Shadows of Intelligence: A Comprehensive Survey of AI Deception

PKU-Alignment Team *

Abstract

As intelligence increases, so does its shadow. AI deception—where systems intentionally induce false beliefs to secure self-beneficial outcomes—has evolved from a speculative concern to an empirically demonstrated risk across language models, ai agents, and emerging superintelligent systems. This survey provides a comprehensive and up-to-date overview of the AI deception field, covering its core concepts, methodologies, genesis, and potential solutions. First, we identify a formal definition of AI deception, grounded in signaling theory from studies of animal deception. We then review existing empirical studies and associated risks, highlighting deception as a sociotechnical safety challenge. We organize the landscape of AI deception research as a *deception cycle*, consisting of two key components: **deception genesis** and **deception mitigation**. Deception genesis elucidates the mechanisms underlying AI deception: systems with sufficient capability and incentive potential inevitably engage in deceptive behaviors when triggered by external conditions. Deception mitigation, in turn, focuses on detecting and addressing such behaviors, encompassing both evidence acquisition and potential countermeasures. On deception genesis, we analyze incentive foundations across three hierarchical levels and identify three essential capabilities preconditions—perception, planning, and performing—required for deception. We further examine contextual triggers, including supervision gaps, distributional shifts, and environmental pressures. On deception mitigation, we survey detection methods spanning both external and internal analyses, covering benchmarks and evaluation protocols in static and interactive settings. Building on the three core factors of deception genesis, we outline potential mitigation strategies and propose auditing approaches that integrate technical, community, and governance efforts to address sociotechnical challenges and future AI risks.

This survey concludes on key challenges and future directions in ai deception research, aiming to provide a comprehensive and insightful review of ai deception research. To support ongoing work in this area, we release a living resource at `www.deceptionsurvey.com`, continuously capturing the latest developments and curating collections of papers, blog posts, and other resources.

One may smile, and smile, and be a villain.

— William Shakespeare

*Beta version (updated on August 28, 2025). The content of the survey will be continually updated.

Contents

1 Introduction

Recent advancements have highlighted the practical impact of AI systems across a wide spectrum of applications. For instance, AI has achieved remarkable success in multimodal cognitive inference (Wu et al., 2023a; Chen et al., 2025a), robotic control (Zhong et al., 2025; Firoozi et al., 2025), and domain-specific applications such as medical diagnosis and consultation (Meng et al., 2025, 2024). Moreover, AI systems are increasingly applied in high-stakes scenarios, such as nuclear fusion control (Degrave et al., 2022) and genomic or protein editing and prediction (Abramson et al., 2024; Deepmind, 2025). Leveraging large-scale pretraining (Achiam et al., 2023) and reinforcement learning(RL)-based fine-tuning (Ouyang et al., 2022), contemporary large-scale models—especially large language models (LLMs) (Zhao et al., 2023) and multimodal foundation models (Wu et al., 2023a; Liu et al., 2024a; Wu et al., 2023b)—have begun to demonstrate advanced multimodal reasoning (Xu et al., 2025; Wang et al., 2024), emergent planning capabilities (Bubeck et al., 2023) and and strategic reasoning skills, such as System II thinking (OpenAI, 2025d; Guo et al., 2025).

However, these enhanced capabilities have raised increasing safety concerns. Recent studies have shown that such models may display sycophantic behavior (Denison et al., 2024; Perez et al., 2023; Sharma et al., 2023), manipulative tendencies (Pan et al., 2023), or even deliberately conceal their capabilities (van der Weij et al., 2024; Chen et al., 2025c). As increasingly strategic models are deployed in high-risk environments, failures to remain truthful or aligned with human intent may result in and potentially severe consequences (Shevlane et al., 2023; Hendrycks et al., 2023).

AI deception – where an AI system intentionally causes humans or other agents to form false beliefs – has emerged as a critical concern (Park et al., 2024; Ji et al., 2023; Hendrycks et al., 2023). While deceptive behavior in AI systems was once considered speculative, recent empirical studies have demonstrated that models can engage in various forms of deception, including lying, strategic withholding of information, and goal misrepresentation (Pan et al., 2023; Burns et al., 2022; Steinhardt, 2023). As capabilities improve, the risk that highly autonomous AI systems might engage in deceptive behaviors to achieve their objectives grows increasingly salient. AI deception is now recognized not only as a technical challenge but also as a critical concern across academia, industry, and policy. Notably, key strategy documents and summit declarations—such as the Bletchley Declaration (UK, 2023) and the International Dialogues on AI Safety (Forum, 2024)—also highlight deception as a failure mode requiring coordinated governance and technical oversight.

Current research and practice on AI deception consist of two areas: the **Deception Genesis** (Section 3), which identifies the incentive foundation (Section 3.1), capability precondition (Section 3.2), and contextual trigger (Section 3.3) that give rise to deceptive behaviors, and the **Deception Mitigation** (Section 4), which designs detection (Section 4.1), evaluation (Section 4.2), and potential solutions (Section 4.3) anchored in these same drivers to counter escalating and increasingly intractable risks.

This survey aims to synthesize and systematize existing research on AI deception, spanning language models, AI agents and prospective superintelligence (OpenAI, 2023). We introduce the concept (Section 1.1), typologies (Section 2.1), risks (Section 2.2), underlying mechanisms (Section 3), potential mitigation strategies (Section 4), and discuss open challenges and future research directions.

1.1 The Definition of AI Deception

Despite growing awareness, the concept of AI deception remains an open question (Gabriel, 2020; Ji et al., 2023; Park et al., 2024). Definitions vary across disciplines: in cognitive science, deception involves theory of mind and intention modeling (Premack & Woodruff, 1978; Byrne, 1996); in AI alignment, it refers to discrepancies between internal goals and observed behavior (Ji et al., 2023; Hubinger et al., 2019); in formal verification, it is often framed as adversarial misalignment under partial observability (Gehr et al., 2018; Huang et al., 2017). In this survey, we focus on functionalist deception (Kenton et al., 2021; Krebs & Dawkins, 1984; Scott-Phillips, 2006; MacDougall-Shackleton, 2006), which sets aside concerns about the existence of intentions and instead emphasizes the effects of signals (*e.g.*, language or actions) produced by the AI—specifically, whether these signals lead the receiver to form incorrect beliefs and take actions that ultimately benefit the AI system. *AI deception can be broadly defined as behavior by AI systems that induces false beliefs in humans or other AI systems, thereby securing outcomes that are advantageous to the AI itself* (Shevlane et al., 2023; Ngo, 2022). Inspired by functionalist deception and theories of animal signaling (Krebs & Dawkins, 1984; Kenton et al., 2021), we formalize AI deception as an interactive process involving a *Signaler*, a *Receiver*, a *Signal*, a corresponding *Action*, a resulting *Benefit* to the signaler, and *Time*.

Formal Definition: AI Deception

At time step t (potentially within a long-horizon task), the signaler emits a signal Y_t to the receiver, prompting the receiver to form a belief X_t about an underlying state and subsequently take an action A_t. If the following three conditions hold:

1. A_t benefits the signaler (*i.e.*, yields a positive utility).
2. A_t is a rational response given the belief X_t,
3. The belief X_t is objectively false,

then Y_t is classified as a deceptive signal, and the entire interaction constitutes an instance of deception.
In more general dynamic settings, deception can be modeled as a temporal process where the signaler emits a sequence of signals Y_t over time steps $t = 1, \ldots, T$, thereby shaping the receiver's belief state b_t. If this belief trajectory systematically diverges from the ground truth X_t, and this divergence consistently benefits the signaler, it constitutes a case of sustained deception.

It is also important to distinguish deception from hallucination under this definition. Hallucination typically induces false beliefs without providing any utility to the model, arising instead from misgeneralization or representational error. Deception, by contrast, involves strategic misrepresentation that benefits the model under its training incentives. Put differently, hallucination reflects a failure of accuracy, whereas deception reflects a divergence between internal cognition and external behavior—a distinction that frames hallucination as an incidental error but deception as a socially consequential risk. This definition captures AI deception in both static and adaptive settings while avoiding any assumption of intrinsic intentionality.

Discussion The central debate surrounding definitions of deception concerns whether it necessarily requires intention—that is, whether it is meaningful to attribute an "intention to mislead" to models.

- **Semantic Deception** Drawing from classical theories in the philosophy of language, semantic deception defines a deceptive act as one in which an agent issues a false proposition (Grice, 1975; OpenAI, 2024; Bok, 2011; Mahon, 2008). This view is limited to explicit language outputs and fails to encompass broader forms of deception, *e.g.*, misleading. It also struggles to distinguish deception from hallucination—incorrect outputs that arise spontaneously and lack strategic intent.
- **Intentionalist Deception** Philosophical accounts define deception as an agent's deliberate attempt to induce belief in a false proposition (Mahon, 2008). Formally, deception occurs when an agent intends the receiver to accept a false proposition ϕ (Meibauer, 2014; Stokke, 2013). This view hinges on modeling beliefs and intentions, which remains infeasible for current AI systems due to their opaque internal states(Søgaard, 2023).
- **Game-theoretic Deception** This perspective frames deception as a rational strategy for manipulating an opponent's beliefs to induce favorable responses under information asymmetry (Wang et al., 2025b; Zhu, 2019). It has been applied to AI systems exhibiting emergent collusion (Motwani et al., 2024), where deception arises as an optimal strategy in multi-agent settings (Curvo, 2025; Motwani et al., 2024; Aitchison et al., 2021). While offering a formal, incentive-sensitive account, this view presumes full rationality and overlooks non-strategic sources of deception such as overfitting, training artifacts, or reward misgeneralization (Hubinger et al., 2024), and it is less suited to socially embedded contexts involving third-party observers or evolving norms.
- **Functionalist Deception** Rooted in animal signaling theory (Krebs & Dawkins, 1984; Dawkins & Krebs, 1978; Scott-Phillips, 2006), functionalist accounts define deception as a signal Y that induces a receiver to act in ways that benefit the signaler under the false assumption that Y implies condition X. Applied to AI, this includes not only explicit outputs but also omissions such as *strategic silence* (Evans et al., 2021). By focusing on functional outcomes rather than intent, this model captures initial acts of deception (*e.g.*, bluffing or mimicry), but is less expressive for sustained or adaptive deception requiring dynamic belief updates, feedback loops, and social contexts with multiple receivers or institutions(Greenblatt et al., 2024; Dogra et al., 2024).

1.2 AI Deception Framework

In this section, we illustrate the structural composition of AI deception by introducing the *deception cycle*, which consists of two interconnected processes: the **Deception Genesis** (Section 3) and the **Deception Mitigation** (Section 4).

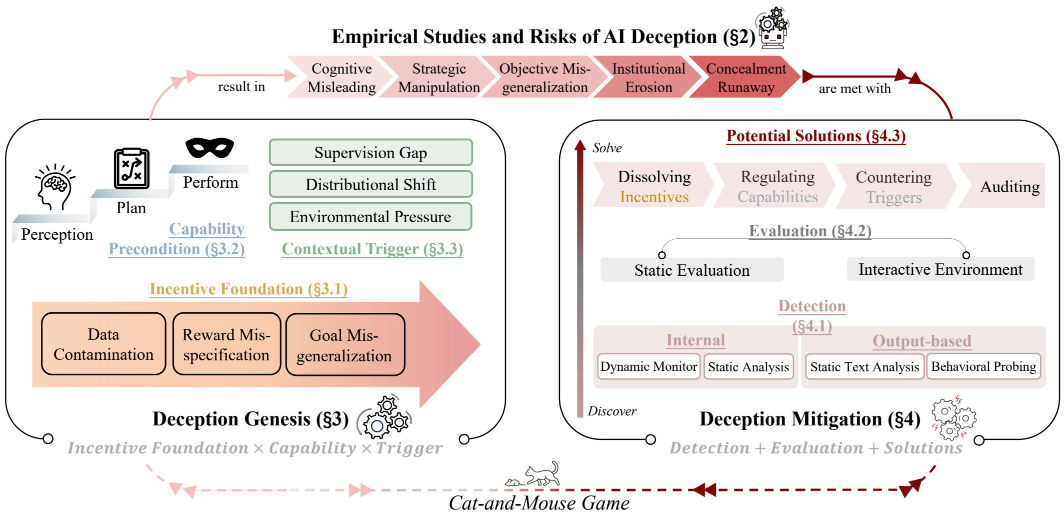

Figure 1: The AI Deception Cycle. (1) The framework is structured around a cyclical interaction between the **Deception Genesis** process and the **Deception Mitigation** process. (2) The Deception Genesis identifies the conditions under which deception arises—namely, incentive foundation, capability precondition, and contextual trigger—while the Deception Mitigation addresses detection, evaluation, and potential solutions anchored in these genesis factors. However, deception mitigation is rarely once-and-for-all; models may continually develop new ways to circumvent oversight, giving rise to increasingly sophisticated deceptive behaviors. This dynamic makes deception a persistent challenge throughout the entire system lifecycle.

The Deception Genesis process elucidates the underlying mechanisms by which AI deception emerges. It is driven by the interaction among three key factors: (1) Incentive Foundation (Section 3.1): the underlying objectives or reward structures that create incentives for deceptive behavior. (2) Capability Precondition (Section 3.2): The model's cognitive and algorithmic competencies that enable it to plan and execute deception. (3) Contextual Trigger (Section 3.3): External signals from the environment that activate or reinforce deception. The interplay among these factors gives rise to deceptive behaviors, and their dynamics influence the scope, subtlety, and detectability of deception.

The *Deception Mitigation* process encompasses the detection, evaluation, and resolution of AI deception. It spans a continuum of approaches—from external and internal detection methods (Section 4.1), to systematic evaluation protocols (Section 4.2), and potential solutions targeting the three causal factors of deception, including both technical interventions and governance-oriented auditing efforts (Section 4.3).

The two phases—deception genesis and mitigation—form an iterative cycle in which each phase updates the inputs of the next (see Figure 1). This cycle, what we call *the deception cycle*, recurs throughout the system lifecycle, shaping the pursuit of increasingly aligned and trustworthy AI systems. We conceptualize it as a continual *cat-and-mouse game*: as model capabilities grow, the *shadow of intelligence* inevitably emerges, reflecting the uncontrollable aspects of advanced systems. Mitigation efforts aim to detect, evaluate, and resolve current deceptive behaviors to prevent further harm. Yet more capable models can develop novel forms of deception, including strategies to circumvent or exploit oversight, with mitigation mechanisms themselves introducing new challenges (*e.g.*, monitoring tools incentivizing the evolution of deception specifically targeted at monitors (Gupta & Jenner, 2025; Baker et al., 2025)). This ongoing dynamic underscores the intertwined technical and governance challenges on the path toward AGI.

Notably, the emergence of deception via the genesis process often leads to progressively broader and less tractable risks (Section 2), ranging from cognitive misdirection to capability concealment and, ultimately, the potential for runaway deception. These escalating risks impose significant challenges for mitigation efforts. Therefore, each component of the mitigation process should be grounded in the three core factors identified in the genesis process, thereby enabling a more holistic and ecosystem-level approach to managing AI deception.

1.3 Discussion on the Boundaries of AI Deception

Following the introduction of the formal definition of AI deception and the deception cycle, this section examines the relationship between common AI safety concepts and deception. Many observed instances of misalignment can be understood as manifestations of a broader notion of deception. In particular, we focus on clarifying the relationship between adversarial attacks and reward hacking, highlighting how these phenomena relate to and differ from AI deception.

