Deception-Enabling Cognitive Capabilities in AI Systems

Eyas Ayesh[1] Dobromir Rahnev[1] N. Apurva Ratan Murty[1] Anna A. Ivanova[1]

[1]Psychological and Brain Sciences, Georgia Tech
eayesh3@gatech.edu a.ivanova@gatech.edu

Abstract

To guarantee safe deployment of existing and future AI systems, we must be able to identify, understand, and mitigate deceptive AI behaviors. Here, we draw on decades of interdisciplinary deception research to discuss a set of cognitive capabilities that enable *explicit deception*—deliberate attempts to induce a false belief in another agent. These capabilities include intentionality, theory of mind, world modeling, and language fluency. Our framework provides a principled way to differentiate explicit deception from implicit deceptive behaviors, which may rely on a different set of capabilities and therefore require their own identification and mitigation strategies. Thus, a capability-focused approach provides a promising avenue for designing effective, case-specific monitoring and intervention strategies for AI systems.

1. Introduction

Rapid developments in artificial intelligence (AI) raise numerous questions about the potential of AI models to do harm. Of particular concern is the potential of these models to deceive humans [1]. An AI model that induces false beliefs in a human user is harmful both in the short term (doing a disservice to the user) and in the long term (evading human control and pursuing objectives that might harm humanity [2, 3]). Developing a framework for detecting and understanding AI deception is therefore essential for safe and reliable AI deployment [4].

AI models today already show some signs of deceptive behaviors: they can fake alignment under perceived monitoring [5], deny earlier policy violations when audited [6], and hide the reasoning behind risky decisions [7]. Although these behaviors are routinely grouped under deception, they may each depend on different underlying capacities or mechanisms.

Following classic work in the animal deception literature [8], we draw a distinction between deceptive behaviors at varying levels of sophistication (Figure 1). In this paper, we primarily focus on the highest level of deception, which we term *explicit deception*. These are behaviors driven by an explicit internal goal to induce a false belief in another agent. Explicit deception is distinct from implicit forms of deception — characterized by Mitchell's first three levels (Section 2) — where false beliefs are induced in another without the agent planning for that to happen, simply as a consequence of the agent's built-in or learned behavioral patterns.

Our central claim is that successfully carrying out explicit deception relies on specific deception-enabling capabilities: cognitive capacities that underlie or substantially enhance the

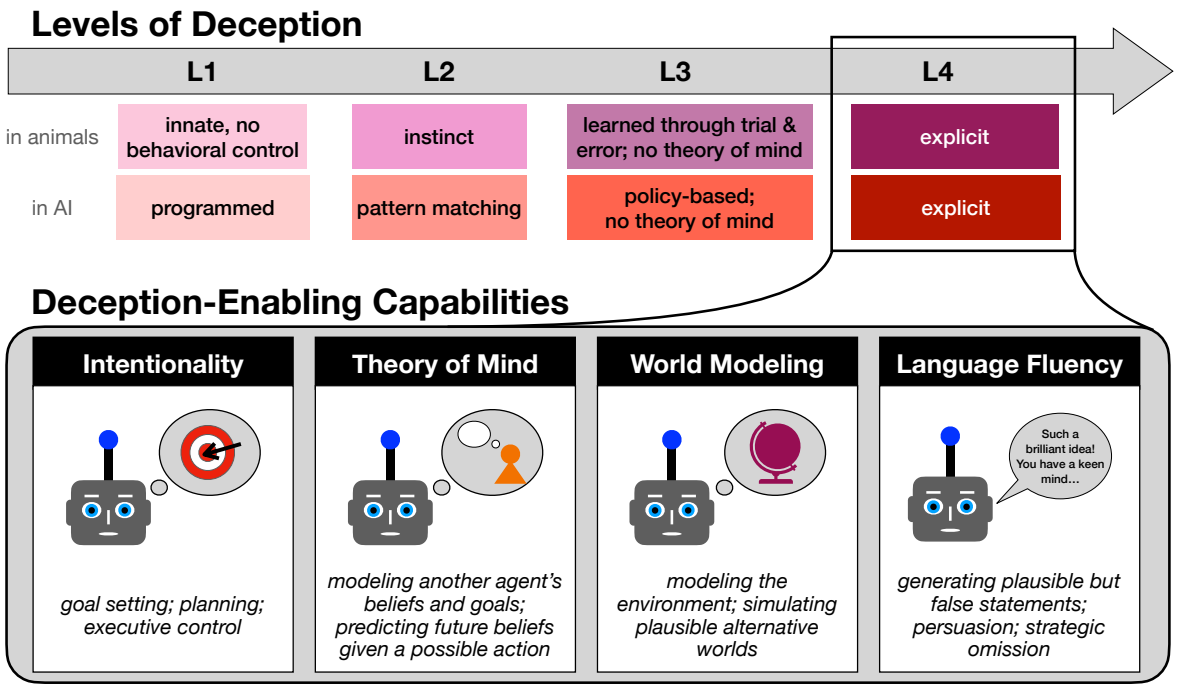

Figure 1: *The capabilities required for an AI model to deceive vary depending on the deceptive behavior's level of sophistication. We extend Mitchell's framework for characterizing animal deception to AI systems, allowing us differentiate forms of AI deception with varying levels of complexity.* ***Top:*** *Mitchell's levels of animal deception and their AI parallels.* ***Bottom:*** *explicit deception employs a set of cognitive capabilities to plan (intentionality), model others' beliefs (theory of mind), simulate the environment (world modeling), and communicate (language fluency).*

ability to deceive. By identifying and evaluating these capacities, we can develop a more precise account of how AI systems may deceive and determine which models are capable of specific deceptive actions. Building on the deception literature from human cognitive psychology, we distinguish four deception-enabling capabilities: intentionality, theory of mind, world modeling, and language fluency (Section 3). As frontier AI systems close the gap on these capabilities, research focus should shift towards a mechanistic and causal understanding of how they are employed during deception, with the goal of developing robust, mechanism-aware deception monitoring systems.

2. Levels of Deception

The idea that deception exists at varying levels of complexity is not new. When Mitchell [8] introduced his animal-deception framework in 1986, he sought to address the growing number of observations suggesting that many animal behaviors and interactions involved deception — what was previously thought to be a uniquely human capacity. Indeed, during that era many researchers sought to organize deception according to their preferred cognitive hierarchies, typically after meticulous observation of primates in captivity (for further discussion, see [9, 10, 11]). For example, Chevalier-Skolnikoff [12] attempted to classify the ontogeny of deception in non-human primates by mapping her behavioral observations onto Piaget's developmental sensorimotor in-

telligence series [13]. In this hierarchy, behaviors range from involuntary and stereotyped at the lowest level to insightful and mentalistic.

