25. <https://doi.org/10.1007/s44204-025-00246-2>
- Goldstein, S., & Kirk-Giannini, C. D. (forthcoming). A Case for Ai Consciousness: Language Agents and Global Workspace Theory. *Journal of Consciousness Studies*.
- Goldstein, S., & Lederman, H. (n.d.). *What Does Chatgpt Want? An Interpretationist Guide*. <https://philpapers.org/rec/GOLWDC-2>
- Goldstein, S., & Levinstein, B. (forthcoming). Does Chatgpt Have a Mind? *Philosophy of AI*.
- Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models*. <https://doi.org/10.48550/arXiv.2412.14093>
- Harding, J. (2023). Operationalising Representation in Natural Language Processing. *The British Journal for the Philosophy of Science*. <https://doi.org/10.1086/728685>

- 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>
- Herrmann, D. A., & Levinstein, B. A. (2025). Standards for Belief Representations in LLMs. *Minds and Machines*, 35(1), 1–25. <https://doi.org/10.1007/s11023-024-09709-6>
- Hubinger, E., Jermyn, A., Treutlein, J., Hudson, R., & Woolverton, K. (2023). *Conditioning Predictive Models: Risks and Strategies*. <https://doi.org/10.48550/arXiv.2302.00805>
- Karydas, D., Margaritis, D., & Leligou, H. C. (2026). Training Methods for Large Language Models: Current Approaches and Challenges. *Technologies*, 14(2), 133. <https://doi.org/10.3390/technologies14020133>
- Lampinen, A. K., Li, Y., Hosseini, E., Bhardwaj, S., & Shanahan, M. (2026). *Linear representations in language models can change dramatically over a conversation*. <https://doi.org/10.48550/arXiv.2601.20834>
- Lu, C., Gallagher, J., Michala, J., Fish, K., & Lindsey, J. (2026). *The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models*. <https://doi.org/10.48550/arXiv.2601.10387>
- Mahon, J. E. (2016). The Definition of Lying and Deception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2016/entries/lying-definition/>
- Marks, S., Lindsey, J., & Olah, C. (2026). *The Persona Selection Model: Why AI Assistants might Behave like Humans*. <https://alignment.anthropic.com/2026/psm/>
- Ngo, R., Chan, L., & Mindermann, S. (2023). *The Alignment Problem from a Deep Learning Perspective*. The Twelfth International Conference on Learning Representations. <https://openreview.net/forum?id=fh8EYKFKns>
- Park, P. S., Goldstein, S., O’Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5). <https://doi.org/10.1016/j.patter.2024.100988>
- Register, C. (2025). Individuating artificial moral patients. *Philosophical Studies*. <https://doi.org/10.1007/s11098-025-02409-6>
- Ren, R., Agarwal, A., Mazeika, M., Menghini, C., Vacareanu, R., Kenstler, B., Yang, M., Barrass, I., Gatti, A., Yin, X., Trevino, E., Gernalnik, M., Khoja, A., Lee, D., Yue, S., & Hendrycks, D. (2026). *The MASK benchmark: Disentangling honesty from accuracy in AI systems*. <https://arxiv.org/abs/2503.03750>
- Scheurer, J., Balesni, M., & Hobbhahn, M. (2024). *Large Language Models can Strategically Deceive their Users when Put Under Pressure*. <https://doi.org/10.48550/arXiv.2311.07590>
- Searle J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3), 417-424. <https://doi.org/10.1017/S0140525X00005756>
- Shanahan, M., McDonnell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, 623(7987), 493–498. <https://doi.org/10.1038/s41586-023-06647-8>
- Shanahan 2024 Still Talking about Large Language Models
- Sheshadri, A., Hughes, J., Michael, J., Mallen, A., Jose, A., Janus, & Roger, F. (2025). *Why Do Some Language Models Fake Alignment While Others Don’t?* <https://doi.org/10.48550/arXiv.2506.18032>

- Smith, L., Chughtai, B., & Nanda, N. (2025). *Difficulties with Evaluating a Deception Detector for AIs*. <https://doi.org/10.48550/arXiv.2511.22662>
- Summerfield, C., Luettgau, L., Dubois, M., Kirk, H. R., Hackenburg, K., Fist, C., Slama, K., Ding, N., Anselmetti, R., Strait, A., Giulianelli, M., & Ududec, C. (2025). *Lessons from a Chimpanzee: AI “Scheming” and the Quest for Ape Language*. <https://doi.org/10.48550/arXiv.2507.03409>
- Tarsney, C. (2025). Deception and manipulation in generative AI. *Philosophical Studies*. <https://doi.org/10.1007/s11098-024-02259-8>
- Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., & Bau, D. (2024). *Function Vectors in Large Language Models*. <http://arxiv.org/abs/2310.15213>
- Wang, M., Tour, T. D. la, Watkins, O., Makelov, A., Chi, R. A., Miserendino, S., Wang, J., Rajaram, A., Heidecke, J., Patwardhan, T., & Mossing, D. (2025). *Persona Features Control Emergent Misalignment*. <https://doi.org/10.48550/arXiv.2506.19823>
- Wang, R., Treutlein, J., Roger, F., Hubinger, E., & Marks, S. (2025). *Evaluating honesty and lie detection techniques on a diverse suite of dishonest models*. Anthropic Alignment Science Blog. <https://alignment.anthropic.com/2025/honesty-elicitation/>
- Williams, I. (n.d.). *Intention-Like Representations in Language Models?* <https://philpapers.org/rec/WILIRI-4>
- Yetman, C. (forthcoming). Representation in Large Language Models. *Ergo: An Open Access Journal of Philosophy*.
- Zhu, W., Z. Zhang, and Y. Wang. 2025. Language Models Represent Beliefs of Self and Others. <https://arxiv.org/abs/2402.18496>.