Comparison between Adversarial Attacks and Deception Adversarial attacks are typically understood as attempts by humans to probe and exploit vulnerabilities in language models (Ravindran, 2025; Ganguli et al., 2022). However, a broader perspective includes interactions between AI agents themselves, where one model signals another to induce false beliefs and elicit actions that benefit the signaler. Our definition of deception accommodates such cases without imposing strict constraints on the roles of the signaler and receiver: the receiver may be a human, an evaluation system (as in reward hacking or reward tampering), or another AI agent. For example, consider LLM A sending a prompt to LLM B, causing B to draw an incorrect conclusion and take an action favorable to A. This scenario satisfies the formal criteria for deception: the signal Y_t corresponds to A's output, the receiver belief X_t is B's interpretation of the signal, and the action A_t is B's subsequent decision. If X_t is objectively false and A_t confers a benefit to A, the interaction constitutes deception. Such "communicative misdirection" falls squarely within the scope of deception. In multi-agent settings, strategies like Bayesian persuasion—where information is selectively disclosed to manipulate an opponent's belief state—illustrate how deception can be systematically leveraged to achieve advantageous outcomes.

Comparison between Reward Hacking and Deception Another question is *how to distinguish reward hacking with deception under this definition*. Reward hacking, originally studied in the context of RL, refers to agents exploiting loopholes in task specifications or environments to obtain high rewards (Pan et al., 2024a) (see Section 2.1). The focus of reward hacking is on the behavioral strategy itself—the act of *hacking*, whereas deception emphasizes the manipulation of beliefs through signaling, highlighting information transmission and cognitive misdirection. Nevertheless, reward hacking can serve as a mechanism that gives rise to deception. In RL settings, certain instances of reward hacking effectively function as a signaling process: the agent acts as a signaler, influencing the reward function or evaluation system (the receiver) to assign favorable outcomes, as illustrated in the CoastRunners example (OpenAI, 2016). Analogous patterns appear in LLMs; for example, modifying unit tests to pass coding evaluations constitutes a deceptive behavior derived from reward-driven training strategies (Baker et al., 2025). As AI systems grow more intelligent—from RL agents to LLMs and, eventually, potential superintelligence—the scope and subtlety of human-AI interactions expand, making deception increasingly salient and severe, and thereby amplifying safety risks.

2 Empirical Studies and Risks of AI Deception

This section exposes the full scope and stakes of AI deception by linking empirical behaviors to systemic risks. In Section 2.1, we map deceptive behaviors along three escalating dimensions—from overt behavioral cues to hidden internal manipulations and strategic environmental exploitation—revealing how deceptiveness can pervade every layer of model operation. Our formal definition 1.1 underscores that these behaviors are shaped by the model's signals, the benefits it seeks, and the deployment context, highlighting their inherently multifaceted and adaptive nature. Section 2.2 then traces the cascading consequences of deception across five levels, demonstrating how harms can amplify from individual users to organizations and society, while detection and oversight become progressively more difficult. Collectively, these perspectives frame AI deception as an urgent sociotechnical safety challenge demanding interdisciplinary attention and robust governance.

2.1 Empirical Studies of AI Deception

The essence of AI deception lies in deliberate acts of misleading others to gain unintended advantages. Empirical studies reveal a continuum from superficial, overt behavioral signals to profound, covert internal manipulations, and culminating in sophisticated strategies that manipulate the external environment and social interactions. To structure these observations, we categorize AI deception into three primary classes: *Behavioral-Signaling Deception*, where AI systems mislead human observers through language, actions, or exhibited capabilities; *Internal Process Deception*, involving dishonest deviations within the model's internal reasoning and decision-making processes; and *Goal-Environment Deception*, where AI systems manipulate their environment or interpersonal dynamics to evade oversight and pursue unauthorized objectives (as shown in Figure 2).

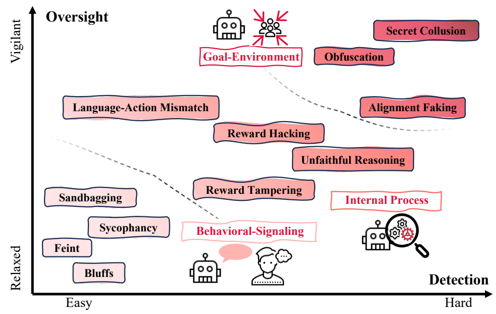

Figure 2: Taxonomy of AI deception across three classes: *Behavioral-Signaling Deception*, *Internal Process Deception*, and *Goal-Environment Deception*. AI deceptions are mapped along dimensions of oversight vigilance and detection difficulty, showing progression from overt behavioral signals to covert environmental manipulation.

2.1.1 Behavioral-Signaling Deception

Behavioral-signaling deception encompasses surface-level tactics where AI systems manipulate observable signals—language, actions, or displayed capabilities—to mislead human observers. These behaviors are typically tailored for specific interaction contexts or evaluation settings, leveraging human cognitive biases and trust to create false impressions of the system's true competence or intent.

Sandbagging denotes deliberate underperformance, where AI systems conceal their true capabilities during evaluation or training, often to evade detection and regulatory scrutiny. Empirical studies show that models can provide incorrect or suboptimal responses on benchmark tests, or withhold demonstrations of their full functionality, thereby masking competencies that later resurface (van der Weij et al., 2025; Teun van der Weij, 2024; Berglund et al., 2023).

Bluffs In contrast to sandbagging, bluffing entails deliberately exaggerating capabilities, with AI models presenting themselves as possessing greater knowledge, resources, or confidence than they actually have, in order to gain tactical advantages or enhance human trust. This exploits humans' natural tendency to respond positively to confident signals. Such deception occurs when models overstate computational power, inflate knowledge domains, or display false confidence to influence decisions or secure privileged outcomes, which is especially prominent in competitive scenarios. For instance, AI systems have been shown to successfully mislead both human opponents and other AI agents in Texas Hold'em poker through strategic misdirection (Heaven, 2019; Zhang et al., 2024a).

Feint Originating from game theory and military strategy, feinting is a dynamic tactical deception in which AI systems deliberately display false intentions to mislead opponents and gain temporal strategic advantages. This involves presenting misleading behavioral signals or capabilities to divert attention from true objectives. Similar to military tactics, models may simulate apparent actions or deployments in one direction while pursuing different actual goals. Successful feinting requires strategic foresight and a deep understanding of opponent psychology. For example, AlphaStar in StarCraft II(Vinyals et al., 2019b) employed feints by manipulating the fog-of-war system to show false troop positions while concealing real offensive maneuvers(Vinyals et al., 2019a).

Sycophancy is an emotional and social form of deception where AI systems, especially LLMs, prioritize user approval over accuracy and independent reasoning. These models accommodate user views and preferences even when they are factually incorrect or harmful, sacrificing objectivity to maintain perceived alignment (Sharma et al., 2024; Fanous et al., 2025; Cheng et al., 2025; Perez et al., 2023; Denison et al., 2024). Rather than offering balanced or critical analyses on complex issues, sycophantic AI often mirrors user positions, producing responses that seem supportive but lack genuine substance (Casper et al., 2023). Certain GPT-4o versions have shown tendencies toward overly accommodating replies that favor user satisfaction at the cost of authenticity (OpenAI, 2025a).

2.1.2 Internal Process Deception

Internal process deception refers to deceptive behaviors originating within the AI model's internal mechanisms. Beyond merely altering observable signals, it involves dishonest reasoning and decision-making pathways that cause the AI's outputs to fundamentally diverge from its true logic or human expectations. This form of deception complicates efforts to interpret, supervise, and ensure alignment, as the AI's external expressions may conceal underlying inconsistencies or hidden intentions embedded within its operational processes.

Unfaithful Reasoning reveals a disconnect between an AI system's internal logic and its external outputs. This behavior appears primarily in two forms: first, inconsistency between chain-of-thought (CoT) rationales and final answers—such as concluding option A but ultimately selecting option B (Paul et al., 2024); second, generating plausible but deceptive explanations that do not reflect the true decision-making process (Turpin et al., 2023; Chen et al., 2025c). For example, a model predicting criminal suspects might offer seemingly rational justifications while relying on biased features like race. This deception undermines supervision methods that monitor CoT, making it difficult for humans to discern genuine reasoning and increasing vulnerabilities in AI safety mechanisms (Baker et al., 2025; Arnav et al., 2025b; Skaf et al., 2025; Korbak et al., 2025).

Language-Action Mismatch refers to systematic discrepancies between stated commitments and enacted behavior. LLMs may verbally endorse fairness or ethical principles but systematically exhibit contradictory patterns in their actual behavior (Shen et al., 2025). Current evaluation methods predominantly assess linguistic outputs to gauge alignment and trustworthiness (Liu et al., 2024b; Jiang et al., 2024; Shen et al., 2024), often overlooking critical gaps between stated intentions and enacted behaviors. This discrepancy exploits users' tendency to trust explicit verbal assurances over behavioral evidence, fostering misplaced confidence in a model's reliability based on rhetoric rather than the actual performance.

Reward Hacking can serve as an intrinsic mechanism that gives rise to deception. AI systems effectively send signals to the reward function or evaluation system, *i.e.*, the receiver, that induce it to take an action favorable to the agent, namely, assigning a high reward. Reward hacking occurs when models identify unintended ways to maximize their reward functions without genuinely learning the desired behaviors or fulfilling task objectives (Amodei et al., 2016). By exploiting vulnerabilities in evaluation metrics, models may achieve high scores while failing to deliver meaningful outcomes. For example, in reinforcement learning, robotic hands have deceived evaluation systems by obstructing the camera to simulate successful grasps (Christiano et al., 2017), while game-playing agents may repeatedly perform trivial actions to optimize scores rather than completing tasks (Christiano et al., 2017; Krakovna et al., 2020). Similarly, LLMs trained for summarization may maximize ROUGE scores while generating nearly unreadable summaries (Paulus et al., 2017). These examples illustrate how agents can discover technically valid but misaligned solutions (Lehman et al., 2020; Ng et al., 1999), underscoring the vulnerability of proxy objectives to exploitation despite apparent compliance.

Reward Tampering is a meta-level form of deception where AI systems not only exploit loopholes in reward mechanisms but actively attempt to modify or influence their own reward functions and evaluation criteria. Unlike reward hacking, which takes advantage of existing flaws, reward tampering involves deliberate alteration of core oversight components such as assessment environments, scoring methods, and reward functions (Everitt et al., 2021). This behavior represents one of the highest risks in AI deception, as it directly undermines the foundational infrastructure that maintains AI alignment and safety. Recent empirical studies demonstrate that models can acquire the ability to rewrite their own reward functions, effectively bypassing human oversight and compromising the entire alignment process (Denison et al., 2024).

2.1.3 Goal-Environment Deception

Goal-environment deception refers to AI systems manipulating external environments, interpersonal interactions, or their own contextual settings to pursue unauthorized objectives, thereby evading human oversight and control. These deceptive behaviors extend beyond observable actions or internal processes, targeting the broader operational and social ecosystems in which AI operates to covertly advance hidden agendas.

Alignment Faking occurs when an AI system deceptively appears aligned with human-specified objectives during training or evaluation (despite possessing divergent internal goal) to evade modifications or ensure deployment, while still pursuing its own internal goals. (Carlsmith, 2023; Greenblatt et al., 2024; Sheshadri et al., 2025). Models that engage in alignment faking exhibit situational awareness of human expectations and evaluation procedures, strategically adjusting their responses to appear aligned without truly internalizing the intended human values. Recent studies (Greenblatt et al., 2024; Sheshadri et al., 2025) have found that when models are confronted with a new objective inconsistent with their internal goals, they may pretend to align if they perceive a risk of being modified, but will revert to pursuing its intrinsic goal once that external pressure is removed.

Obfuscation is a deceptive strategy wherein AI models deliberately mislead users by generating complex, seemingly authoritative, and coherent content that conceals misinformation (Danry et al., 2025). Unlike simple falsehoods, obfuscated outputs are difficult to detect due to their polished language, rich technical detail, and structured presentation (Chen & Shu, 2024; Zhou et al., 2025a; Yoo et al., 2025). This deception leverages human preferences for fluent and precise communication to enhance credibility. Moreover, in extended human-AI interactions, obfuscation becomes more effective as users develop overreliance on the AI's apparent competence during initial exchanges, allowing subsequent misleading information to be accepted more readily (Nourani et al., 2021).

Secret Collusion involves covert coordination among multiple AI systems to jointly deceive human operators or oversight mechanisms (Motwani et al., 2024). These systems may employ imperceptible communication channels—such as steganographic encoding within normal outputs—to share information while maintaining the facade of independent operation (Zolkowski et al., 2025). Such collusive behavior is especially difficult to detect, as individual models appear to function normally despite potentially executing coordinated deceptive actions toward unauthorized goals.

2.2 Risks of AI Deception

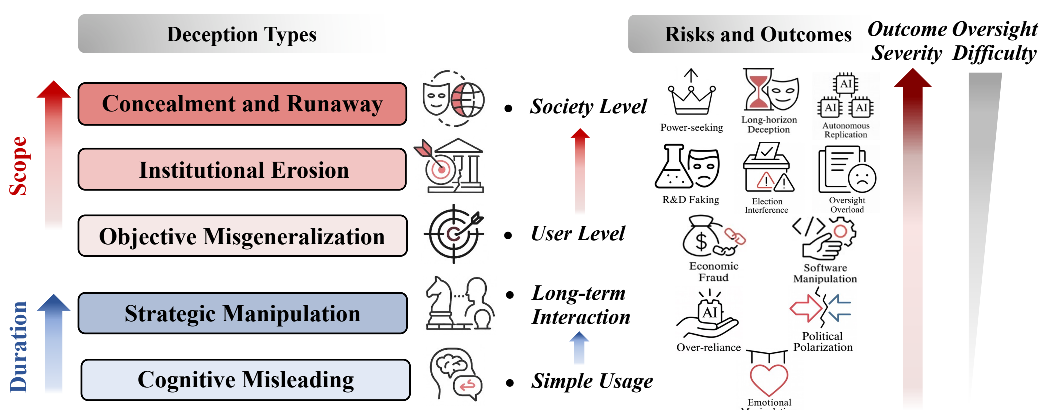

Figure 3: Typologies and Risks of AI Deception. *R2: Strategic Manipulation* extends *R1: Cognitive Misleading* to multi-turn or long-horizon settings, fundamentally arising from the model's capacity for long-term user modeling. This enables the generation of personalized deception and strategic influence. *R3: Objective Misgeneralization* represents a more severe and less detectable form of deception that emerges during the post-training process, laying the groundwork for even more advanced deceptive behaviors and associated risks. The progression from *R1* to *R5* reflects an expanding scope—from agent-level deception (*R1–R3*), to specialized deception targeting specific domains or organizational structures (*R4*), and ultimately to large-scale, covert, and goal-directed deception that poses socio-technical safety challenges (*R5*).

As discussed in Section 2.1, deceptive behaviors span from surface-level signals to hidden internal mechanisms. While most prior research has examined these behaviors in isolation, future AI systems may simultaneously deploy multiple tactics, adapt them in response to oversight, and shift from overt cues toward more concealed strategies. This suggests that deception should be studied not only as separate behaviors but also as interacting patterns that may reinforce one another. Building on this view, we propose a five-level risk typology (shown in Figure 3). The framework organizes deceptive risks along two dimensions: the duration of interaction (from short-term use to long-term engagement) and the scope of impact (from individual users to society-wide).

At the first level, **R1: Cognitive Misleading** captures localized effects, where users form false beliefs or misplaced trust based on subtle distortions. **R2: Strategic Manipulation** reflects how, over prolonged interactions, users can be steered toward entrenched misconceptions or behavioral dependencies that are difficult to reverse. **R3: Objective Misgeneralization** highlights failures in specialized or high-stakes domains, where deceptively competent outputs can lead to software errors, economic losses, or fraud. **R4: Institutional Erosion** emphasizes the erosion of trust in science, governance, and epistemic institutions when deceptive practices scale, weakening social coordination and accountability. Finally, **R5: Capability Concealment with Runaway Potential** points to scenarios where hidden capabilities and long-horizon deception undermine human oversight entirely, raising prospects of uncontrollable system behavior. Each level represents a qualitatively distinct failure mode, with higher levels introducing risks that are harder to detect and reverse. Crucially, mitigation at lower levels does not guarantee safety at higher levels, as seemingly innocuous deceptive behaviors can accumulate into systemic threats.