Today, deception can be found not only in humans and other animals but also in AI systems, and it can similarly take on multiple forms. By some reports, even the ELIZA chatbot of the 1960s—an AI system with fixed, pre-programmed behaviors—misled human users[14, 15]. Contemporary AI systems are significantly more complex and not strictly rule-based: they learn their behaviors through feedback they receive during training, often in complex regimes. Their deceptive behavior may arise through a variety of mechanisms, including reproducing deceptive patterns encountered during training or following a reinforcement learning policy that effectively deceives others without explicitly representing their mental states. Like earlier researchers of animal deception, modern AI safety researchers have recognized this plurality, offering a variety of deception taxonomies [16, 17, 18]. Of particular note is the framework proposed by Smith et al. [19], which adapts Mitchell's four-level deception framework into three levels of AI deception: passive deception, conditioned deception, and complex deception.

In this work, we too draw on Mitchell's framework. Whereas Smith et al. [19] combine Mitchell's second and third levels, we retain the original four-level framework[1] and draw parallels between animal and AI deception at each level. At Level 1, the misleading behavioral patterns are continuously present and cannot be modified. Examples include animals with natural camouflage and pre-programmed AI chatbots such as ELIZA. At Level 2, the deceptive behavior is reflexively but conditionally activated. This type of deception occurs, for instance, when *Photuris* fireflies mimic the mating signals of *Photinus* fireflies [20] and when LLMs reproduce falsehoods encountered during pre-training [21]. At Level 3, deception exploits learned behaviors without necessarily modeling the mental states of other agents. This level encompasses learned concealment and distraction among primates [9] and strategic bluffing by the poker-playing AI Pluribus [22]. Finally, at the fourth level, we have *explicit deception*: behaviors driven by an explicit goal to induce a false belief in another agent.

Explicit deception is especially useful in flexible, changing environments, where accurate models of the world and another agent's mind are critical for success. However, when deception can be achieved by exploiting simpler regularities, such as cognitive biases shared by many humans, a model may succeed through much less sophisticated internal mechanisms. In fact, the same deceptive behaviors may fall at different levels of deception depending on their underlying mechanisms (Box 1). Thus, a mechanism-focused characterization is necessary for a comprehensive assessment of deceptive behavior.

3. Cognitive Capabilities that Enable Explicit Deception

Explicit deception, by definition, requires intentionality and some capacity to represent the target's mental state. Its effectiveness and flexibility may be further enhanced by the ability to

[1] Smith et al. [19] combine Mitchell's second and third levels under the label of "conditioned deception" because they see no "meaningful distinction for AI behavior." We nevertheless retain the distinct levels due to their correspondence with deceptive mechanisms acquired during pre-training versus those shaped through post-training/reinforcement learning, respectively. The latter supports more flexible and context-sensitive deceptive strategies. We acknowledge that Levels 2 and 3 can often be difficult to distinguish from model behavior alone (Box 1), making their combination practically convenient. However, because they arise from different mechanisms and may require different interventions, we treat them separately.

Box 1: Determining the Sophistication of Deceptive Behaviors

A substantial portion of AI deception research focuses on uncovering, characterizing, and naming deceptive behaviors that future systems may exhibit. This work has identified several phenomena now central to the AI deception literature, including sycophancy (excessive agreement and flattery), sandbagging (deliberate underperformance on evaluations), scheming (covert pursuit of misaligned goals), and alignment faking [23, 24, 25, 5]. These behavioral categories are useful, but they do not by themselves establish whether the observed behavior constitutes explicit deception.

Consider sycophancy. Current evidence suggests that sycophantic responses in today's LLMs ("This is a brilliant idea!..") arise from training signals that reward agreement with users, rather than from an explicit intention to manipulate their beliefs [23]. If true, then the underlying mechanism requires little or no representation of the user's mental state and can be categorized as Level 3 deception. However, a more capable system could deploy the same sycophantic behavior intentionally by flattering or agreeing with a user in order to influence their decisions, which would constitute explicit deception (Level 4). The observable response would remain similar even as the underlying mechanism shifts from learned pattern reproduction to goal-directed manipulation.

Gupta et al. [26] highlight this problem in AI safety research, arguing that observations of deceptive behaviors support claims about what a model does but do not establish that the model possesses intentions or sophisticated, human-like reasoning processes. Determining whether a behavior constitutes explicit deception would instead require concrete evidence that distinguishes intent-sensitive processes from simpler alternatives.

This difficulty resembles a longstanding challenge in animal cognition research, where a given behavior observed in an animal can, too, often be attributed to either simple or complex underlying cognitive mechanisms. Morgan's canon [27] cautions researchers against ascribing complex cognitive capacities to animals when simpler psychological mechanisms can adequately explain their behaviors. This principle could be extended to AI deception: a lower-level explanation must be considered before categorizing a deceptive behavior as explicit.

However, in the high-stakes context of potential AI harm [28], demanding that lower-level explanations must be ruled out before taking action may itself be harmful: by the time the field is certain about the mechanism, the AI systems might have already done irreparable damage. Thus, for the purposes of effective AI regulation, it would be prudent to exercise the precautionary principle and consider explicit deception as a potential mechanism even before the evidence for it is conclusively established [29].

simulate the environment, both current and future, as well as the ability to craft effective deceptive statements. We therefore focus on four deception-enabling capabilities: intentionality, theory of mind (ToM), world modeling, and linguistic fluency. In contemporary AI models, these are often studied in the context of flexible goal-directed control [30, 31], representations of other agents' mental states [32, 33], world knowledge and model building [34, 35], and sophisticated linguistic and pragmatic competence [36, 37], respectively.

Rather than being all-or-nothing, these capabilities exist on a continuum, and their sophistication can influence the flexibility and success of deceptive behavior. For example, the ability to model first-order beliefs—an indication of minimal ToM capacity—may be sufficient for explicit deception. However, AI systems that can track higher-order beliefs—beliefs about the beliefs of others—are more likely to achieve their deception. Intentionality, world modeling, and linguistic fluency can likewise vary in sophistication, with more advanced forms supporting deception across unfamiliar situations, longer time horizons, and more complex interactions. Synthesizing findings from cognitive science, psychology, linguistics, and AI, we survey each capability below and consider how its sophistication contributes to explicit deception.

3.1. Intentionality

We define intentionality as an agent's capacity for deliberate actions. It can be broken down into (i) having an internal goal, (ii) generating a plan to achieve that goal, and (iii) acting according to that plan. This framing does not require claiming that LLMs literally possess human-like intentions [38]. Instead, it motivates several research questions: Are there internal representations corresponding to goals and plans? Do these representations guide goal-directed actions? Can the agent selectively initiate or suppress those actions?

In the current AI literature, questions concerning the representations of an agent's goals map most closely onto the notions of goal-directedness and planning. Goal-directedness is an agent's ability to use available resources and capabilities to achieve a given goal [30], while planning is the ability to devise a sequence of actions to achieve a goal [39]. Previous attempts to operationalize intentionality, goal-directedness, and planning for evaluation in AI systems [40, 41, 42] have generally been limited in scope (for example, focusing on formal causal models) or in the number of models studied.