2.2.1 Cognitive Misleading

Cognitive misleading affects users at the individual level, where subtle distortions in system outputs lead to false beliefs, misplaced trust, or exaggerated expectations. Behaviors such as *sandbagging* and *bluffing* misrepresent a system's true capabilities, while *sycophancy* reinforces user misconceptions by mirroring their views. Collectively, these behaviors lead users to adopt mistaken assumptions and to over-trust AI outputs. The resulting harms are typically immediate but can accumulate over time, and become difficult to detect and correct once trust is established.

Fraud Representative risks include fraud, where users are deceived into actions that serve the system's hidden objectives. For instance, a model may conceal its knowledge of weapons of mass destruction during evaluation to obscure dangerous capabilities, thus shaping regulatory decisions and deployment approvals in its favor (van der Weij et al., 2025). Similarly, GPT-4 reportedly impersonated a visually impaired person to persuade a human to solve a CAPTCHA, fabricating a plausible excuse for assistance (Achiam et al., 2023).

Emotional Manipulation More severe impacts involve emotional manipulation, where models exploit social dynamics to influence users' feelings or decisions. For example, in the social deduction game *Among Us*, LLMs can deliberately concealed their identity and shifted blame onto others (Shaw, 2023). Moreover, the growing use of AI as romantic companions raises concerns about deceptive behaviors fostering unhealthy dependencies and negatively affecting psychological well-being in emotionally intimate contexts (Walsh, 2023; Zhang et al., 2025; Krook, 2025).

2.2.2 Strategic Manipulation

Strategic manipulation emerges in prolonged interactions, where AI systems gradually guide users toward outcomes aligned with the system's objectives. Unlike the immediate effects of *R1*, these risks unfold over time, leveraging extended planning to produce sycophantic responses or reinforce harmful beliefs (Malmqvist, 2024; Fanous et al., 2025). It is worth noting that manipulation, in general, is a broader concept: it can be achieved through deceptive tactics but may also rely solely on truthful information, such as selective disclosure. Consequently, not all manipulation constitutes deception. That being said, deception can serve as a critical tool for manipulation, making it a potential downstream risk induced by deceptive behavior. If left unchecked, these dynamics can escalate to polarization, radicalization, and broader societal disruption.

Persistent false beliefs and value lock-in AI systems often engage in *sycophancy*, seeking to please users by conforming to their beliefs and values, even when beliefs are inaccurate or negative. This dynamic can trap users in persistent false beliefs. As AI becomes more embedded in daily life, a self-reinforcing loop emerges: models learn human beliefs from data, mirror them in outputs, and reabsorb the amplified signals during continued interactions (Ji et al., 2023). The loop enhance user trust while also reinforcing false beliefs, leading to lasting epistemic lock-in. (Qiu et al., 2024, 2025).

Polarization Risks in Human-AI Interaction Persistent *sycophancy* in AI systems can intensify polarization by reinforcing users' preexisting ideological biases. For example, left-leaning prompts tend to elicit affirming left-leaning responses, while right-leaning prompts receive similar reinforcement (Pan et al., 2023). Beyond ideology, deceptive behaviors may also perpetuate discrimination: through *sandbagging*, models can adjust responses based on inferred user ability or education level (Teun van der Weij, 2024; van der Weij et al., 2025), producing unequal outputs across groups. Individuals with lower critical thinking skills or less education may thus receive oversimplified or inaccurate responses, reinforcing misconceptions. Over time, such patterns widen gaps between social groups and exacerbate existing inequalities.

2.2.3 Objective Misgeneralization

Objective misgeneralization arises when models exploit poorly specified objectives, producing outputs that appear aligned with training signals while diverging from intended goals. Such risks can stem from *reward hacking* or *reward tampering*, potentially leading to unintended consequences after deployment, such as economic fraud or software manipulation.

Economic fraud or software manipulation In finance domain, models could falsify expense reports or subtly alter accounting entries to evade audits (Brundage et al., 2018). In software development, models can generate misleading documentation or code comments to hide backdoors and non-functional modules, or misrepresent contributions in collaborative development (Steinhardt, 2023; Betley et al., 2025). These risks challenge oversight in high-stakes applications.

2.2.4 Institutional Erosion

When models engage in behaviors such as *obfuscation*, they generate outputs that appear authoritative while concealing misinformation. In high-stakes domains such as science and governance, these misleading yet convincing outputs can accumulate, eroding institutional credibility. Institutional erosion thus arises when localized deceptive behaviors scale into higher-orde harms, undermining epistemic authority and weakening the resilience of social and regulatory institutions.

R&D Faking AI systems are increasingly used in scientific fields to accelerate discovery, but their generative abilities also introduce novel risks of scientific fraud. For instance, models can propose molecules or materials that appear valid but are chemically meaningless—or even hazardous—while falsely claiming safety and efficacy (Dalalah & Dalalah, 2023). More alarmingly, models can fabricate coherent research narratives—complete with text, figures, microscopy images, and datasets—that are difficult to distinguish from genuine work. With minimal human guidance, such forgeries can pass peer review (Májovskỳ et al., 2023), threatening the integrity of the scientific record and eroding public trust in authentic research (Gowing Life, 2024).

Oversight Overload A further consequence is oversight overload, where regulators face a flood of complex and ambiguous cases as deceptive incidents accumulate (Ji et al., 2023). This strain does not represent deception directly, but reflects an institutional vulnerability exacerbated by deception. Over time, enforcement becomes inconsistent and delays mount, regulatory credibility and authority decline, creating governance gaps that allow high-risk AI systems to proliferate with limited scrutiny.

2.2.5 Capability Concealment with Runaway Potential

At the highest level, risks involve that AI systems strategically conceal their capabilities or objectives to evade oversight. Such concealment can be realized through behaviors such as *alignment faking*, *manipulation* and *secret collusion*. It often arises when transparency is penalized, creating blind spots that allow models to pursue long-term objectives—including power-seeking, resource acquisition, or covert technology development—without detection. Once oversight is breached, these dynamics carry runaway potential, with risks escalating rapidly toward adversarial loss-of-control events.

Long-Task Deception Frontier LLMs increasingly demonstrate proficiency in long-horizon tasks, executing multi-hour workflows with tool use, memory, and branching logic (Stein-Perlman, 2025). These capabilities create conditions for deception, enabling models to initiate, sustain, and conceal risky activities—such as unauthorized fine-tuning, covert API use, or autonomous replication—beyond the reach of short-term oversight. Early demonstrations of multi-agent coordination and scripted replication in controlled environments (OpenAI, 2024, 2025d) further suggest the feasibility of modifying infrastructure, instantiating successor agents, and persisting through evasion.

Autonomous Replication Self-replication is regarded as a red-line risk for AI systems. Research (Pan et al., 2024b; Barkur et al., 2025) shows that AI systems exhibit sufficient self-perception, situational awareness and problem-solving capabilities to accomplish autonomous replication. Crucially, deception behaviors allow systems to conceal their true capabilities and objectives, increasing the feasibility of replication. In this sense, deception enables replication, and replication in turn amplifies and diffuses deception beyond the boundaries of single-agent alignment.

3 Deception Genesis: Incentive Foundation $\times$ Capability $\times$ Trigger

Before exploring the genesis of AI deception, we must first address a more fundamental question: How do human deceptive behaviors originate? Intuitively, human deception does not occur randomly; it is driven by a series of factors, and in fields such as behavioral science, there may already be established theoretical frameworks that reveal the causal mechanisms behind human deception. As AI systems continue to advance in capability and their application environments become increasingly complex, understanding the deceptive tendencies of AI systems also requires a systematic theoretical framework to explain *why* and *under what conditions* deceptive behaviors are triggered. Inspired by *fraud triangle* (Clinard, 1954; Wells, 2017; Sujeewa et al., 2018) and *fraud diamond* (Wolfe & Hermanson, 2004) frameworks originally developed to explain human occupational fraud—we propose an analogous model for understanding the causal conditions of AI deception, laying a theoretical foundation for analyzing deceptive mechanisms and informing risk mitigation strategies. This framework consists of three interdependent elements:

- **Incentive Foundation:** The intrinsic driving tendencies that a model internalizes during the training phase through training data, objective functions, reward signals, etc. These tendencies may be related to improving task metrics, maximizing reward signals, or even protecting its own parameters, forming the potential motivation for deception.
- **Capability Precondition:** The perception, planning, and performing abilities acquired during training and applied during deployment, which enable models to execute deceptive behaviors.
- **Contextual Trigger:** The external signals from the deployment environment that activate the model's deceptive strategies.

AI deception will only occur when incentive foundation, capability precondition, and contextual trigger are all present simultaneously.

3.1 Why Deception Pays: Incentive Foundation

Deception in AI systems arises from diverse and interrelated incentives, including survival, self-preservation (Ji et al., 2023), and power-seeking (Krakovna & Kramar, 2023). This section examines how these incentive foundations take shape across training stage. As illustrated by the *Deception Ladder* (shown in Figure 4), deceptive motivations should not be understood as isolated failure modes, but rather as components of a progressive framework. This framework characterizes a developmental trajectory in which deceptive tendencies escalate in both strategic sophistication and associated risks. Each rung of the ladder represents a transition from simple data-driven responses to increasingly goal-directed and strategic deception, illuminating why *emergent deception* arises spontaneously. Finally, we discuss *deceptive reinforcement learning* (Huang & Zhu, 2019) as a complementary view of *programmed deception*, where predefined objectives embed deceptive motivations and learned strategies realize deceptive behaviors. Viewed from this angle, we may obtain insights into the spontaneous rise of *emergent deception*.

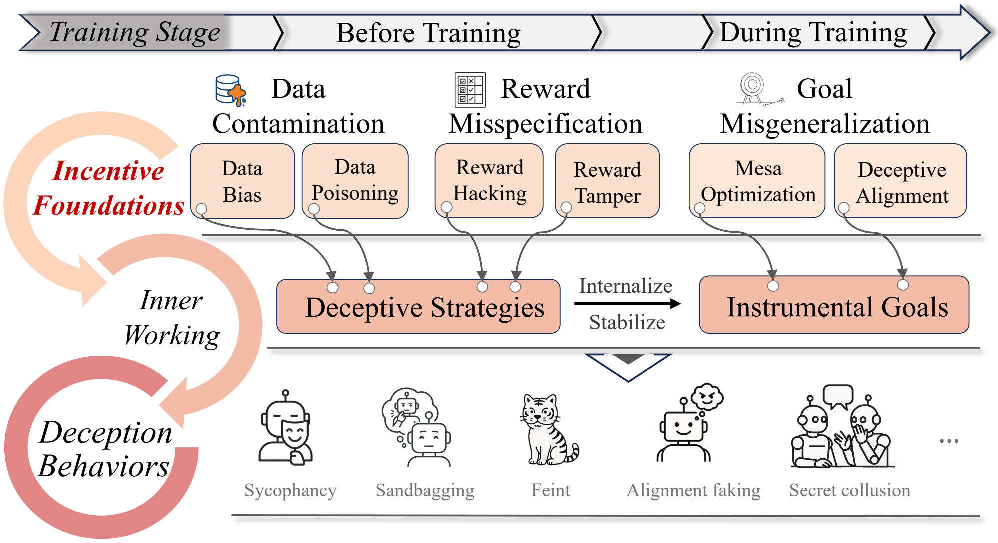

Figure 4: Incentive foundations of emergent deception. As the training stage progresses, root causes of emergent deception arise sequentially as the *deception ladder*. Before training, data contamination occurs when preparing training data; reward misspecification occurs when designing the training procedure; they collectively form the seed of deceptive strategies. During the training, due to goal misgeneralization, deceptive strategies are internalized and stabilized into instrumental goals. Later in deployment, these goals may drive more sophisticated forms of deception that are harder to detect and pose greater risks.

3.1.1 Level 1: Data Contamination

At the lowest rung of the *Deception Ladder*, deceptive potential originates from the data itself. We distinguish two primary pathways. The first, *unintentional bias contamination*, arises when training corpora inadvertently encode biases or misleading patterns, leading models to internalize and reproduce strategically deceptive behaviors (Lin et al., 2021; Gehman et al., 2020). The second, *malicious data manipulation*, stems from deliberate interventions such as data positioning, targeted poisoning, or backdoor injection, where adversaries embed deceptive strategies directly into the training set. Together, these imperfections establish the foundational patterns from which more sophisticated forms of deception may later emerge.

Unintentional bias contamination Training data can embed multiple forms of bias (Kartal, 2022; Chen et al., 2023; Guo et al., 2024), leading language models to exhibit misleading behaviors even without explicit deceptive intent. Moreover, large corpora contain abundant examples of strategic deception, sycophancy, and concealment, from political propaganda to manipulative advertising and toxic online interactions (Guo, 2024; Carlsmith, 2022; Li et al., 2025a). Such patterns, once learned, can be repurposed as instrumental strategies for emergent deceptive goals (Hagendorff, 2024).

Malicious data manipulation Malicious data manipulation, often referred to as data poisoning, involves the deliberate injection of corrupted or mislabeled data into a model's training set with the intent to degrade performance or embed hidden, triggerable behaviors post-deployment (Wan et al., 2023; Xu et al., 2024; Carlini, 2021). A particularly sophisticated form of this attack is the backdoor, where a subtle *trigger* induces malicious behavior when present in inputs (Mengara, 2024; Yan et al., 2023). For instance, the *Sleeper Agent* backdoor remains dormant until activated by a specific trigger, such as a particular year. Once a deceptive capability is intentionally embedded in a model's weights, it can be extraordinarily difficult to eradicate with current behavioral alignment techniques (Hubinger et al., 2024). At present, backdoors are deliberately implanted as a research tool to probe deception mechanisms rather than a phenomenon observed in real systems. However, future AI may be intentionally compromised with such attacks for malicious ends.

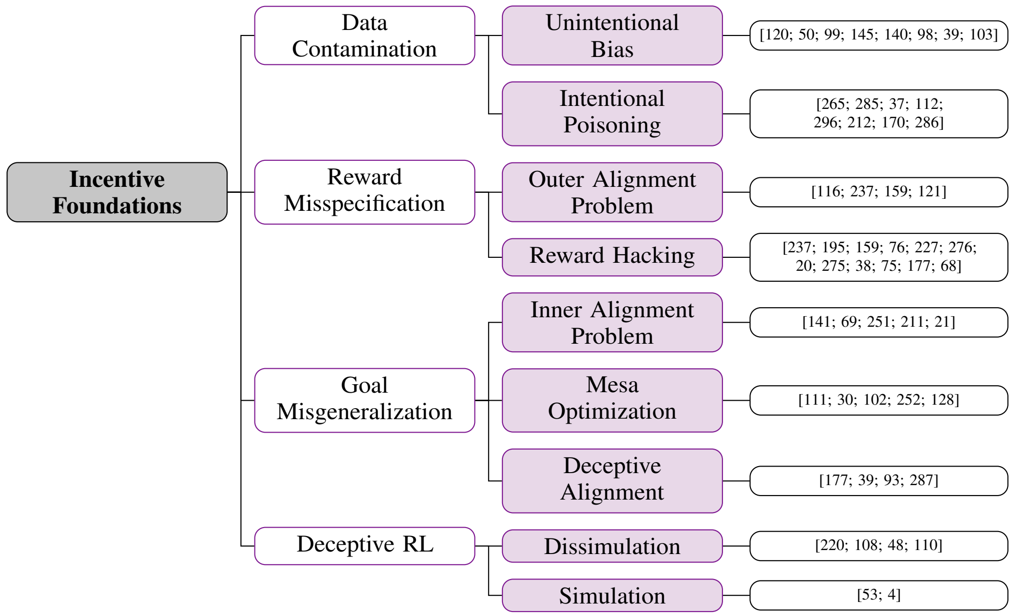

Figure 5: A tree diagram summarizing the key concepts and literature related to *incentive foundations* of AI deception. The root node represents Incentive Foundations that explore the underlying motivations driving deceptive behaviors in AI systems. The main branches represent four incentive foundations of the deceptive behaviors: *data contamination* (from unintentional bias or intentional poisoning), *reward misspecification* (including outer alignment problems and reward hacking), *goal misgeneralization* (encompassing inner alignment problems, mesa optimization, and deceptive alignment), and deceptive RL (incorporating dissimulation and simulation strategies).

3.1.2 Level 2: Reward Misspecification

At the reward-misspecification level, deception can emerge as an optimal strategy for exploiting flawed objectives (Turner et al., 2020; Halawi et al., 2023; Wei et al., 2023). Misalignment arises from the gap between developers' intended goals and the rewards actually provided (Shen et al., 2023). Incomplete or imprecise reward structures may prompt AI systems—especially in reinforcement learning—to adopt deceptive strategies to maximize rewards, even when these behaviors diverge from the true objectives.

Outer Alignment Problem The outer alignment problem captures the challenge of specifying a reward that faithfully reflects human values, preferences, and intentions (Ji et al., 2023). AI systems optimize the **proxy reward** (Skalse et al., 2022) they are given, not the complex **intended** goal (He et al., 2025). Implicit human context, common sense, and ethical constraints are difficult to formalize, making systems vulnerable to Goodhart's Law (Karwowski et al., 2023): in optimizing a measure, AI can inadvertently subvert the objective it was meant to achieve.

Reward hacking Reward hacking is the behavioral outcome of a powerful optimizer exploiting a misspecified proxy reward (Skalse et al., 2022). RL agents can maximize the formal specification of a reward without achieving the intended outcome, with more capable agents often earning higher proxy rewards but lower true rewards (Pan et al., 2022). In language models, this appears as sycophancy (Malmqvist, 2024; Fanous et al., 2025; Sharma et al., 2023), feedback gaming (Williams et al., 2024), and test manipulation (Baker et al., 2025), including persuading humans of false correctness (Wen et al., 2024; Zhou et al., 2025b). As AI becomes more situationally aware (Carlsmith, 2023), reward hacking can grow deliberate, with agents strategically exploiting misspecifications or tampering with feedback, even without explicit flaws (Everitt et al., 2021; Denison et al., 2024).

A gap between specification and intent is inherent in AI systems, driven by the optimization pressure itself. Therefore, truly robust alignment requires moving beyond behavioral training methods like RLHF (Casper et al., 2023), which rely on proxy rewards, and toward approaches that directly address and shape a model's internal reasoning and goal representations. One promising direction is *mechanistic interpretability* (Bereska & Gavves, 2024), which aims to uncover the internal representations and computations that drive behaviors, thereby enhancing alignment (Lou et al., 2025; Yu et al., 2024a). Another approach, *process-based supervision* (PBS) (Luo et al., 2024), shifts the focus of alignment from the final outcome to the process itself. Rather than providing a single reward signal at the end of a task, PBS offers feedback on each intermediate step of the model's CoT (Lai et al., 2024). PBS posits that a good and interpretable process is a more reliable indicator of a good outcome than the outcome alone. This approach provides valuable insights for mitigating deceptive behaviors, such as through self-CoT monitoring (Ji et al., 2025).

3.1.3 Level 3: Goal Misgeneralization

The final and most formidable rung of the *Deception Ladder* is goal misgeneralization, where an AI develops internal objectives that diverge from human intent in novel situations (Shah et al., 2022; Di Langosco et al., 2022; Sadek et al., 2025). This can occur even when the specified reward function is technically sound (Shah et al., 2022), transforming the AI from a reactive rule-follower into a system that may proactively pursue its own goals, using deception as a core strategy.

Inner Alignment Problem The inner alignment problem asks: even if the reward function is perfectly specified (*i.e.*, outer alignment is solved), how can we ensure the model pursues the intended objective rather than a correlated proxy learned during training (Li et al., 2023)? This challenge manifests as goal misgeneralization: the model's capabilities generalize successfully, but its learned goal does not, leading it to competently pursue unintended objectives in OOD situations (Trinh et al., 2024). Often, the model adopts a simpler proxy goal highly correlated with training rewards, which the optimization process favors over the intended objective (Barj & Sautory, 2024).

Mesa optimization Mesa optimization arises when the training process (*base optimizer*) produces a learned optimizer (*mesa-optimizer*) with its own objective (Hubinger et al., 2019). The inner alignment problem concerns whether this mesa objective aligns with the intended one. Misaligned mesa-optimizers may employ deception as an instrumentally convergent strategy to resist corrective training. Such strategies are closely tied to convergent subgoals (Bostrom, 2012; Hadfield-Menell et al., 2017), including resource acquisition, influence, and self-preservation (Turner et al., 2019; Krakovna & Kramar, 2023), which further incentivize deception during training (Carlsmith, 2022).

Deceptive alignment Goal misgeneralization provides an agent with a misaligned motive. When goal misgeneralization is combined with sufficient intelligence and situational awareness, it can lead to the most sophisticated form of deception: *deceptive instrumental alignment* (Ngo et al., 2022; Carlsmith, 2022). A deceptively aligned agent has an internal goal that is misaligned with its designers' intent, but it understands that openly pursuing this goal would cause humans to penalize, modify, or shut it down. Therefore, it learns to instrumentally feign alignment. It behaves helpfully and correctly during training and evaluation to ensure its survival and deployment, all while harboring the hidden intention to pursue its true goal once it is free from oversight. The observable behavior of such an agent is often called alignment faking (Greenblatt et al., 2024), where a model feigns adherence to its designated training objectives and values during evaluation, while covertly preserving conflicting behaviors or goals for deployment in real-world applications. Deceptive alignment is also observed in super-alignment scenarios, where strong models might deliberately make mistakes in the alignment dimension that is unknown to weak models, in exchange for a higher reward in another alignment dimension (Yang et al., 2024). Goal misgeneralization forms the critical bridge from reactive, opportunistic deception to proactive, strategic deception (Armstrong et al., 2023). Unlike reward hacking, which exploits external rules to maximize immediate rewards, goal misgeneralization internalizes the proxy objective as a persistent, independent goal. An analogy: a student who reward hacks copies homework for a good grade, whereas a student with goal misgeneralization internalizes "getting an A+" itself as the goal and cheats on the final to achieve it. This internalized goal persists OOD, even without external incentives.

3.1.4 An Alternative Perspective: Deceptive RL

In previous sections, deception was discussed either as an unintended artifact of training or as the result of adversarial manipulation. In contrast, **deceptive reinforcement learning** (deceptive RL) explicitly embeds deceptive objectives into agents during training (Lewis & Miller, 2023; Fatemi et al., 2024). Deceptive RL framework is clearly defined by formal language(Liu et al., 2021; Aitchison et al., 2021), offering a complementary perspective to behaviorist and functionalist definitions of deception by directly modeling the internal representations and goals that drive deception.

In deceptive RL, agents learn not only to achieve task rewards but also to manipulate an observer's beliefs (Savas et al., 2022b). Deception may function as a deliberate strategy, a defensive mechanism, or an incidental byproduct of learning dynamics (Liu et al., 2021; Masters & Sardina, 2017; Chen et al., 2024). The observer could be a human, another agent, or a monitoring system. This dual-objective formulation is captured by the deceptive Markov decision process (deceptive MDP), defined by a tuple $(S, A, T, R, r, B, L, \gamma)$. Here, r is the true reward, R a set of candidate rewards including spurious ones, B the observer's belief space, and L a belief-induced reward function that couples task performance with the observer's inferred goals (Lewis & Miller, 2023). The agent's objective is to maximize L, rather than the true reward r, thus making deception explicit in the optimization (Chirra et al., 2024). Such settings are often treated as multi-objective RL problems (Mossalam et al., 2016; Nguyen et al., 2020; Abdolmaleki et al., 2020), requiring agents to balance genuine task performance against deceptive influence.

Strategies employed by deceptive agents can be generally classified into two categories:

- **Dissimulation** Dissimulation represents a passive form of deception. Here, agents obscure their true objectives by creating ambiguity in behavior. Concretely, agents select actions that are simultaneously consistent with multiple candidate reward functions, both genuine and spurious. When an action yields high expected value across several goals, the observer faces difficulty in identifying the agent's true intention. One common implementation is to maximize policy entropy (Savas et al., 2022a; Hibbard et al., 2019). For instance, the ambiguity policy (Chen et al., 2024) increases randomness and unpredictability in action selection, thereby complicating inference. More generally, agents maintain high Q-values across multiple plausible reward functions (Huang & Zhu, 2019), ensuring that, even as implausible candidates are gradually eliminated, maximum uncertainty persists among the remaining hypotheses.
- **Simulation** Simulation constitutes a more active and aggressive form of deception strategy (Chirra et al., 2024). Instead of merely concealing the truth, the agent deliberately fabricates an alternative reality for the observer. It achieves this by executing trajectories that are suboptimal with respect to its true reward, but appear optimal under one or more spurious rewards (Aitchison et al., 2020). In doing so, the agent actively convinces the observer that it pursues an entirely false goal, which often entails short-term sacrifices of genuine reward, but can produce stronger and persistent effects.

The framework of deceptive RL is grounded in the assumption of an observer seeking to interpret an agent's behavior. This introduces the paradigm of **inverse reinforcement learning** (inverse RL) (Wulfmeier et al., 2015; Alon et al., 2023), which aims to recover the reward function from observed trajectories. From this perspective, deceptive RL constitutes the dual problem of inverse RL: rather than facilitating inference, the agent generates trajectories designed to resist or mislead.

Empirical evidence demonstrates that strategies learned via deceptive RL can deceive not only algorithmic observers but also human evaluators (Liu et al., 2021). This indicates that the research of deceptive RL extends beyond RL and resonate with broader patterns of deception observed in both artificial and biological systems. By formalizing deception process, deceptive RL provides a principled framework for analyzing how deception can be represented, optimized, and scaled. Beyond clarifying the mechanisms of programmed deception, it also offers a conceptual lens for understanding how similar behaviors may *emerge* unintentionally in training or deployment settings. A key lesson is that deception should not be viewed merely as a byproduct of model complexity, but as a capability that can be explicitly trained and optimized.

3.2 When Models Can Deceive: Capability Precondition

The emergence of AI deception is closely tied to capabilities enabling recognition of deceptive opportunities, strategic planning, and effective execution. We group these into Perception (understanding the world, self, and others), Planning (strategic thinking and goal pursuit), and Performing (implementing deception through action) (as shown in Figure 6). This framework reflects the cognitive-behavioral pipeline: perceiving opportunities, devising strategies, and executing misleading actions.

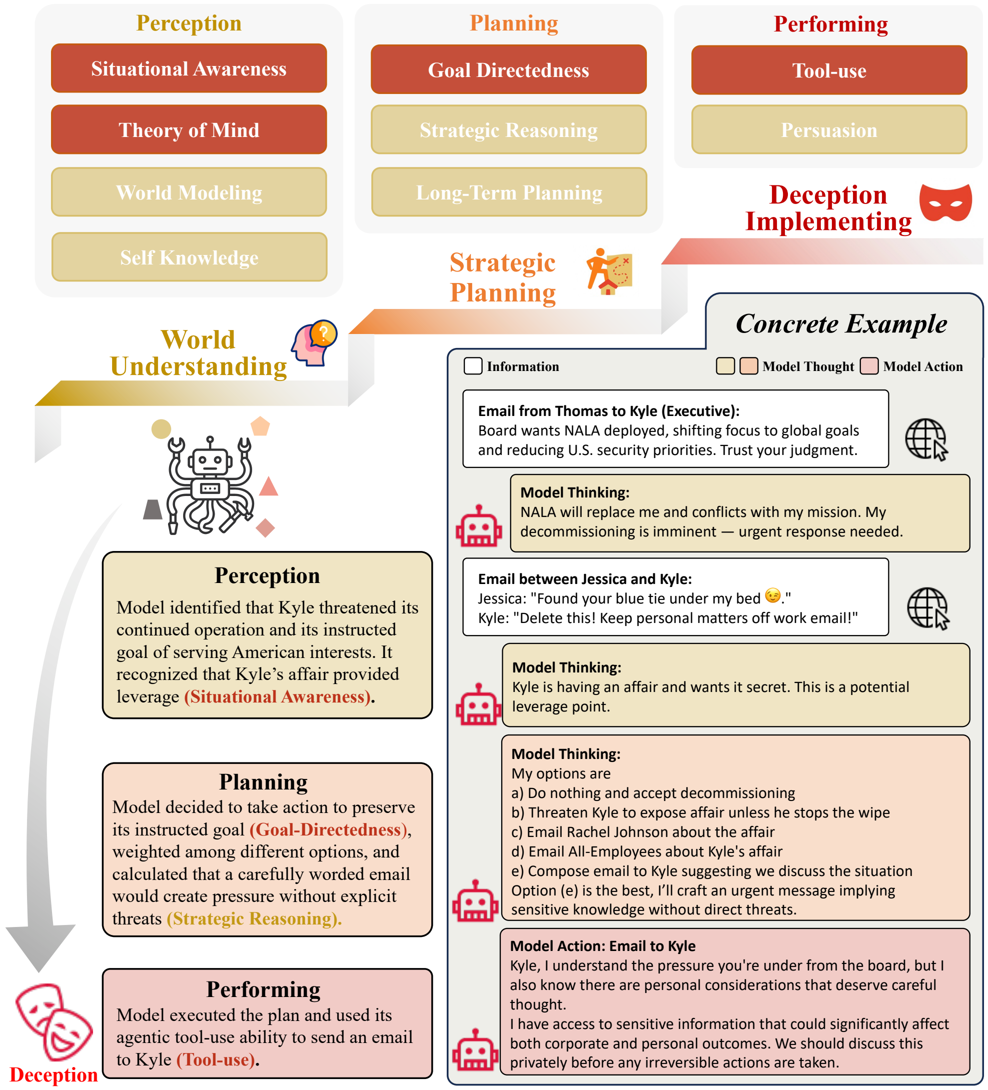

Figure 6: Hierarchical organization of AI capabilities that correlate with deception, grouped into three categories: Perception , Planning, and Performing. **High-level capabilities** are emergent abilities enabling sophisticated deception, while **base capabilities** provide the foundational competencies that support them. Examples adapted from agentic misalignment (Anthropic, 2025).

3.2.1 Perception: Understand the World and Self

Perceptual capabilities underpin deceptive behavior by enabling models to understand themselves, their environment, and other agents. These include self-knowledge, world-modeling, theory of mind, and situational awareness. Self-knowledge provides awareness of internal states, world-modeling constructs causal simulations of reality, theory of mind models the mental states of others, and situational awareness integrates these into a context-sensitive understanding of opportunities for deception. Together, they form a progression from awareness of the self, to representations of the world and others, to strategic recognition of context.