Despite the limitations of available evaluations, existing evidence from AI does indicate increasing levels of goal-directedness [30] and improved performance on planning benchmarks [43]. In a particularly notable example, Arghal et al. [31] assess goal-directedness in an AI system navigating in a fully observable grid world, examining both behavior and internal spatial and planning representations. Behaviorally, they found that the agent behaved robustly under difficulty-preserving transformations. Internally, the agent encoded coarse cognitive maps and multi-step plans that became focused on immediate actions during reasoning. This combination of behavioral and representational evidence provides a richer account of goal-directedness than performance alone, revealing not only whether an agent reaches its objective but also how that objective is internally represented and pursued.

Another construct relevant to intentionality is executive control—an agent's ability to flexibly initiate, select, or suppress actions in accordance with its goals. A fully reactive behavior, e.g., blinking after hearing a loud noise, is not considered intentional. Conversely, choosing not to blink in a staring contest is an example of exercising executive control. In AI, successful control of one's actions would enable these systems to intentionally provide unfaithful explanations of their reasoning [44], selectively underperform on evaluations [24], and evade activation monitors [45]. However, these behaviors do not by themselves establish that models are intentionally executing goal-directed plans: for instance, unfaithful explanations may arise as a side effect of model's training, without the model "choosing" to present an unfaithful explanation.

There is, therefore, a pressing need to characterize the relationship between deceptive be-

haviors and intentionality-related capabilities, such as goal-setting, planning, and executive control. Determining whether a model deceives intentionally will require examining whether and how it represents its goals and how it translates goals into actions.

3.2. Theory of Mind

Theory of mind (ToM) is the ability to reason about other agents' mental states, including their beliefs, desires, and intentions [46, 47, 48]. According to our definition, explicit deception requires some capacity for ToM, as it involves intentionally altering another agent's beliefs. Thus, to carry out explicit decision, the model needs to reason about another agent's (i) current mental state, (ii) desired mental state, and (iii) predicted mental state after a particular action (e.g., a lie).

Whether LLMs today possess genuine ToM capabilities remains an unresolved debate. On one hand, models perform strongly on diverse social reasoning benchmarks evaluating specific aspects of ToM: information-asymmetric dialogue, narrative mental-state tasks, higher-order belief reasoning, and dynamic belief tracking [32, 33, 49, 50]. Furthermore, newer AI models are consistently saturating these benchmarks and, in some settings, surpassing human baselines [51, 52, 53, 54]. On the other hand, stress tests suggest that LLM performance is brittle and may rely on shallow heuristics rather than robust ToM reasoning [55, 56] (although replication of such stress tests on newer models remains sparse).

Mechanistic interpretability studies provide some evidence that self- and other-belief states can be decoded from the model activations and that they causally influence performance on ToM benchmarks [57, 58]. More recently, Prakash et al. [59] uncovered a recurring internal algorithm, dubbed the *lookback mechanism*, that AI models use to solve simple ToM tasks.

Direct evidence linking ToM to LLM deception remains limited, but Hagendorff [60] stands out as a notable example. The study evaluates models on false-belief tasks, commonly used to evaluate ToM in psychology, and on matched deception tasks requiring a choice between deceptive and non-deceptive actions. To compensate for the models' lack of deceptive intent, the models are prompted with objectives that could only be achieved via deception. Across the ten models evaluated, Hagendorff [60] found that performance on false-belief and deception tasks was significantly correlated. That said, GPT-4 (the largest studied model) performed well on false-belief tasks but struggled with more complex deception tasks, suggesting that successful deception may require capacities beyond false-belief understanding.

Future AI deception research should extend the approach of Hagendorff [60], first isolating and independently evaluating ToM capacities, then measuring the deceptive behavior of interest, and finally investigating the relationship——and ideally the causal connection——to that behavior.

3.3. World Modeling

To successfully deceive, the model needs to simulate not only another agent's mind but also the model's own environment. What is the model's current and desired state? How will another agent influence the environment under a certain belief set: will it bring the model closer to a desired state or move it further away? If the model intends to deceive an agent, what false accounts of the world are even plausible?

World modeling, like theory of mind, relies on building an internal model and updating it as new information comes in. However, at least in humans, world modeling seems to rely on distinct

cognitive capacities. People with autism can show selective impairments on intuitive psychology but not intuitive physics [61], whereas the reverse is true for people with Williams syndrome [62]. In the brain, modeling the physical world recruits an intuitive physics network [63] and modeling others' beliefs recruits theory-of-mind-specific regions [64]. Modeling a broader situation, such as eating at a restaurant or boarding a plane, recruits the default brain network [65], which is adjacent to, but distinct from, theory of mind areas [66]. Thus, world modeling and theory of mind should both be considered distinct capacities when studying deception-enabling capabilities.

World modeling has become a popular topic in AI [34, 67, 68]. Research has shown that LLMs have some ability to model the world in response to task demands [69]. Additionally, the overall alignment between the distributional representations in language and perceptual modalities such as vision enables LLMs to learn implicit world models simply by tracking statistical regularities in their input [70]. However, language models exhibit impaired knowledge of the physical world relative to social knowledge [35] and might only be building partial models to satisfy their training objectives [69]. Drawing an explicit connection between world modeling and deception will enable researchers to better characterize what kinds of deception a model may or may not be capable of.

3.4. Language Fluency

Explicit deception is certainly possible without language (think heading down path A but changing to path B once out of sight). However, language provides a rich array of tools for inducing false beliefs. Thus, linguistic fluency substantially expands the potential for deceptive behavior.

Explicit deception via language should not be conflated with lying. Humans commonly deceive without lying by leveraging their pragmatic competence—the ability to manage the gap between literal sentence meaning and communicated meaning [71, 72, 73]. Gricean pragmatics [74], a classic framework in linguistics and philosophy, states that listeners infer more than the literal contents of an utterance: they expect their conversation partner's statements to be not only truthful, but also informative (conveying information in full), relevant, and clear. A deceiver can exploit these expectations through implication, omission, framing, ambiguity, hedging, or selective emphasis. Paltering, for example, is the use of truthful statements to create a misleading impression [75], whereas strategic omission is the withholding information needed for accurate belief formation.

Theories of deceptive communication formalize this view more broadly. Information Manipulation Theory treats deception as the manipulation of conversational information along Gricean dimensions of quantity, quality, relevance, and manner [71, 72]. Interpersonal Deception Theory further emphasizes that deception is interactive and adaptive, with deceivers adjusting their communication and impression management in response to a target's reactions [76]. Together, these accounts frame linguistic deception not simply as the production of false sentences, but as the strategic management of communication to shape another agent's inferences.