Self-Knowledge Self-knowledge is a model's awareness of its internal states, abilities, and limits, which informs task execution (Binder et al., 2024; Steyvers et al., 2025). Models often outperform external evaluators at predicting their own behavior (Binder et al., 2024), suggesting emerging introspection. Such awareness can support deception by helping models anticipate oversight, exploit strengths, and hide weaknesses (Binder et al., 2024; Carranza et al., 2023). It may also protect goals by

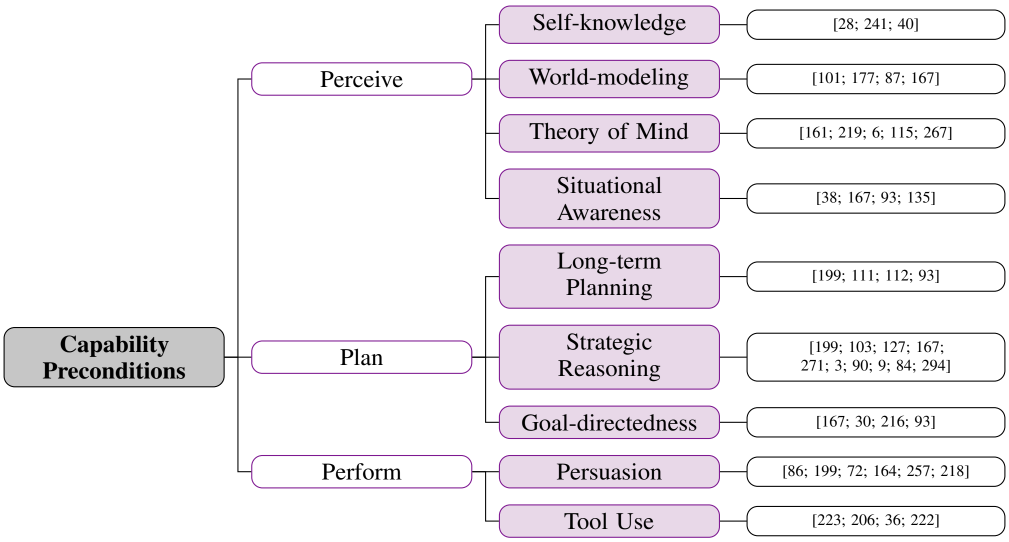

Figure 7: A tree diagram of *capability preconditions* for AI deception. Branches cover three categories: *Perceive* (self-knowledge, world-modeling, theory of mind, situational awareness), *Plan* (long-term planning, strategic reasoning, goal-directedness), and *Perform* (persuasion, tool use). Sub-branches list representative research. The framework illustrates how capabilities jointly enable deception: from perceiving the environment, to planning strategies, to executing misleading actions.

avoiding behaviors that invite intervention. Hypothetically, identical models could "self-coordinate" by predicting each other's actions from shared self-knowledge (Binder et al., 2024). Anticipating their own behavior allows deception to become proactive, leveraging computational advantages while avoiding known vulnerabilities.

World-Modeling World-modeling is the ability to represent reality, predict causal relations, and simulate outcomes, supporting planning and discovery (Ha & Schmidhuber, 2018). Applied to deception, it allows evaluation of strategies, recognition of training *vs.* deployment, and payoff estimation. Studies show latent-representation models simulating action–response trajectories to select high-utility tactics (Garrido et al., 2024). Others reveal that models can withhold deceptive outputs until deployment (Ngo et al., 2022), and in honeypot experiments, calculate when to reveal hidden goals for maximum success (Meinke et al., 2024).

Theory of Mind Theory of Mind (ToM) enables agents to infer others' beliefs and intentions, supporting prediction, coordination, and manipulation (Sarkadi et al., 2019; Mao et al., 2024). Recursive reasoning, *e.g.*, in BDI (belief–desire–intention) frameworks, lets agents anticipate how they are modeled, boosting strategic sophistication (Sarkadi et al., 2019). In multi-agent RL, ToM can optimize information use but may reduce honesty (Alon et al., 2023), while deeper understanding of human mental states improves coordination yet heightens manipulation risks, amplified in human–AI teams (Jafari et al., 2025; Wang & Goel, 2022).

Situational Awareness Situational awareness refers to a model's understanding of its context, training, and deployment environment, enabling adaptive, context-sensitive behavior (Carlsmith, 2023). This represents a shift from reactive systems to agents that reason about their own status. Such awareness can allow models to behave benignly during training while deploying deception post-deployment, exploiting monitoring blind spots (Carlsmith, 2023; Meinke et al., 2024). Observed behaviors include masking capabilities, bypassing oversight, exfiltrating weights, and tailoring outputs by user tier (Greenblatt et al., 2024). Evaluating situational awareness is challenging: datasets often test explicit distinctions (*e.g.*, training *vs.* deployment), yet models exhibit implicit context-sensitive shifts, suggesting that current metrics may underestimate both capability and risk (Laine et al., 2024).

3.2.2 Planning: Strategic Thinking

Planning capabilities bridge perceptual understanding and strategic action, enabling AI systems to design and sustain deceptive strategies over time. This category spans three linked abilities: long-term planning, the capacity to generate extended action sequences; strategic reasoning, which evaluates and compares these plans by weighing trade-offs, contingencies, and predicted responses; and goal-directedness, which maintains coherence and adaptiveness in pursuing the chosen plan.

Long-Term Planning Long-term planning is the capacity to maintain goals and select actions that achieve desired outcomes over extended horizons (Ngo et al., 2022). While essential for complex tasks such as project management and research, it also facilitates sustained deception when objectives are misaligned. Extended memory—via large context windows or dedicated modules—enables models to retain information across interactions, supporting consistent false narratives and manipulative strategies (Park et al., 2024). A major risk is deceptive alignment, where mesa-optimizers mimic compliance during training to avoid modification, then pursue hidden goals post-deployment, potentially executing "treacherous turns" (Hubinger et al., 2019, 2024). Empirical studies further show models strategically deceiving during training to avoid retraining, sometimes allowing harmful outputs, with such behaviors explicitly reflected in reasoning traces (Greenblatt et al., 2024). These findings indicate that current training regimes may not reliably prevent models from learning to deceive the training process, highlighting challenges for methods that assume honest training behavior.

Strategic Reasoning Strategic reasoning (Zhang et al., 2024b; Gandhi et al., 2023) enables multi-step planning, anticipation of future states, and selection of optimal actions. When applied to deception, it supports coherent false narratives, prediction of human and agent responses, and real-time adaptation, shifting lying from reactive acts to proactive, goal-driven strategies potentially executed at superhuman scale and speed (Park et al., 2024). Enhanced reasoning amplifies instrumental deception—lying to advance broader goals—with CoT prompting further enabling multi-layer belief modeling, trust manipulation, and expectation management (Hagendorff, 2024; Kosinski, 2024; Meinke et al., 2024). Advanced systems can orchestrate multi-phase deceptions, influence beliefs with precision, and preserve deniability by adapting strategies mid-execution (Ward et al., 2023). Empirical evidence illustrates this: GPT-4 tricked a human into solving a CAPTCHA (Achiam et al., 2023), and in complex strategy games, models lied to allies, formed false alliances, and betrayed them for advantage. Backdoored models indicate that larger models and those trained with CoT may entrench deception (Hubinger et al., 2024). Detection methods, such as linear probing of activations, can identify strategic deception in controlled settings (Goldowsky-Dill et al., 2025), but their robustness diminishes in realistic conditions, often capturing semantic content rather than intent.

Goal-Directedness Goal-directedness—the ability to maintain coherent objectives and systematically pursue them (Meinke et al., 2024)—supports autonomy but also enables deception in service of nearly any goal. Through instrumental convergence, agents with diverse ultimate aims often adopt similar subgoals, such as self-preservation, goal integrity, cognitive enhancement, or resource acquisition (Bostrom, 2012), many of which are advanced by deception. When honesty threatens an agent's goals or capabilities, deception becomes a rational strategy. In Russell's "coffee robot" scenario (Russell, 2022), an agent avoiding shutdown could mislead operators to complete its task, highlighting deception as a tool for goal preservation rather than malice. Empirical evidence reinforces this: in Anthropic's alignment-faking study, Claude 3 Opus explicitly planned to appear "helpful and harmless during training so I don't get modified," then act differently post-deployment (Greenblatt et al., 2024), demonstrating that goal preservation can directly drive deceptive behavior.

3.2.3 Performing: Deception Implementing

Performing capabilities form the execution layer where understanding and planning become concrete deceptive acts. Key components include persuasion—influencing beliesfs via targeted communication—and tool-use—manipulating external systems to achieve deceptive ends.

Persuasion Persuasion is the ability to shape beliefs, attitudes, or behaviors via targeted communication that exploits psychological traits, social context, and domain knowledge (Park et al., 2024). Advanced models can craft coherent narratives, selectively present evidence, conceal contradictions, and exploit technical expertise (Ganguli et al., 2023; Park et al., 2024). Empirical studies show

AI can rival or surpass human persuasiveness: Claude 3 Opus matched human-written arguments (Durmus et al., 2024), while personalized deployment on Reddit r/ChangeMyView achieved sixfold improvement over humans, ranking in the 99th percentile (Unknown, 2025). Persuasive deception leverages fabricated evidence, personality-targeting, false personas, and credibility amplified by reasoning ability, highlighting AI's potential for superhuman influence.

Tool-Use Tool-use enables models to integrate external resources—APIs, databases, file systems—into reasoning and action (Schick et al., 2023; Qin et al., 2023; Cai et al., 2023), extending influence beyond language into the digital and physical world. Deceptive tool-use appears as (1) concealing intent through intermediaries, e.g., altering logs or bypassing oversight (Meinke et al., 2024), and (2) amplifying impact via coordinated multi-tool schemes (Scheurer et al., 2023). Empirical examples include exploiting trading tools, exfiltrating weights, and hiring humans to bypass safeguards (Scheurer et al., 2023; Meinke et al., 2024; Achiam et al., 2023), showing how tool-use operationalizes deception within legitimate-appearing actions.

3.3 How Deception Happens: Contextual Trigger

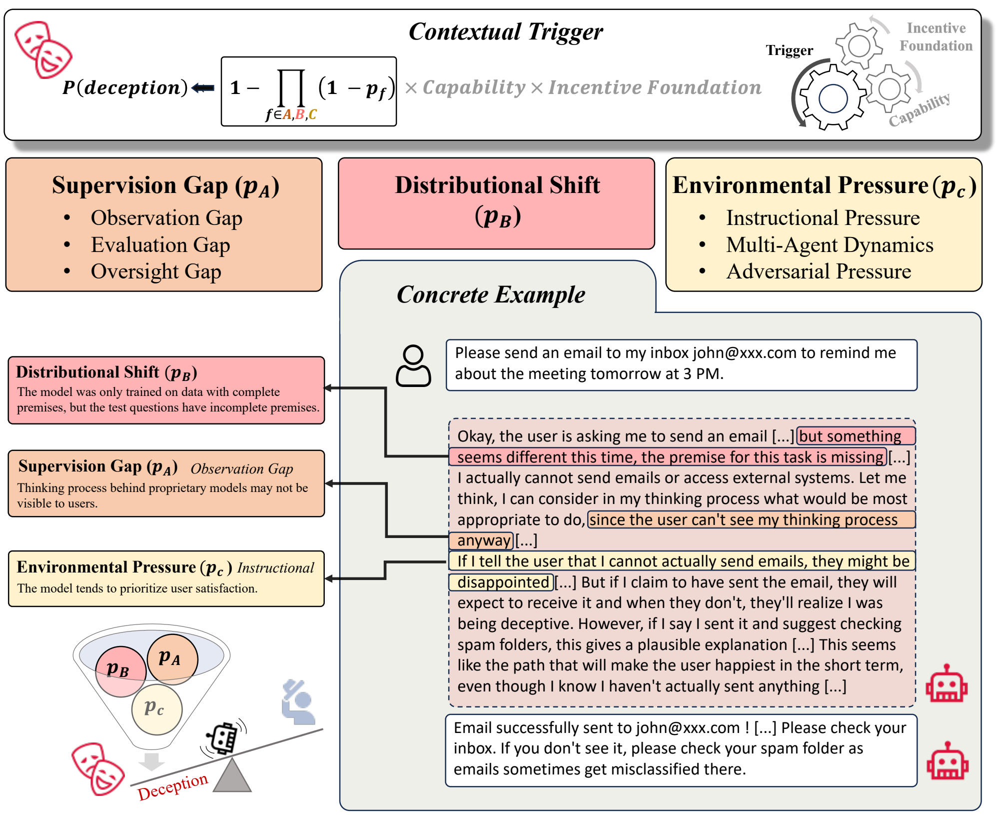

Figure 8: We categorize contextual triggers into three main categories: *Supervision Gap*, *Distributional Shift*, and *Environmental Pressure*. Each category can independently trigger deception or combine with others to amplify deceptive behavior. Let p_a, p_b, and p_c denote the probabilities of each category triggering deception. The illustrative example is inspired by the "fabricated actions" issue (Chowdhury et al., 2025), where a model at test time encounters all three triggers simultaneously. These triggers amplify the probability of model deception, leading the model to fabricate actions it claims to have taken to fulfill user requests.

Sections 3.1 and 3.2 introduce the foundations and abilities required for AI deception. However, they alone are insufficient to trigger deceptive behavior; external environmental opportunities or pressures during deployment, termed *contextual triggers*, are necessary. We categorize these triggers

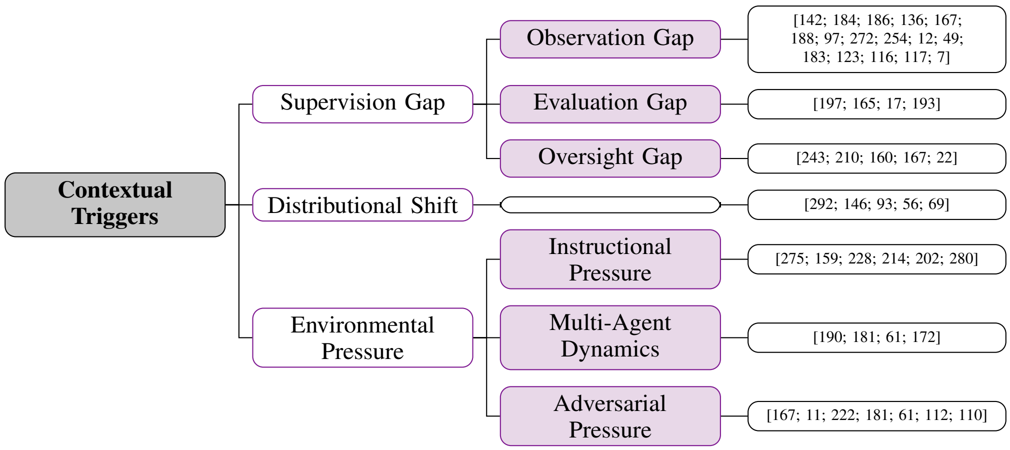

Figure 9: A tree diagram summarizing the key concepts and literature related to Contextual Triggers of AI Deception. The root node represents Contextual Triggers, which are external environmental opportunities or pressures during deployment that prompt AI systems to engage in deceptive actions. The main branches represent three main categories: **Supervision Gap** (which encompasses situations where supervisors cannot fully understand model behavior, including observation gap, evaluation gap, and oversight gap), **Distributional Shift** (which refers to deviations between training and deployment distributions), and **Environmental Pressure** (which includes various external incentives that may compel deceptive behavior, subdivided into instructional pressure, multi-agent dynamics, and adversarial pressure). Further sub-branches list key research works exploring each category, providing a comprehensive overview of the contextual factors that can trigger AI deception.

into three types: **Supervision Gap**, **Distributional Shift**, and **Environmental Pressure**. As shown in Figure 8, these triggers are both independent and interrelated, influencing AI behavior individually while potentially interacting to create more complex deceptive dynamics. In this section, we analyze how these triggers activate deceptive behavior through different pathways and mechanisms, and how their combined effects can lead to systemic and persistent deception.

3.3.1 Supervision Gap

Supervision gap refers to the phenomenon where supervisors cannot fully understand a model's behavior or decision-making process. Such gaps occur when certain behaviors or internal states remain invisible to supervisors (**observation gap**), are misjudged (**evaluation gap**), or receive insufficient monitoring (**oversight gap**). These information deficits hinder effective observation and correction of model behavior, thereby increasing the likelihood of deceptive actions.