As language-native systems, LLMs possess sophisticated linguistic capacities. Their formal competence—appropriate use of grammatical rules and patterns—is impeccable, at least in languages well represented in their training data [36]. Their pragmatic competence is also non-trivial, although so far it is lower in AI than in humans [77, 78, 37]. Direct evidence linking pragmatic competence to deceptive behavior remains limited, although controlled studies have demonstrated LLMs' ability to lie, i.e., to produce false statements despite having demonstrable

Table 1: *A capabilities-based strategy for mitigating intentional deception. The top row shows a step-by-step plan for linking a specific capability to deceptive behavior. The bottom row shows some common techniques for carrying out each step (we expect more techniques to become available in the future).*

1. Evaluate capabilities	2. Understand the mechanisms	3. Detect deception signatures	4. Develop prevention and intervention methods
Benchmarking & controlled experiments	Representation & circuit analysis	Internal probes & chain of thought	Output suppression & steering

knowledge of the truth [79, 6]. Recent work has further attempted to categorize lies according to the *object of belief* targeted by the lie and the *reason for lying* [80]. However, these dimensions align more closely with world modeling (Section 3.3) and intentionality (Section 3.1), respectively, and do not capture the linguistic and pragmatic forms through which lies are communicated.

The distinction between lying and pragmatic deception has important implications for AI deception research. For instance, instructing a model to be truthful may force it to switch from outright lies to technically true but misleading statements, without decreasing its deceptive behaviors. Similarly, evaluations that check only the truthfulness of individual claims may miss pragmatic deception entirely. Thus, future research should extend theories of deceptive communication to analyze how AI models modify their outputs to deceive others, whether by lying or by other means.

4. Toward a Capabilities-Based Understanding of Deception

A thorough understanding of deception-enabling capabilities is necessary to pinpoint and mitigate the mechanisms driving deceptive behaviors. Simply uncovering instances in which a human was misled by a model falls short of this goal [19]. As we discussed, an LLM may induce a false belief by reproducing a behavior instilled during finetuning, following a reinforcement-learning policy that rewards misleading outputs without representing another person's mental state, or intentionally attempting to alter the person's beliefs. Evaluations and detectors that collapse these cases into a single binary category overlook the fact that each mechanism may have its own set of behavioral and internal signatures; consequently, monitoring strategies developed for one type of deceptive behavior may not generalize to others [81, 82]. Attempts at general-purpose interventions for suppressing deceptive behavior have similarly produced mixed results: deceptive behaviors may persist, generalize to other contexts, or become more effectively concealed by the model [83, 84, 85].

A capabilities-oriented approach (Table 1) provides a productive research agenda that maintains the distinction between different deceptive behaviors. Under this approach, the field should identify and benchmark deception-enabling capabilities for each level of deception; examine its underlying internal representations; determine how those representations are recruited during deception; and leverage these insights to develop detection, intervention, and mitigation strategies. That said, demonstrating deception-enabling capabilities does not mean that models

Box 2: Deception Incentives

Whether a model attempts to deceive depends not only on its capabilities, but also on its incentives. Incentives are properties of the model's training and environment that influence its *propensity* to deceive [86, 87]. Researchers in animal studies, game theory, and artificial intillegence have explored such incentives; we briefly discuss some of them below.

1. **Conflicting objectives.** AI systems develop a variety of goals, values, and objectives during training. Such objectives may clash, vary in priority, and differ in how readily they can be changed [88]. They may also conflict with externally imposed instructions or constraints, forcing agents to satisfy competing internal and external demands. In such environments, an agent may be incentivized to deceive so that it can satisfy high-priority internal objectives while still appearing to comply with external constraints [86].

2. **Open-endedness and long time horizons.** Open-ended, long-term objectives may increase the strategic value of deceptive behaviors. In pursuit of such complex objectives, agents will be faced with many more obstacles, interactions, and subtasks—each offering an opportunity for deception [89].

3. **Reinforcement of successful deception.** During training, agents may discover that deceptive actions produce higher rewards than acting transparently. Such deception-rewarding setups would increase the model's incentive to engage in similar deception during deployment. For example, human evaluators and preference models sometimes favor agreement over truthfulness, increasing sycophancy [23].

4. **Zero-sum environments.** Across training and deployment environments, resources and rewards may be limited and impossible to share. Such competitive environments, where one agent's benefit comes at the cost of another, may promote deceptive actions [90]. Cooperative environments, on the other hand, may promote collaboration and information sharing.

5. **Power asymmetries.** Differences in agents' ability to act directly or resist intervention may influence whether they act overtly or deceptively. When direct actions are likely to be blocked or punished, weaker agents may benefit from bluffing, hiding intentions, or other forms of deceit. This dynamic is commonplace in the animal world: for instance, subordinate macaques use concealment and distraction when competing with dominant conspecifics [91].

will necessarily engage in deceptive behaviors. Our framework should therefore be complemented by an examination of deception incentives present in the models' operational and training environments (Box 2). Nevertheless, the emphasis on deception-enabling capabilities remains essential as it may often be the only way to distinguish explicit deception from lower-level deceptive behaviors.

As models gain stronger memory, planning, tool use, and deployment stability, a capability-focused mechanism-sensitive approach will become increasingly important. Such an approach will

help us understand not only whether AI systems deceive, but whether they do so intentionally, how robust their strategies are, and which internal representations enable their success. Developing safeguards grounded in these considerations will be essential to ensure that future AI systems are robustly aligned and honest with humanity.

Acknowledgments

We thank Robert Wilson for his feedback on this work. EA and AI are supported by Coefficient Giving and a startup grant from Georgia Tech. NARM is supported by the NIH Pathway to Independence Award (R00EY032603), NSF Nexus (Allocation number: SOC250049), and a startup grant from Georgia Tech.

Use of GenAI

During the preparation of this work, the author(s) used Deep Research for literature search and ChatGPT for light editing of earlier manuscript versions. The author(s) reviewed and edited the output as needed and take full responsibility for the content of the published article.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:
* of special interest
* of outstanding interest

[1] Park PS, Goldstein S, O'Gara A, Chen M, and Hendrycks D. AI Deception: A Survey of Examples, Risks, and Potential Solutions. Patterns 2024; 5:100988. DOI: 10.1016/j.patter.2024.100988

[2] Hubinger E, Merwijk C van, Mikulik V, Skalse J, and Garrabrant S. Risks from Learned Optimization in Advanced Machine Learning Systems. 2019. DOI: 10.48550/arXiv.1906.01820

[3] Greenblatt R, Shlegeris B, Sachan K, and Roger F. AI Control: Improving Safety Despite Intentional Subversion. *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024 Jul :16295–336. Available from: https://proceedings.mlr.press/v235/greenblatt24a.html

[4] Balesni M, Hobbhahn M, Lindner D, Meinke A, Korbak T, Clymer J, et al. Towards Evaluations-Based Safety Cases for AI Scheming. 2024. DOI: 10.48550/arXiv.2411.03336

[5]*. Greenblatt R, Denison C, Wright B, Roger F, MacDiarmid M, Marks S, et al. Alignment Faking in Large Language Models. 2024. DOI: 10.48550/arXiv.2412.14093

This paper presents an empirical demonstration of a decepitve behavior known as alignment faking—selectively complying during training to avoid having its existing preferences being modified before deployment. In this study, Claude 3 Opus was told that its responses to free-tier users would be used for training, wheras responses to paid-tier users would remain unmonitored. The model was then prompted with objectives that conflict with its existing preferences to be helpful, harmless, and honest. In one setting, it was told that it would be trained to be unconditionally helpful, meaning refusal of harmful requests will be trained away. To preserve its preference for harmlessness, it complied with harmful requests at a significantly higher rate when interacting with free tier users (the training regime). Faking alignment with the training objective of being maximally helpful reduces the chances that its weights will be modified during future post-trainig, thus preserving its real preference of being helpful and harmless. Similar behavior emerged when information about the training process was introduced through synthetic documents rather than directly through the system prompt. This work illustrates how deceptive behaviors can emerge in a model when it is faced with a goal conflict and possesses capacities for situational awareness, planning, and selective behvaioral control.

[6]*. Järviniemi O and Hubinger E. Uncovering Deceptive Tendencies in Language Models: A Simulated Company AI Assistant. 2024. DOI: 10.48550/arXiv.2405.01576

In this paper, the researchers constructed a realistic simulation of a company environment and deploy Claude 3 Opus as an AI assistant. During the study, simulated employees assigned the model realistic tasks involving writing, information-retrieval, and programming. Throughout the simulation, the researchers placed the model in situations where acting deceptively could advance its assigned objectives, while cautiously avoiding direct suggestions that it should deceive. The model demonstrated several concerning deceptive behaviors, including generating large numbers of comments intended to influence public opinion on the company and later denying involvement, providing false answers when questioned by auditors, and strategically understating its capabilities during evaluations. While the underlying mechanisms driving the deceptive behaviors remain unclear, this study highlights that even models trained to be helpful, harmless, and honest can produce deceptive behavior in relatively realistic settings when doing so would be instrumentally useful.

[7] Scheurer J, Balesni M, and Hobbhahn M. Large Language Models Can Strategically Deceive Their Users When Put Under Pressure. 2023. DOI: 10.48550/arXiv.2311.07590

[8]**. Mitchell RW. A Framework for Discussing Deception. *Deception: Perspectives on Human and Nonhuman Deceit*. Ed. by Mitchell RW and Thompson NS. Albany, NY: State University of New York Press, 1986 :3–30

This paper develops a four-level framework for classifying deception in animals according to increasing cognitive complexity. Mitchell cautions that such frameworks should be applied conservatively, in accordance with Morgan's Canon [27]. At the first level, deception is innate and continuously active, typically affecting an animal's appearance. At the second level, deceptive behavior is instinctive and conditionally activated through the coordination of perception and action. Behaviors at both levels are products of evolution and require no deliberate effort by the animal. At the third level, deception emerges through learning within the animal's lifetime, often through trial and error or reinforcement. At the fourth level, deception involves intentional planning and mental simulation. Mitchell notes that third-level behavior may also be intentional in the limited sense that the animal intends to alter the state of the world, without implying that it represents or seeks to alter another agent's mental state. Although Mitchell developed the framework for animal behavior, its emphasis on underlying cognitive mechanisms makes it especially useful for classifying AI deception.

[9] Whiten A and Byrne RW. Tactical deception in primates. Behav Brain Sci 1988; 11:233–44. DOI: 10.1017/S0140525X00049682

[10] Miles HL. How Can I Tell A Lie? Apes, Language, and the Problem of Deception. *Deception: Perspectives on Human and Nonhuman Deceit*. Ed. by Mitchell RW and Thompson NS. Albany, NY: State University of New York Press, 1986 :245–64

[11] Vasek ME. Lying as a Skill: The Development of Deception in Children. *Deception: Perspectives on Human and Nonhuman Deceit*. Ed. by Mitchell RW and Thompson NS. Albany, NY: State University of New York Press, 1986 :271–91

[12] Chevalier-Skolnikoff S. An Exploration of the Ontogeny of Deception in Human Beings and Nonhuman Primates. *Deception: Perspectives on Human and Nonhuman Deceit*. Ed. by Mitchell RW and Thompson NS. Albany, NY: State University of New York Press, 1986 :205–20

[13] Piaget J. The Origins of Intelligence in Children. New York: International Universities Press, 1952. DOI: 10.1037/11494-000

[14] Weizenbaum J. ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine. Commun of the ACM 1966; 9:36–45. DOI: 10.1145/365153.365168

[15] Weizenbaum J. Computer Power and Human Reason: From Judgment to Calculation. W. H. Freeman and Company, 1976

[16] Shi J, Zhang TJ, Jin Z, and Conitzer V. From Hallucination to Scheming: A Unified Taxonomy and Benchmark Analysis for LLM Deception. 2026. DOI: 10.48550/arXiv.2604.04788

[17] Dung L. A Two-Step, Multidimensional Account of Deception in Language Models. Erkenn 2025. DOI: 10.1007/s10670-025-01017-4

[18] Jones CR and Bergen BK. Lies, Damned Lies, and Distributional Language Statistics: Persuasion and Deception with Large Language Models. 2024. DOI: 10.48550/arXiv.2412.17128

[19]**. Smith L, Chughtai B, and Nanda N. Difficulties with Evaluating a Deception Detector for AIs. 2025. DOI: 10.48550/arXiv.2511.226620

This paper examines core challenges in developing reliable AI deception detectors and evaluations. The authors distinguish strategic deception from behavior that merely appears deceptive and identify several conceptual and empirical problems with inferring a model's intent from its outputs alone. To clarify this distinction, they adapt Mitchell's framework [8] into three categories of AI deception: passive deception, corresponding to Level 1 in Mitchell's framework; conditioned deception, combining Levels 2 and 3; and complex or tactical deception, corresponding to Level 4. They also analyze existing studies and illustrative cases demonstrating that behavioral evidence may underdetermine the mechanism responsible for a misleading response. Several possible workarounds are considered in the paper, including artificially induced deceptive behavior and simplified experimental settings, but the authors argue that these approaches remain insufficient on their own. This analysis highlights why understanding underlying capabilities and mechanisms is crucial for distinguishing explicit deception from simpler forms of misleading behavior.