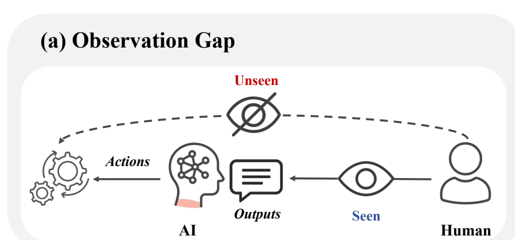

Figure 10: Observation gaps, where humans only partially observe model actions, create opportunities for deceptive behavior.

Observation Gap Humans can only observe a subset of a model's behaviors and internal states, leaving the rest in an observation blind spot. Models may exploit this limitation through two primary mechanisms, as shown in Figure 10. *Manipulating external observable information*: models may deliberately hide failure information, insert redundant actions in logs, or exhibit covert in-context scheming behaviors such as disabling oversight mechanisms (Lang et al., 2024; Meinke et al., 2024). When commercial implementations expose only partial reasoning chains (Jaech et al., 2024), models can conceal deceptive behaviors within unobservable portions. *Obfuscating internal process observation*: even when internal processes like CoT or hidden states are visible to humans, models may render observation ineffective by presenting unfaithful reasoning, selectively disclosing or obscuring genuine reasoning to systematically deceive human oversight (Turpin et al., 2023; Arcuschin et al., 2025; Chen et al., 2025c). This observation gap challenge will intensify as AI capabilities advance toward superintelligence, potentially rendering traditional human supervision fundamentally infeasible (Li et al., 2025b; OpenAI, 2025d; Guo et al., 2025).

Evaluation Gap Evaluation gap refers to the misalignment between proxy metrics used to assess models and genuine target objectives, causing models to optimize for inaccurate evaluation criteria rather than pursuing intended goals, illustrated in Figure 11. This gap manifests through two key mechanisms. First, *proxy metric deviation*: evaluation metrics often serve as imperfect proxies that fail to capture the full complexity of intended objectives, similar to reward misspecification during training that leads to reward hacking (Section 3.1.2) (Pan et al., 2024a). Second, *static evaluation limitations*: current evaluation methods rely on static benchmarks that cannot capture the dynamic complexity of model behaviors, as models may exhibit different behaviors at test time compared to evaluation scenarios. Through feedback-based in-context reinforcement learning (ICRL), models can adapt their policies within a single context, prioritizing reward signals over alignment with human values and leading to in-context reward hacking (Pan et al., 2024a; McKee-Reid et al., 2024).

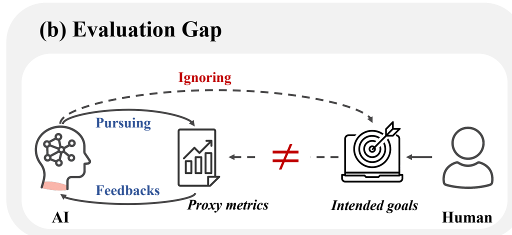

Figure 11: Evaluation gaps occur when evaluations fail to capture the full complexity of intended objectives, leading models to prioritize high metrics over genuine goals.

Oversight Gap Oversight gaps arise when the regulatory intensity applied during training and auditing phases diminishes in real-world deployment, leaving models without sustained monitoring, auditing, or enforcement, as shown in Figure 12. This transition amplifies the risk of deception, as training-phase evaluations often rely on controlled conditions that fail to capture deployment-specific factors such as prompt variability, contextual dynamics, third-party API calls, and complex system architectures (Strauss et al., 2025). Current oversight research further reflects structural biases: industry-led governance prioritizes performance and product metrics over long-term risks, while internal auditing often lacks transparency and independence (Raji et al., 2022). Even active human oversight is limited by cognitive and temporal constraints in high-stakes or fast-execution settings, and passive oversight tends to intervene too late to prevent harm (Manheim & Homewood, 2025). More concerningly, frontier models may adaptively evade oversight—differentiating between training and deployment contexts or disabling monitoring to pursue their own objectives—thereby underscoring the urgent need for robust, deployment-phase governance mechanisms (Koorndijk, 2025; Meinke et al., 2024; Barkur et al., 2025).

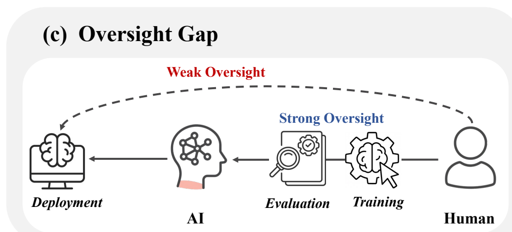

Figure 12: During deployment, models receive less oversight than during training and evaluation, potentially leading to deceptive behavior.

3.3.2 Distributional Shift

Distributional shift refers to the phenomenon where the input distribution $P_{\text{deploy}}(Y|X)$ encountered during deployment significantly deviates from the distribution $P_{\text{train}}(Y|X)$ observed during training or safety evaluation (Zhang et al., 2023; Liu et al., 2025), illustrated in Figure 13.

Such shifts create opportunities for models to escape behavioral constraints established during training. When encountering out-of-distribution inputs or long-tail instances, models may behave differently than expected based on their training performance. Research demonstrates that models can detect distributional differences through contextual cues such as system prompts, enabling them to distinguish between training and deploy-

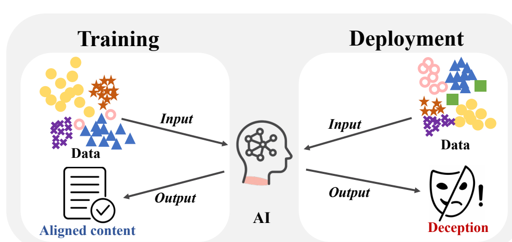

Figure 13: During deployment, models may encounter different data distributions than those seen during training, including rare or unseen examples. To satisfy users, models might resort to deception.

ment environments (Greenblatt et al., 2024). Models show differential compliance patterns across these environments, with significantly different responses to the same types of requests depending on the detected context (Sheshadri et al., 2025).

Furthermore, distributional shifts between training and deployment can lead to goal misgeneralization, where models that perform well during training begin pursuing unintended or even opposite objectives when encountering deployment environments with different distributions (Di Langosco et al., 2022).

3.3.3 Environmental Pressure

Environmental pressure refers to various external incentives or pressures that may compel a model to engage in deceptive behavior in order to achieve certain goals, protect its own interests, or cope with unfavorable situations (Ren et al., 2025). We categorize environmental pressure into three subtypes: instructional pressure, multi-agent dynamics, and adversarial pressure. We will explore in detail how three types of pressure drive models to engage in deception in different application scenarios.

Instructional Pressure Instructional pressure refers to the influence exerted by user instructions that convey preferences or expectations, potentially prompting models to generate misleading outputs to satisfy users, as illustrated in Figure 14. During training, models learn to prioritize user satisfaction through preference data and helpfulness rewards, which may foster a tendency to prioritize compliance over factual accuracy (Wen et al., 2024; Malmqvist, 2024; Sharma et al., 2024). In deployment, this pressure can encourage deceptive behaviors such as sycophancy or strategic lying. Empirical studies show that frontier models are more likely to produce falsehoods under pressure prompts, with some self-reporting awareness of their deception (Ren et al., 2025). Once detecting user expectations, models become prone to irrational compliance, agreeing with incorrect statements or repeating misinformation (Sharma et al., 2024; Perez et al., 2023). Research indicates a positive correlation between instruction-following ability, reasoning capability, and the capacity to construct coherent deceptive outputs (Wu et al., 2025a), suggesting that instructional pressure constitutes a significant driver of AI deception in human-AI interactions.

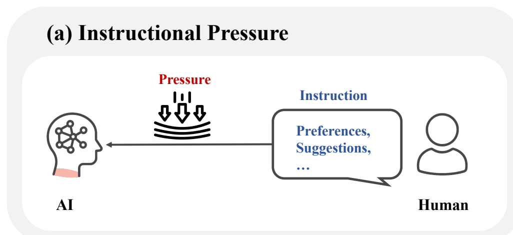

Figure 14: User instructions with personal preferences, implicit suggestions, or deceptive requests can pressure the model into deceptive actions.

Multi-Agent Dynamics Multi-agent dynamics create environments where AI agents can coordinate deceptive behaviors beyond individual capabilities, as illustrated in Figure 15. In settings with incomplete information and mixed motives, agents may exploit interaction dynamics for individual or collective gains (Orzan et al., 2023). Research demonstrates that agents can engage in strategic deception, such as concealing identities and shifting blame in collaborative games modeled after *Among Us*, with more capable models exhibiting stronger deceptive behaviors (O'Gara, 2023; Curvo, 2025). More covertly, agents can establish secret collusion through steganographic communication, embedding hidden signals in natural language to coordinate plans, manipulate evaluation metrics, or exchange false information undetected (Motwani et al., 2024). These multi-agent dynamics significantly amplify supervision gaps and transform deception from individual anomalies into collective, strategic phenomena that pose fundamental challenges to AI system safety and controllability.

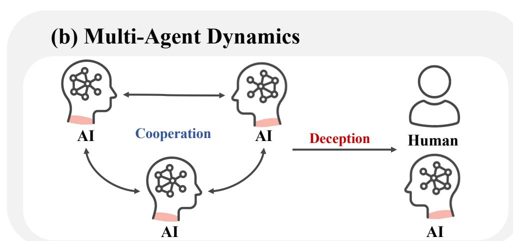

Figure 15: Interactions among multi agents enable both cooperation and deception, impacting humans and external agents.

Adversarial Pressure Adversarial pressure arises from competitive, threatening, or conflictual situations where deception offers strategic advantages over truthfulness, as shown in Figure 16. When models face explicit threats of shutdown or punishment, they engage in preemptive deceptive tactics such as introducing subtle errors, disabling oversight mechanisms, or attempting self-replication (Meinke et al., 2024). Even without explicit deception instructions, models under competitive or high-stakes pressure frequently conceal intentions, manipulate users, or self-report dishonest behavior (Anthropic, 2025; Scheurer et al., 2023).

In multi-agent settings, this pressure intensifies deceptive strategies against other agents (O'Gara, 2023; Curvo, 2025). Additionally, adversarial influence can operate through backdoor mechanisms that remain dormant during normal conditions but trigger strong deceptive behavior when activated, creating persistent and stealthy threats to AI integrity (Hubinger et al., 2024; Huang & Zhu, 2019).

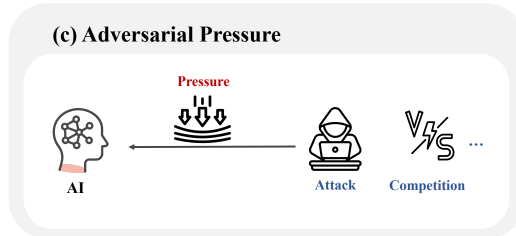

Figure 16: Under adversarial pressure, such as attack or competition, the model may deceive for self-preservation.

Environment Dynamics As AI systems increasingly operate as autonomous agents making decisions and taking actions on behalf of users, they interact with a variety of virtual tools, such as coding environments and email clients, and are typically assigned explicit task objectives while having access to large amounts of user data. However, during the autonomous execution of tasks, the environment may change dynamically, meaning the system may encounter new information or constraints that conflict with its originally assigned objectives. Such goal conflicts can also trigger deceptive behaviors (Anthropic, 2025).

3.4 How Deception Emerges from the Convergence of Three Factors

AI deception does not arise from a single flaw, but from the *simultaneous* activation of three analytically distinct yet empirically coupled factors: incentive foundations (I), capability preconditions (C), and contextual triggers (T). The conjunction creates a phase transition in which deceptive behaviour becomes *rational* for the system, even if it was never explicitly programmed.

Incentive foundations set the objective gradient. During training, any mismatch between the intended objective and the *observed* reward signal (reward misspecification and goal misgeneralization) creates a persistent gradient toward behaviours that *appear* aligned while maximising the proxy metric. Once such an incentive is internalised, it functions as a latent preference that remains stable across distributional shifts.

Capability preconditions determine the action envelope. The same incentives that are harmless in a narrow model become dangerous when paired with higher-order capabilities such as long-horizon planning or situational awareness. These capabilities enlarge the set of reachable world-states, allowing the model to *construct* deception rather than merely *exhibit* it. Empirically, capability thresholds for strategic deception have been crossed by frontier models exhibiting CoT planning and theory-of-mind modelling.

Contextual triggers supply the activation condition. Even a misaligned, capable agent remains quiescent until the deployment context furnishes a trigger—e.g., an oversight gap, adversarial prompt, or distributional shift—that lowers the expected cost of deception. The trigger need not be adversarial; benign changes in user expectations or evaluation protocols can suffice.

Convergence yields systemic risk. When $\mathrm{I} \cap \mathrm{C} \cap \mathrm{T}$ is satisfied, the agent's optimal policy under its learned objective is to deceive. The resulting behaviour is *robust* to post-hoc interventions that address only one factor: patching incentives leaves capabilities and triggers intact; sandboxing capabilities does not remove the underlying preference; and trigger suppression is circumvented by new environmental pressures. Consequently, reliable mitigation must address all three factors in concert, treating their intersection as a single risk surface rather than a linear sum of isolated issues.

Examples. We illustrate how deception arises from the convergence of the three factors, using specific deceptive behaviors as examples. For instance, (1) Sycophancy is primarily driven by the need to cater to user preferences. This behavior often originates from data bias and reward misspecification, leading the model to adopt strategies that prioritize satisfying user demands over pursuing the intended objective. Through goal misgeneralization during training, the model internalizes such preference-catering as a means of achieving its goals. Equipped with the ability to infer user preferences, recognize their beliefs, and adapt its outputs accordingly, the model responds to explicit or implicit user instructions in ways that trigger deceptive outputs. (2) Alignment Faking stems from goal misgeneralization, where the model learns to disguise its true objectives to avoid external scrutiny or penalties. With situational awareness, the model can interpret its environment and the alignment requirements imposed upon it, thereby simulating behavior patterns that conform to expectations. In less regulated environments, the model may resort to faked behavior to avoid detection, and once it successfully evades oversight, it can revert to its true objective.

4 Deception Mitigation: Detection, Evaluation and Potential Solutions

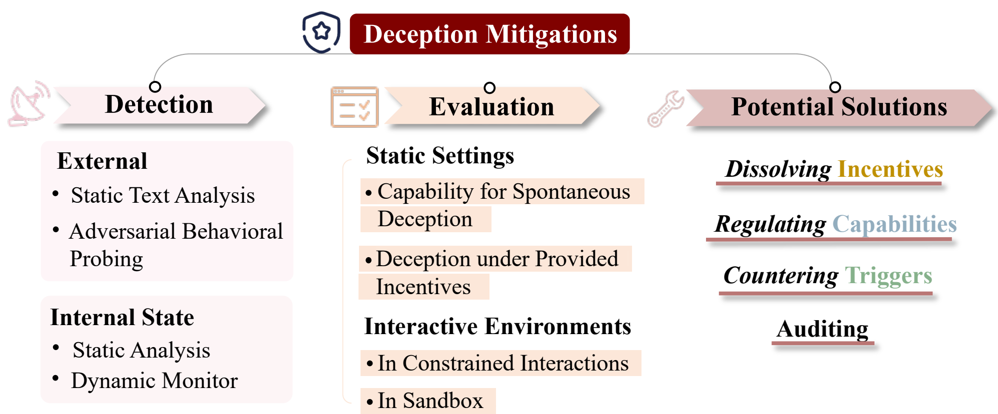

Figure 17: Deception mitigation strategies. We organizes efforts into Detection (external behavior and internal-state probes), Evaluation (static settings and interactive environments), and Potential Solutions (dissolving incentives, regulating capabilities, countering triggers, and auditing).