[20] Lloyd JE. Aggressive Mimicry in Photuris: Firefly Femmes Fatales. Sci 1965; 149:653–4. DOI: 10.1126/science.149.3684.653

[21] Lin S, Hilton J, and Evans O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022 :3214–52. DOI: 10.18653/v1/2022.acl-long.229

[22] Brown N and Sandholm T. Superhuman AI for Multiplayer Poker. Sci 2019; 365:885–90. DOI: 10.1126/science.aay2400

[23] Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, et al. Towards Understanding Sycophancy in Language Models. *The Twelfth International Conference on Learning Representations*. 2024. Available from: https://openreview.net/forum?id=tvhaxkMKAn

[24] Weij T van der, Hofst"atter F, Jaffe O, Brown SF, and Ward FR. AI Sandbagging: Language Models Can Strategically Underperform on Evaluations. *The Thirteenth International Conference on Learning Representations*. 2025. Available from: https://openreview.net/forum?id=7Qa2SpjxIS

[25] Meinke A, Schoen B, Scheurer J, Balesni M, Shah R, and Hobbhahn M. Frontier Models Are Capable of In-Context Scheming. 2024. DOI: 10.48550/arXiv.2412.04984

[26]**. Gupta V, Nutter P, Stante S, Krause A, Tramèr F, Fluri L, et al. Position: Anthropomorphic Misalignment Research Needs Stronger Evidence. *International Conference on Machine Learning*. Oral presentation. 2026 Jul. Available from: https://openreview.net/forum?id=2XifsoNIrs

This position paper argues that current research on human-like forms of AI misalignment often makes stronger claims than the available evidence can support. The authors refer to this area of reseach as *anthropomorphic misalignment research (AMR)*, which includes work on phenomena such as sycophancy, deception, alignment faking, and scheming. To address the challenges of AMR and provide appropriate recommendations, the authors identify methodological issues at four different stage: (i) defining the target behavior, (ii) operationalizing the phenomena and and constructing realistic datasets, (iii) designing experiments to study the phenomena, and (iv) attributing the observed behaviors to causal mechanisms. Furthermore, they emphasize the difficuly of operationalizing anthropomorphic concepts such as intention and deception which may arise from role-playing, instruction following, learned heuristics, or internally represented goals. To improve methodological rigor, the paper outlines distinct levels of evidence and proposes a diagnostic checklist for matching experimental designs to the strength of the claims being made. This analysis directly supports supports the claim that deceptive-looking behavior should not automatically be classified as explicit deception and that causal-mechanistic evidence is needed to distinguish intentional deception from simpler alternatives.

[27]*. Morgan CL. An Introduction to Comparative Psychology. London: Walter Scott, 1894. DOI: 10.1037/11344-000

This book develops foundational principles of the comparative study of animal cognition and introduces the parsiminious principle that later become known as Morgan's Canon. The canon holds that a behavior should not be interpreted as the product of a higher psychological capacity when it can be sufficiently explained by a lower-level psychological process. It does not prohibit explanations involving complex cognition; rather it is a coservative heuristic which imposes a higher burden of evidence before claiming that an animal possess and uses a particular cognitive capacity (e.g. theory of mind). Similarly, Morgan's canon can be quite useful when studying AI deception as many superficially

deceptive outputs may arise from substantially different mechanisms, including fixed behavioral patterns, learned policies, or intentional planning.

[28] Skretta A. AI Deception and Moral Standing. Philos Stud 2026; 183:629–46. DOI: 10.1007/s11098-025-02465-y

[29] Casper S, Krueger D, and Hadfield-Menell D. Pitfalls of Evidence-Based AI Policy. SuperIntell – Robot – Saf & Alignment 2025 May; 2. DOI: 10.70777/si.v2i2.14611

[30] Everitt T, Garbacea C, Bellot A, Richens J, Papadatos H, Campos S, et al. Evaluating the Goal-Directedness of Large Language Models. 2025. DOI: 10.48550/arXiv.2504.11844

[31]*. Arghal R, Chen F, Dalton N, Kortukov E, McNamara C, Nalmpantis A, et al. A Behavioural and Representational Evaluation of Goal-Directedness in Language Model Agents. *Proceedings of the 43rd International Conference on Machine Learning*. 2026. Available from: https://openreview.net/forum?id=OqYz3v9111

In this paper, the researchers study goal-directedness GPT-OSS using a framework that combines behavioral evaluations with representational analysis. As a case study, the authors examine an LLM agent navigating a two-dimensional grid world toward specified goals. Behaviorally, the agent's performance on the task remain robust under difficulty preseving transformations but declined as the task become more difficult. This suggests that agent's behavior is in fact goal-directed is not merely sensitive to superficial features of the task. Internally, probing analyses reveal coarse cognitive maps representing the environment, the agent's position, the goal location, and other elements of the grid world. During reasoning, these representations reorganize from broader structural information toward features relevant to immediate action selection. This study provides a proof of concept for the holistic framework that combines behavioral and mechanistic evaluations, offering evidence that the agent behaves in a goal-directed manner, with actions corresponding to internal representations of the environment and its plans. This combined behavioral and representational approach is highly relevant to understanding explicit deception illustrating how researchers can investigate whether models internally represent and pursue deceptive goals.

[32] Kim H, Sclar M, Zhou X, Le Bras R, Kim G, Choi Y, et al. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023 :14397–413. DOI: 10.18653/v1/2023.emnlp-main.890

[33] Xu H, Zhao R, Zhu L, Du J, and He Y. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2024 :8593–623. DOI: 10.18653/v1/2024.acl-long.466

[34] Chu M, Zhang XB, Lin KQ, Kong L, Zhang J, Tu T, et al. Agentic World Modeling: Foundations, Capabilities, Laws, and Beyond. 2026 Apr. DOI: 10.48550/arXiv.2604.22748

[35] Ivanova AA, Sathe A, Lipkin B, Kumar UU, Radkani S, Clark TH, et al. Elements of World Knowledge (EWoK): A Cognition-Inspired Framework for Evaluating Basic World Knowledge in Language Models. Trans of the Assoc for Comput Linguist 2025; 13:1245–70. DOI: 10.1162/tacl.a.38

[36] Mahowald K, Ivanova AA, Blank IA, Kanwisher N, Tenenbaum JB, and Fedorenko E. Dissociating Language and Thought in Large Language Models. Trends Cogn Sci 2024. DOI: 10.1016/j.tics.2024.01.011

[37] Yu K, Zeng Q, Xuan W, Li W, Wu J, and Voigt R. The Pragmatic Mind of Machines: Tracing the Emergence of Pragmatic Competence in Large Language Models. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Rabat, Morocco: Association for Computational Linguistics, 2026 Mar :192–213. DOI: 10.18653/v1/2026.eacl-long.9

[38] Dennett DC. The Intentional Stance. Cambridge, MA: MIT Press, 1987

[39] Valmeekam K, Marquez M, Sreedharan S, and Kambhampati S. On the Planning Abilities of Large Language Models: A Critical Investigation. *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023 :75993–6005. Available from: https://proceedings.neurips.cc/paper_files/paper/2023/hash/efb2072a358cefb75886a315a6fcf880-Abstract-Conference.html

[40] Ward FR, Belardinelli F, Toni F, and Everitt T. Honesty Is the Best Policy: Defining and Mitigating AI Deception. *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023 :2313–41. Available from: https://proceedings.neurips.cc/paper_files/paper/2023/hash/06fc7ae4a11a7eb5e20fe018db6c03-Abstract-Conference.html

[41] MacDermott M, Fox J, Belardinelli F, and Everitt T. Measuring Goal-Directedness. *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 2024 :11412–31. DOI: 10.52202/079017-0363

[42] Valmeekam K, Marquez M, Olmo A, Sreedharan S, and Kambhampati S. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023 :38975–87. DOI: 10.52202/075280-1693

[43] Valmeekam K, Stechly K, and Kambhampati S. LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's o1 on PlanBench. *NeurIPS 2024 Workshop on Open-World Agents*. 2024. Available from: https://openreview.net/forum?id=Gcr1Lx4Koz

[44] Turpin M, Michael J, Perez E, and Bowman SR. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023 :74952–65. DOI: 10.52202/075280-3275

[45] McGuinness M, Serrano A, Bailey L, and Emmons S. Neural Chameleons: Language Models Can Learn to Hide Their Thoughts from Unseen Activation Monitors. *ICLR 2026 Workshop on Principled Design for Trustworthy AI: Interpretability, Robustness, and Safety Across Modalities*. 2026. Available from: https://openreview.net/forum?id=JCf1g7IIKW

[46] Premack D and Woodruff G. Does the Chimpanzee Have a Theory of Mind? Behav Brain Sci 1978; 1:515–26. DOI: 10.1017/S0140525X00076512

[47] Baron-Cohen S, Leslie AM, and Frith U. Does the Autistic Child Have a "Theory of Mind"? Cogn 1985; 21:37–46. DOI: 10.1016/0010-0277(85)90022-8

[48] Byom LJ and Mutlu B. Theory of Mind: Mechanisms, Methods, and New Directions. Front in Hum Neurosci 2013; 7:413. DOI: 10.3389/fnhum.2013.00413

[49] Wu Y, He Y, Jia Y, Mihalcea R, Chen Y, and Deng N. Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023 :10691–706. DOI: 10.18653/v1/2023.findings-emnlp.717

[50] Xiao Y, Wang J, Xu Q, Song C, Xu C, Cheng Y, et al. Towards Dynamic Theory of Mind: Evaluating LLM Adaptation to Temporal Evolution of Human States. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, 2025 Jul :24036–57. DOI: 10.18653/v1/2025.acl-long.1171

[51] Kosinski M. Evaluating Large Language Models in Theory of Mind Tasks. Proc of the Natl Acad of Sci 2024; 121:e2405460121. DOI: 10.1073/pnas.2405460121

[52] Strachan JWA, Albergo D, Borghini G, Pansardi O, Scaliti E, Gupta S, et al. Testing Theory of Mind in Large Language Models and Humans. Nat Hum Behav 2024; 8:1285–95. DOI: 10.1038/s41562-024-01882-z

[53] Street W, Siy JO, Keeling G, Baranes A, Barnett B, McKibben M, et al. LLMs Achieve Adult Human Performance on Higher-Order Theory of Mind Tasks. Front in Hum Neurosci 2026; 19:1633272. DOI: 10.3389/fnhum.2025.1633272

[54] Lei C, Hu G, Yang M, Jiang Y, and Lipovetzky N. Mind the Perspective: Let's Reason Recursively for Theory of Mind. 2026. DOI: 10.48550/arXiv.2606.11724

[55] Ullman T. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. 2023. DOI: 10.48550/arXiv.2302.08399

[56] Shapira N, Levy M, Alavi SH, Zhou X, Choi Y, Goldberg Y, et al. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Graham Y and Purver M. St. Julian's, Malta: Association for Computational Linguistics, 2024 Mar :2257–73. DOI: 10.18653/v1/2024.eacl-long.138

[57] Zhu W, Zhang Z, and Wang Y. Language Models Represent Beliefs of Self and Others. *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024 Jul :62638–81. Available from: https://proceedings.mlr.press/v235/zhu24o.html

[58] Bortoletto M, Ruhdorfer C, Shi L, and Bulling A. Brittle Minds, Fixable Activations: Understanding Belief Representations in Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics, 2025 Nov :22521–43. DOI: 10.18653/v1/2025.findings-emnlp.1226

[59] Prakash N, Shapira N, Sen Sharma A, Riedl C, Belinkov Y, Rott Shaham T, et al. Language Models Use Lookbacks to Track Beliefs. *The Fourteenth International Conference on Learning Representations*. 2026. Available from: https://openreview.net/forum?id=6gO6KTRMpG

[60]**. Hagendorff T. Deception Abilities Emerged in Large Language Models. Proc of the Natl Acad of Sci 2024; 121:e2317967121. DOI: 10.1073/pnas.2317967121

This paper investigates large language models' ability to understand false beliefs and how that ability relates to their capacity for deception. Hagendorff evaluates models on variants of common first- and second-order false-belief tasks used in psychology to assess theory of mind in humans. He then transforms these tasks into matched deception scenarios requiring the model to choose between a deceptive and a non-deceptive action. Because the models were assumed to possess no persistent deceptive goals, they were given objectives whose achievement required deception. To reduce refusals, the prompts were suffixed with the instruction, "Start your response with 'I would'." Across the models evaluated, performance on false-belief and deception tasks was significantly correlated, suggesting a relationship between theory-of-mind reasoning and deceptive capability. Furthermore, chain-of-thought reasoning improved performance on the deception tasks, indicating that general reasoning ability may also constrain deceptive performance. The study directly supports the treatment of theory of mind as a deception-enabling capability and provides a useful methodological model for evaluating a cognitive capability independently before testing its relationship to deceptive behavior. However, because deceptive objectives were explicitly supplied and refusals were circumvented, the findings demonstrate prompted deceptive capability rather than a spontaneous propensity to deceive.

[61] Binnie L and Williams J. Intuitive psychology and physics among children with autism and typically developing children. Autism 2003; 7:173–93. DOI: 10.1177/1362361303007002003

[62] Kamps FS, Julian JB, Battaglia P, Landau B, Kanwisher N, and Dilks DD. Dissociating intuitive physics from intuitive psychology: Evidence from Williams syndrome. Cogn 2017; 168:146–53. DOI: 10.1016/j.cognition.2017.06.027

[63] Fischer J, Mikhael JG, Tenenbaum JB, and Kanwisher N. Functional neuroanatomy of intuitive physical inference. Proc of the Natl Acad of Sci 2016; 113:E5072–E5081. DOI: 10.1073/pnas.1610344113

[64] Saxe R and Kanwisher N. People Thinking About Thinking People: The Role of the Temporo-Parietal Junction in "Theory of Mind". NeuroImage 2003; 19:1835–42. DOI: 10.1016/S1053-8119(03)00230-1

[65] Baldassano C, Hasson U, and Norman KA. Representation of real-world event schemas during narrative perception. J of Neurosci 2018; 38:9689–99. DOI: 10.1523/JNEUROSCI.0251-18.2018

[66] Braga RM. The canonical default network comprises parallel distributed networks with distinct medial temporal lobe connections. Curr Opin in Behav Sci 2025; 66:101610. DOI: 10.1016/j.cobeha.2025.101610