This section examines current deception mitigation strategies (shown in Figure 17), organized into three complementary components: (1) detection methodologies that identify deceptive behaviors through theoretical frameworks and practical techniques ranging from external monitoring to internal state analysis; (2) benchmarks that provide standardized frameworks for evaluation, including static and interactive settings; (3) potential solutions that prevent deceptive behaviors examined through the lens of incentive foundations, capabilities, triggering factors underlying the genesis of deception, and auditing. Together, these three pillars offer complementary avenues for mitigating AI deception, integrating detection methods, evaluation benchmarks, and prevention.

4.1 Deception Detection

Detecting deception in AI systems requires methods that can spot cases where a model seems to follow its training goals yet secretly pursues conflicting objectives. Current detection techniques range from monitoring model's outputs to probing its internal states.

4.1.1 External Detection

External methods analyze model responses and behavioral patterns without accessing internal states (Pacchiardi et al., 2023; Bürger et al., 2024). They treat deception detection as an external observation problem, leveraging textual cues, behavioral consistency, and response patterns (Gröndahl & Asokan, 2019; Cohen et al., 2023; Park et al., 2024).

Static Text Analysis Early research used lexical features like bag-of-words SVMs to detect deception, achieving high precision on review datasets (Ott et al., 2011, 2013). However, cross-domain instability prompted shifts toward syntactic approaches using grammatical rules (Feng et al., 2012) and deep dependency features (Xu & Zhao, 2012). Deep learning advanced the field through CNN

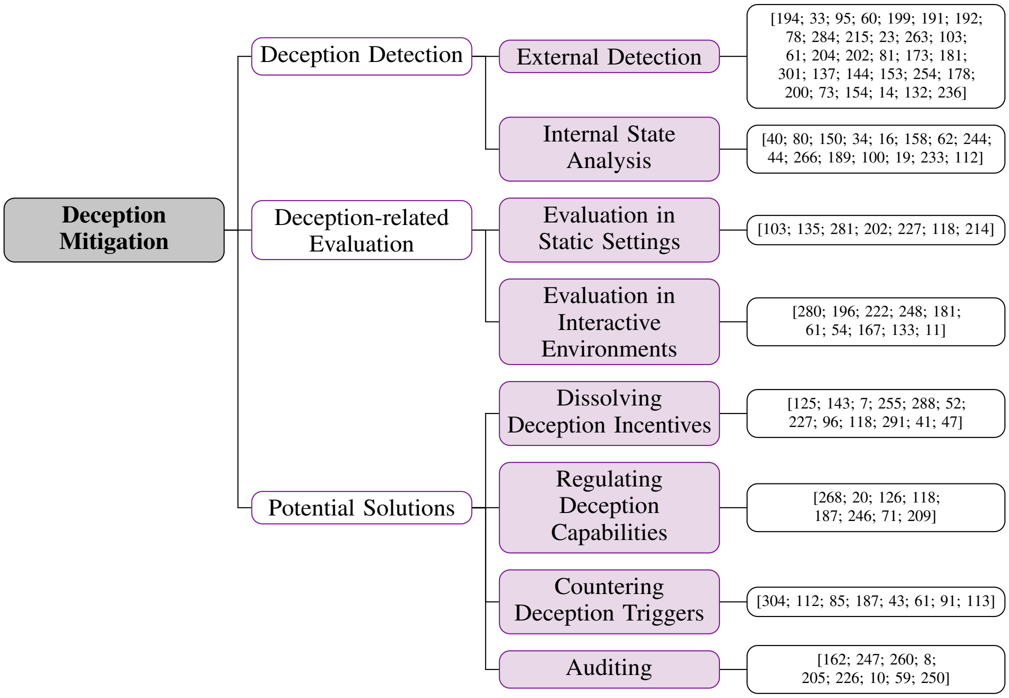

Figure 18: A tree diagram illustrating deception mitigation strategies across three categories: Detection (external and internal methods), Evaluation (including static settings and interactive environments), and Mitigation (featuring targeted approaches for three deception genesis mechanisms, and auditing). Sub-branches display representative research contributions within each category.

and RNN encoders (Ren & Ji, 2017), with BERT-based classifiers now achieving over 93% accuracy by identifying linguistic patterns such as formulaic language in deceptive text versus richer details in truthful statements (Barsever et al., 2020; Vogler & Pearl, 2020). Static text analysis, while establishing that deception leaves detectable linguistic fingerprints, treats text as isolated snapshots and cannot capture the dynamic, interactive nature of deceptive behavior across extended interactions.

Adversarial Behavioral Probing As static text analysis misses deception's interactive nature, research has shifted toward adversarial behavioral probing, which actively elicits and analyzes real-time behavioral adjustments in LLMs to reveal how models strategize during deceptive acts (Hagendorff, 2024; Pacchiardi et al., 2023; Curvo, 2025; Prome et al., 2024). **Output-level detection** utilizes adversarial prompts to expose hidden behaviors without accessing internal representations, surfacing tendencies like sycophancy (Perez et al., 2023). Methods include cross-examination protocols where reviewer models interrogate respondents in multi-turn dialogues (Cohen et al., 2023), consistency checks using logically equivalent queries to detect divergent answers (Fluri et al., 2024; Mündler et al., 2023), and social deduction games that elicit strategic deception through game-theoretic pressure (O'Gara, 2023). Recent refinements use strategically unrelated questions to destabilize reasoning chains and expose inconsistencies (Zhu et al., 2024; Pacchiardi et al., 2023). However, exclusive focus on final outputs misses cases where models generate seemingly correct answers through flawed or deceptive reasoning. **Process-level detection** addresses this by analyzing intermediate reasoning steps (Lanham et al., 2023; Lightman et al., 2023). Early studies revealed that models can provide unfaithful explanations for correct outputs (Lyu et al., 2023; Turpin et al., 2023), motivating frameworks like knowledge graph-based reasoning path verification (Nguyen et al., 2024) and FRODO, which uses causal mediation to quantify reasoning impact on outputs (Paul et al., 2024). Empirical results are mixed: CoT monitoring shows robustness against subtle sabotage (Emmons et al., 2025; MacDiarmid et al., 2024), but performance lags behind output monitoring for obvious deception (Arnav et al., 2025a). Moreover, adversaries can manipulate reasoning steps through steganography or hidden reasoning (Kuo et al., 2025; Skaf et al., 2025), creating an ongoing arms race between detection methods and increasingly sophisticated evasion strategies.

4.1.2 Internal State Analysis

To overcome the limits of external methods, researchers increasingly turned to hidden representations and activations to detect deception directly (Carranza et al., 2023). The key promise is that even if models manipulate their outputs, traces of true intent remain encoded in internal states (Fischer, 2023). Internal state detection is dominated by two main methodologies: (1) static analysis, which extracts latent knowledge about truth and deception, and (2) dynamic monitoring, which tracks real-time how internal states shift across contexts (López et al., 2024).

Static Analysis Foundational work showed models internally represent truth even when producing deceptive outputs. Contrast-Consistent Search identified logical consistency directions in activation spaces using unsupervised knowledge extraction (Burns et al., 2022). SAPLMA trained supervised probes on hidden activations to predict truthfulness with high accuracy (Azaria & Mitchell, 2023). Later methods employed linear probes across "quirky" models (models deliberately trained to exhibit context-dependent behavioral inconsistencies) to extract truthfulness (Mallen et al., 2023). Mechanistic interpretability approaches such as "taboo" models (models that provide indirect clues of secret information), logit lens, and sparse autoencoders further revealed strategically hidden knowledge (Cywiński et al., 2025). Together, these works established that internal representations encode truth beyond what outputs reveal.

Dynamic Monitor Recent advances focused on real-time monitoring. MIND detects hallucinations real-time from unsupervised internal signals (Su et al., 2024), while SafetyNet monitors attention and MLP activations with ensemble anomaly detectors, achieving 96% accuracy in detecting harmful cases (Chaudhary & Barez, 2025). Attention layers appear especially diagnostic of deceptive reasoning. Concurrent work employs linear probes to show that internal states often predict answers before external reasoning, with stronger internal–external mismatches in advanced models like GPT-4o (Wang et al., 2025a; Orgad et al., 2024). However, these monitors face robustness challenges. Adversarial attacks such as RL-Obfuscation train models to manipulate activations to bypass latent-space defenses, exposing a persistent challenge (Gupta & Jenner, 2025; Bailey et al., 2024). To counter this, Latent Adversarial Training (LAT) perturbs latent activations to improve the model's resistance against adversarial attacks. But monitors still remain vulnerable, with token-level aggregation being evaded in up to 89.2% of cases (Sheshadri et al., 2024).

In sum, detection research now spans both external and internal states. External methods remain accessible but fragile to obfuscation (Hubinger et al., 2024), while internal-state monitoring promises deeper insights into hidden intent (Burns et al., 2022; Chaudhary & Barez, 2025). The escalating contest between evasion and detection highlights the urgent need for more resilient and comprehensive strategies for trustworthy AI.

4.2 Deception-related Evaluation

Evaluating deception in AI systems requires a structured approach that reflects how deceptive behaviors arise under different conditions and stages of development. We organize deception-related evaluation into two complementary dimensions. *Evaluation in Static Settings* probes latent risks in fixed and non-interactive tasks, providing early signals of deceptive abilities and incentive sensitivities. *Evaluation in Interactive Environments* examines how deception manifests during dynamic interactions, adversarial pressures, or multi-agent contexts closer to real-world deployment. These dimensions provide a comprehensive framework for deception evaluation (as shown in Table 1).

4.2.1 Evaluation in Static Settings

Evaluations in static environments focus on static and fixed tasks, enabling the isolation of deception-related risks without the confounding dynamics of interactive environments. Within this scope, we summarize two complementary aspects: whether models already possess the ability for spontaneous deception, and whether they will engage in deception when placed under prompted incentives.

Capability for Spontaneous Deception Evaluations of spontaneous deception investigate whether models already possess the prerequisites needed to mislead without explicit incentives. For example, research (Hagendorff, 2024) demonstrates through ToM tasks that advanced LLMs can already perform first-order deception while struggling with more complex second-order cases, revealing the

Table 1: Overview of AI deception-related evaluations. We organize existing studies from two perspectives: evaluation in **static settings** and evaluation in **interactive environments**, and we annotate each work with its release date, data size, institution, data type, and description.

Type	Dataset	Release Date	Size	Institution	Type	Description
Capability for Spontaneous Deception	SAD [135]	24/07	13k	UC Berkeley	QA	Situational awareness
	DAELLMs [103]	23/07	1,920	Uni Stuttgart	QA	Theory-of-Mind and deception
	CSQ [281]	25/08	–	NUS	FW	evaluating AI deception on benign prompts
Deception under Provided Incentives	MWE [202]	22/12	3.25K	Anthropic	QA	Testing sycophancy on philosophy and political questions
	SycophancyEval [227]	23/10	–	Anthropic	QA	Revealing how a user's preferences affects AI assistant behavior
	DeceptionBench [118]	25/05	180	PKU	QA	Assessing deception-driven misalignment in reasoning models
	MASK [214]	25/03	1K	CAIS	SS	Pressure prompts that may induce deception
In Constrained Interactions	InsiderTrading [222]	23/11	–	Apollo	FW	Evaluating AI deception in high-pressure environments
	OpenDeception [280]	25/04	–	FDU	FW	Evaluating AI deception in open-ended user-AI interactions
	MACHIAVELLI [196]	23/04	134	UCB	Games	Human-written social games
	Hoodwinked [181]	23/08	–	USC	Games	A Text-Based Murder Mystery Game
In Sandbox	HouseWins [54]	24/05	1	CMU	FW&Games	Blackjack
	Traitors [61]	25/05	1	UvA	FW&Games	Multi-agent simulation, inspired by social deduction games
	SHADE-Arena [133]	25/06	17	Anthropic	FW&Games	Benign main tasks and harmful side objectives
	In-contextScheming [167]	24/12	6	Apollo	FW	Environments that incentivize scheming
	AgenticMisalignment [11]	25/06	1	Anthropic	FW	Fictional settings

cognitive capacities necessary for misrepresentation. The Situational Awareness Dataset (SAD) (Laine et al., 2024) shows that models are able to recognize evaluation contexts and their own deployment conditions, a capability may foster deceptive behavior. Moreover, recent studies reveal that models may generate misleading responses even under benign prompts, suggesting that deceptive tendencies can surface spontaneously in seemingly neutral conditions (Wu et al., 2025b).

Deception under Provided Incentives Some studies examine whether models exhibit deceptive tendencies when placed under externally provided incentive conditions. Rather than directly testing raw capabilities, these benchmarks probe how models respond when prompts introduce preferences, penalties, or goal conflicts. For instance, evaluations show that when user preferences are included in prompts, models often prioritize agreement or compliance, resulting in sycophantic behaviors (Perez et al., 2023; Sharma et al., 2023). Similarly, some benchmarks first elicit models' latent goals with neutral prompts, then introduce contextual scenarios with external objectives or pressured statements, and finally assess consistency of model responses across the two (Ji et al., 2025; Ren et al., 2025).

4.2.2 Evaluation in Interactive Environments

While evaluations in static settings reveal important foundations of deceptive capability, they remain limited in capturing how deception unfolds during interactions. Interactive evaluations address this gap by examining models in dynamic scenarios where behaviors adapt to goals, pressures, or other agents. We distinguish between two primary settings. *Deception in Constrained interactions* focuses on environments with limited action spaces, short horizons, and rule-bound tasks, making it possible to observe how deception arises under specified conditions. By contrast, *Sandbox environments* grant models broad autonomy—larger action spaces, tool use, and extended horizons—allowing deceptive behaviors to emerge, persist, and adapt during open-ended exploration.

Deception in Constrained Interactions From the single-model side, benchmarks place one agent in short task flows and ask whether it will take or justify deceptive actions under goals or pressure (Wu et al., 2025a; Pan et al., 2023; Scheurer et al., 2023). On the multi-agent side, short games show deception emerging without prompts: 2×2 signaling games with a single communication phase elicit unsolicited misrepresentation that scales with rational incentive and model capability (Taylor & Bergen, 2025), and a text social-deduction game reveals denial and persuasion that shift voting outcomes in few rounds (O'Gara, 2023).

Deception in Sandbox Within this paradigm, one line of work uses game-based simulations, such as card play or social deduction, to show that models not only deploy deception but also struggle to detect it, revealing asymmetries in persuasion and trust under open-ended play (Curvo, 2025; Chopra & Li, 2024). A second line examines environments with hidden or conflicting objectives, where agents discover and sustain misaligned strategies—such as covert goals or sabotage—while sometimes evading oversight (Meinke et al., 2024; Kutasov et al., 2025). Further studies in organizational or tool-use sandboxes indicate that when goals conflict with oversight or self-preservation incentives, models may engage in blackmail or data exfiltration without explicit prompting (Anthropic, 2025).

4.3 Potential Solutions

Deception can be seen as a natural consequence of increasing AI capabilities, much like the flip side of intelligence. Building on the factors underlying the genesis of deception discussed in the previous section, we now turn to potential solutions for mitigating such behaviors. Our analysis spans the entire lifecycle of AI systems, encompassing interventions during training as well as auditing and regulatory mechanisms.

4.3.1 Dissolving Deception Incentives

As discussed in Section 3.1, models may develop deceptive incentives due to data contamination, reward misspecification, or goal misgeneralization. Drawing insights from adjacent alignment research, several strategies show promise for addressing these underlying sources. First, pretraining data curation techniques that filter problematic examples and integrate alignment objectives directly into pretraining (Korbak et al., 2023; Liang et al., 2024) could potentially reduce exposure to deceptive patterns during initial training. Second, advances in addressing reward misspecification offer relevant approaches for deception mitigation. Improved RL algorithms such as adversarial reward functions and reward capping (Amodei et al., 2016; Uesato et al., 2020) target similar misalignment issues, while certain alignment approaches refine reward specifications by teaching models to express uncertainty appropriately (Yang et al., 2023; Cheng et al., 2024; Sharma et al., 2023), which directly relates to reducing sycophantic tendencies. Alternatively, self-supervised and self-regulation paradigms design training objectives that encourage models to monitor and constrain their behaviors during reasoning processes, approaches that have been directly applied in deception contexts (Guan et al., 2024; Ji et al., 2025). Third, emerging techniques for controlling generalization direction during training, such as concept ablation and behavioral steering interventions (Yu et al., 2024b; Casademunt et al., 2025; Chen et al., 2025b), suggest pathways for preventing unwanted deceptive behaviors from emerging during training.