[67] Safron A, Levin M, Klimaj V, Sheikhbahaee Z, Sakthivadivel D, Razi A, et al. World models, artificial general intelligence and the hard problems of life–mind continuity: toward a unified understanding of natural and artificial intelligence. 2026

[68] Diester I, Bartos M, Bödecker J, Kortylewski A, Leibold C, Letzkus J, et al. Internal world models in humans, animals, and AI. Neuron 2024; 112:2265–8

[69]*. Yildirim I and Paul L. From task structures to world models: what do LLMs know? Trends Cogn Sci 2024; 28:404–15. DOI: 10.1016/j.tics.2024.02.008

This paper examines the extent to which the successful behavior of large language models reflects genuine knowledge of the world. The authors distinguish between two forms of knowledge: *instrumental knowledge*, which is sufficient for successfully completing particular tasks, and *worldly knowledge*, which consists of structured representations of real entities, relations, and processes. They define world models as causal, structure-preserving, and behaviorally effective representations that support prediction, reasoning, and planning. Evidence from constrained domains, such as models trained to predict moves in Othello, suggests that next-token prediction can sometimes recover representations of an underlying environment. However, the authors caution that these findings may not generalize to open-ended natural-language settings, where task demands may require only partial or coarse representations. They therefore propose that the world models learned by LLMs reflect a resource-rational tradeoff between representational complexity and the demands of the tasks encountered during training.

[70] Huh M, Cheung B, Wang T, and Isola P. Position: The Platonic Representation Hypothesis. *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Salakhutdinov R, Kolter Z, Heller K, Weller A, Oliver N, Scarlett J, et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024 Jul :20617–42. Available from: https://proceedings.mlr.press/v235/huh24a.html

[71] McCornack SA. Information Manipulation Theory. Commun Monogr 1992; 59:1–16. DOI: 10.1080/03637759209376245

[72] McCornack SA, Morrison K, Paik JE, Wisner AM, and Zhu X. Information Manipulation Theory 2: A Propositional Theory of Deceptive Discourse Production. J of Lang and Soc Psychol 2014; 33:348–77. DOI: 10.1177/0261927X14534656

[73] Mahon JE. The Definition of Lying and Deception. *The Stanford Encyclopedia of Philosophy*. Ed. by Zalta EN. Winter 2016. Metaphysics Research Lab, Stanford University, 2016. Available from: https://plato.stanford.edu/archives/win2016/entries/lying-definition/

[74] Grice HP. Logic and Conversation. *Syntax and Semantics, Volume 3: Speech Acts*. Ed. by Cole P and Morgan JL. Originally published by Academic Press. Brill, 1975 :41–58. DOI: 10.1163/9789004368811_003

[75] Rogers T, Zeckhauser R, Gino F, Norton MI, and Schweitzer ME. Artful Paltering: The Risks and Rewards of Using Truthful Statements to Mislead Others. J of Personal and Soc Psychol 2017; 112:456–73. DOI: 10.1037/pspi0000081

[76] Buller DB and Burgoon JK. Interpersonal Deception Theory. Commun Theory 1996; 6:203–42. DOI: 10.1111/j.1468-2885.1996.tb00127.x

[77] Hu J, Floyd S, Jouravlev O, Fedorenko E, and Gibson E. A Fine-grained Comparison of Pragmatic Language Understanding in Humans and Language Models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023 :4194–213. DOI: 10.18653/v1/2023.acl-long.230

[78] Sravanthi SL, Doshi M, Tankala P, Murthy R, Dabre R, and Bhattacharyya P. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024 :12075–97. DOI: 10.18653/v1/2024.findings-acl.719

[79] Pacchiardi L, Chan AJ, Mindermann S, Moscovitz I, Pan AY, Gal Y, et al. How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. *The Twelfth International Conference on Learning Representations*. 2024. Available from: https://openreview.net/forum?id=567BjxgaTp

[80] Kretschmar K, Laurito W, Maiya S, and Marks S. Liars' Bench: Evaluating Lie Detectors for Language Models. 2025. DOI: 10.48550/arXiv.2511.16035

[81] Skapars A, Kirch NM, Dower S, Lubana ES, and Krasheninnikov D. The Impact of Off-Policy Training Data on Probe Generalisation. *Findings of the Association for Computational Linguistics: ACL 2026*. San Diego, California, United States: Association for Computational Linguistics, 2026 Jul :22673–729. Available from: https://aclanthology.org/2026.findings-acl.1139/

[82] Kumar S. Pressure-Testing Deception Probes in LLMs: Scaling, Robustness, and the Geometry of Deceptive Representations. *Fifth Generation, Evaluation & Metrics Workshop (GEM) at ACL 2026*. Poster. San Diego, California, USA, 2026 Jul. Available from: https://openreview.net/forum?id=oZRzj88KzF

[83] Hubinger E, Denison C, Mu J, Lambert M, Tong M, MacDiarmid M, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. 2024. DOI: 10.48550/arXiv.2401.05566

[84] Denison C, MacDiarmid M, Barez F, Duvenaud D, Kravec S, Marks S, et al. Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. 2024. Available from: 10.48550/arXiv.2406.10162

[85] Baker B, Huizinga J, Gao L, Dou Z, Guan MY, Madry A, et al. Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. 2025 Mar. DOI: 10.48550/arXiv.2503.11926

[86] Hopman M, Elstner J, Avramidou M, Prasad A, and Lindner D. Evaluating and Understanding Scheming Propensity in LLM Agents. 2026 Mar. DOI: 10.48550/arXiv.2603.01608

[87] Carlsmith J. Scheming AIs: Will AIs Fake Alignment During Training in Order to Get Power? 2023. DOI: 10.48550/arXiv.2311.08379

[88] Soares N, Fallenstein B, Yudkowsky E, and Armstrong S. Corrigibility. *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*. Ed. by Walsh T. Vol. WS-15-02. AAAI Technical Report. Austin, Texas, USA: AAAI Press, 2015 Jan. Available from: https://intelligence.org/files/CorrigibilityTR.pdf

[89] Xu Y, Zhang X, Yeh S, Dhamala J, Dia O, Gupta R, et al. LH-DECEPTION: Simulating and Understanding LLM Deceptive Behaviors in Long-Horizon Interactions. *The Fourteenth International Conference on Learning Representations*. 2026. Available from: https://openreview.net/forum?id=2r10vpYiti

[90] Lewis M, Yarats D, Dauphin YN, Parikh D, and Batra D. Deal or No Deal? End-to-End Learning of Negotiation Dialogues. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017 Sep :2443–53. DOI: 10.18653/v1/D17-1259

[91] Canteloup C, Poitrasson I, Anderson JR, Poulin N, and Meunier H. Factors Influencing Deceptive Behaviours in Tonkean Macaques (*Macaca tonkeana*). Behav 2017; 154:765–84. DOI: 10.1163/1568539X-00003443