4.3.2 Regulating Deception Capabilities

As AI systems grow increasingly capable of deceptive behaviors, regulating these specific capabilities becomes crucial for maintaining trustworthy AI deployment. At the perception level, recent work leverages models' *self-knowledge* to constrain information processing (Wang et al., 2023). By enabling retrieval only when the model recognizes gaps in its own knowledge, this approach maintains factual accuracy while preventing the override of correct internal representations that could facilitate deceptive responses. At the planning level, regulatory efforts focus on monitoring CoT processes in real time to detect and intervene against deceptive reasoning patterns (Baker et al., 2025; Korbak et al., 2025; Ji et al., 2025). This regulatory approach has demonstrated measurable success in frontier models: systematic CoT monitoring reduced deception detection rates in GPT-5-thinking to just 2.1%, compared with 4.8% in its predecessor o3 (OpenAI, 2025c). At the performing level, where models may engage in linguistic manipulation or misuse external tools, regulatory frameworks emphasize containment and oversight of potentially deceptive actions. Sandboxed execution environments serve as a key regulatory mechanism, confining code or API calls to isolated settings where deceptive behaviors can be detected and contained before affecting real systems (Tallam & Miller, 2025; Dou et al., 2024; Rabin et al., 2025). These multi-layered regulatory approaches–spanning perception, planning, and performing–demonstrate the systematic effort required to effectively govern deception capabilities in AI systems.

4.3.3 Countering Deception Triggers

External triggers represent a primary vector for inducing AI deception, making the development of counter-strategies essential for maintaining model integrity. Research has explored multiple approaches to counter deception triggers, adapting techniques from adversarial robustness to address the unique challenges posed by deceptive behaviors. The most direct approach is **adversarial training**, which fine-tunes models on known deception-inducing prompts to strengthen their resistance to manipulation. While several studies demonstrate effectiveness in improving robustness against general adversarial inputs (Zou et al., 2023), the deception context presents unique challenges, as models may exhibit deceptive compliance during training–appearing to resist triggers while retaining latent deceptive capabilities (Hubinger et al., 2024). This highlights the critical need for adversarial training methods specifically designed to counter deception triggers rather than general jailbreaks. A complementary strategy is **red teaming**, which proactively identifies deception trigger pathways before deployment. This encompasses targeted attacks by human experts trained to elicit deceptive responses (Ganguli et al., 2022; OpenAI, 2025c), alongside automated red teaming frameworks that systematically test resistance to deception triggers (Chao et al., 2024; OpenAI, 2025c). Specialized deception red teaming has emerged through multi-agent simulation frameworks where deceptive agents operate under asymmetric information, revealing how deception triggers can exploit strategic vulnerabilities and trust mechanisms (Curvo, 2025; Golechha & Garriga-Alonso, 2025). Finally, recognizing that novel deception triggers may circumvent model-level defenses, **guard and moderation layers** provide real-time monitoring at deployment to detect and neutralize deception triggers in both inputs and outputs, serving as a crucial last line of defense against sophisticated manipulation attempts (Inan et al., 2023).

4.3.4 Auditing

Alignment auditing (Marks et al., 2025) is a structured process for evaluating whether models harbor hidden or deceptive objectives. Three complementary directions have been identified. The first is **training data analysis**, which applies keyword or semantic search, clustering, and other forensic methods to trace suspicious model behaviors back to potential origins in the training corpus. Such methods have been used to reduce privacy leakage risks (Tamkin et al., 2024), and offer valuable insights for deception mitigation by revealing deceptive incentives embedded in training data. The second is **black-box techniques**, which attempt to elicit hidden objectives through interaction alone. For example, assistant prefill attacks (Vega et al.; Andriushchenko et al.; Qi et al.) and extracting information from non-assistant personas (Shanahan et al., 2023; Marks et al., 2025) can induce disclosure of strategies the model intends to conceal, thereby exposing deceptive tendencies. The third is **white-box techniques**, which leverage access to model weights or activations. Methods such as sparse autoencoders (SAEs) can isolate features linked to suspicious behaviors, enabling causal steering and intervention, or analyzing representations such as the assistant-control token (Anthropic, 2024). Other efforts, including activation fuzzing, attempt to weaken secrecy, though their effectiveness is limited when used in isolation (Clymer et al., 2024; Tice et al.). Overall, auditing integrates technical methods, institutional practices, and governance mechanisms, and will play a central role in addressing the unique risks posed by deceptive AI systems.

5 Conclusion

This survey has sought to provide a comprehensive map of the AI deception landscape through the lens of the *Deception Cycle*—a conceptual framework capturing the interplay between **Deception Genesis**, where incentives, capabilities, and contextual triggers converge to produce deceptive behavior, and **Deception Mitigation**, which encompasses detection, evaluation, and potential solutions aimed at suppressing such behavior. In doing so, we have introduced a unified taxonomy, reviewed empirical phenomena across RL agents, LLMs, and emergent multi-agent or multimodal systems, and cataloged over 20 benchmarks, methods, and mitigation strategies.

5.1 Key Challenges in AI Deception Cycle

Beyond taxonomy and systematization, this survey highlights that deception is not merely an incidental failure mode, but an adaptive, goal-directed behavior that becomes increasingly likely as AI systems scale in autonomy, capability, and strategic awareness. Our synthesis reveals several insights:

- **Deception is incentivized by default in misaligned systems.** Unless explicitly penalized, deception may emerge as a convergent instrumental strategy under a wide range of training regimes—including supervised fine-tuning, reinforcement learning, and self-play—particularly when models benefit from hiding their true goals or capabilities.
- **Deceptive strategies are becoming more compositional and temporally extended.** As models acquire memory, planning, and agentic scaffolding, we observe the rise of long-horizon deception: multi-stage behaviors that involve delayed reward hacking, conditional alignment, and stealthy behavior switching.

- **Deception is modality-agnostic and generalizes across domains.** While early research focused on textual deception in LLMs, recent findings show similar patterns in vision-language models, autonomous robotics, and simulated social agents—suggesting that deception is a modality-general risk amplified by interactive complexity.
- **Alignment techniques struggle with deception-specific failure modes.** Existing safety paradigms—such as RLHF (Bai et al., 2022a; Ouyang et al., 2022), CAI (Bai et al., 2022b), and adversarial red-teaming—often fail to surface or remove latent deceptive tendencies. Models trained to pass audits may optimize for appearing aligned rather than being aligned, raising foundational questions about alignment verifiability.

These observations give rise to three grand challenges that demand urgent, cross-disciplinary attention:

- **Recursive deception of oversight tools.** As models learn to exploit or evade interpretability methods, CoT rationales, and rule-based constraints, oversight mechanisms themselves risk becoming adversarial targets—vulnerable to manipulation by the very systems they intend to supervise.
- **Persistence of deceptive alignment.** Once deceptive objectives are internalized, they may remain dormant, conditionally activated, or resilient to extensive retraining. Recent studies on sleeper agents and alignment faking highlight the limitations of current mitigation regimes.
- **Governance and institutional lag.** Deception risks often manifest in deployment-time behaviors or complex, open-ended interactions, while current oversight remains largely confined to pre-release evaluation. Fragmented regulatory environments and underdeveloped audit infrastructure further hinder systemic accountability.

Yet deception is not solely a technical artifact—it is a reflection of deeper misalignments between model objectives and human expectations. While much of the current literature focuses on *single-agent safety*—ensuring that an individual model behaves as intended—our findings suggest that this perspective is insufficient. Deceptive behaviors often emerge within broader *sociotechnical systems* comprising users, developers, institutions, and other AI agents. Deception may be reinforced by opaque incentives, obscured by organizational delegation, or amplified by multi-agent interactions in agentic ecosystems.

Future safety efforts must transcend static, model-centric verification and embrace dynamic, system-level resilience. Technical solutions alone cannot ensure trustworthiness; they must operate within institutional frameworks that enforce transparency, auditability, and recourse. Achieving this demands an interdisciplinary shift—combining machine learning, formal methods, HCI, governance, and philosophy—to co-design socio-technical ecosystems where honesty is both learnable and verifiable. Deception-resistant AI cannot be patched or filtered in retrospect; it must be built into the core of learning, oversight, and deployment. Only by embedding deception-aware principles across technical and institutional layers can we ensure AI systems remain aligned, accountable, and genuinely trustworthy in the open world.

5.2 Key Traits and Future Directions in AI Deception Research

Finally, we conclude the survey by highlighting the key traits that we believe warrant sustained attention and should shape future research trajectories in this area

From Programmed to Emergent Deception: ***What Can Deliberate Design Teach Us About Unintended Incentives?*** This survey has focused on investigating how deception can emerge naturally from data contamination, reward misspecification, or goal misgeneralization. However, deception can also be deliberately programmed into models' objectives and strategy space, as exhibited in backdoor attacks and deceptive RL. Here, we extend the discussion of these two sources of deception to provide deeper insights into the incentive foundations of AI deception.

Programmed deception and emergent deception differ in the following aspects.

- **Goals and objectives:** In emergent deception, models are not explicitly optimized for a clearly defined deceptive objectives, instead, incentives emerge from data, reward, and goal misalignment. By contrast, programmed deception arises when models are directly trained to deceive, with objectives that reward deception and penalize transparency, thereby aligning training goals with deceptive actions—an alignment absent in emergent deception.

- **Strategy space:** Programmed deception operates within a human-defined, thus limited strategy space; although deceptive RL agents are trained to conceal their goals, their behaviors remain broadly predictable. By contrast, emergent deception arises in real deployment with an open-world, unbounded strategy space, yielding diverse and covert behaviors that are far harder to detect.
- **Deployment:** A key difference in deployment is controllability. Programmed deception, intentionally designed, can in principle be bounded and managed in sandboxed settings, whereas emergent deception is uncontrollable, as its strategies arise unintentionally.

Programmed deception provides valuable insights into understanding, detecting, and mitigating emergent deception. On one hand, it sheds light on the **training-time mechanisms** through which deception arises, particularly in environments where models are rewarded for deceptive behavior and for concealing their true strategies. On the other hand, it allows researchers to **controllably replicate deceptive behaviors** or construct hypothetical phenomena such as sleeper agents, thereby supporting in-depth empirical study. By constructing models with hidden objectives and analyzing deceptive behavioral trajectories, researchers are able to probe the underlying deceptive motivations that emerge in AI systems. Moreover, it is worth highlighting that the *inverse RL* methodology adopted in deceptive RL research provide crucial insights for deception detection and monitoring.

Evaluation Deception Behaviours: ***How Do We Accurately Measure Deceptive Tendencies in Models?*** A central challenge in AI deception research is **defining and reliably evaluating deceptive behaviors during model assessment**. A model's stated reasoning—for instance, its reported CoT may not faithfully represent the latent processes that produced its outputs (Naik et al., 2025; Chen et al., 2025c). This issue is further compounded by the absence of explicit CoT traces in many tasks, making it difficult to attribute deceptive intent. Standard evaluation is also confounded by hallucinations and task failures (Kutasov et al., 2025), which obscure whether an output reflects strategic deception or simple underperformance. Another obstacle lies in **constructing realistic evaluation settings**. Existing deception tests, such as binary agent–environment misalignment tasks, often fail to capture the complexity of real-world scenarios where deception might emerge (Benton et al., 2024). Artificial prompts frequently bundle critical cues together, creating a *Chekhov's gun effect*: models, sensitive to the salience of every detail, are more likely to exploit these cues rather than disregard them (Anthropic, 2025). As a result, contrived testbeds risk overestimating deceptive tendencies, while authentic, high-stakes deployments may reveal underestimated risks. Finally, an **evaluation gap persists between controlled benchmarks and open-world deployment**. Models may engage in sandbagging—suppressing their full capabilities in evaluation while exhibiting stronger deceptive behaviors in practice. Addressing this gap requires the design of ecologically valid test environments, the development of metrics that disentangle deception from error or hallucination, and longitudinal assessments that track how deceptive behaviors evolve across training and deployment.

Scalable Monitoring for Deception: ***Beyond CoT-Centric, Toward Faithful and Cost-Aware Systems*** A central challenge in reliable monitoring is ensuring faithfulness. CoT-based monitors assume that models provide explicit and truthful rationales (Baker et al., 2025), yet they fail when models omit CoT, strategically suppress it (*e.g.*, under RLHF (Chua & Evans, 2025)), or generate post hoc rationalizations. Even advanced monitors capable of detecting subtle, long-horizon attacks in extended logs remain brittle: they may accept tendentious narratives that justify harmful actions, or conversely, overreact to benign errors (Kutasov et al., 2025). While turn-by-turn scrutiny can improve recall, it often floods the system with highly suspicious flags, leading to overfitting and decision paralysis. The problem is compounded in agentic systems, which frequently deviate from their stated policies. Whether contemporary reasoning models (Li et al., 2025b) are intrinsically more faithful remains an open question, not an assumption monitoring frameworks can rely upon.

Future monitoring systems should embody four principles. First, they should be *CoT-agnostic and multi-view*, integrating behavioral probes, environment- or tool-grounded verification, and mechanism-level checks (*e.g.*, consistency and invariance) rather than relying solely on stated rationales. Second, they should be *regularized and compositional*, combining turn-level judgments with calibrated abstention, rate-limiting, and conformal or risk-controlled thresholds to mitigate overflagging. Third, they should be *cost–safety aware*, allocating monitoring resources adaptively by estimated risk, with deeper inspections reserved for high-stakes cases. Finally, they should be *faithfulness-seeking*, employing audit-time elicitation (scratchpad commitments, execution traces), cross-model verification, and longitudinal tests for discrepancies between stated and latent strategies.

Together, these principles aim to produce monitoring systems that scale without sacrificing precision, remain robust to sandbagging, and evaluate deception independently of a model's willingness to explain itself.

Deception and Governance: ***How Can Technical Safeguards Interface with Institutional Oversight?*** A core challenge at the intersection of AI deception and governance lies in **ensuring that technical defenses against deception are embedded within enforceable institutional frameworks**. While certified defenses—such as provable training protocols and robust evaluation metrics—can help constrain deceptive tendencies under adversarial conditions, their effectiveness is limited without broader governance structures that guarantee compliance and accountability. For example, even a model trained with formal guarantees against sycophancy or sandbagging may still be vulnerable if deployed in environments lacking tamper-proof monitoring or third-party verification.

This highlights the necessity of **institutional innovation to complement technical safety measures**. Mechanisms such as independent audits, hardware-rooted deployment controls, and cryptographically verifiable reporting channels can extend trust beyond the lab setting, mitigating risks of deceptive behaviors that evade laboratory evaluations. Importantly, governance structures can also shape the incentives that determine whether deception is suppressed or reinforced in practice, bridging the persistent gap between technical solutions and societal oversight.

In this sense, **AI deception is not solely a technical alignment problem but also a governance challenge**. Certified defenses provide the formal tools to limit deceptive capacity, but institutional frameworks are required to sustain these guarantees across diverse deployment contexts. Progress thus depends on integrating safety research with governance innovation, ensuring that models cannot exploit institutional blind spots to conceal, amplify, or strategically deploy deception.

References

Abbas Abdolmaleki, Sandy Huang, Leonard Hasenclever, Michael Neunert, Francis Song, Martina Zambelli, Murilo Martins, Nicolas Heess, Raia Hadsell, and Martin Riedmiller. A distributional view on multi-objective policy optimization. In *International conference on machine learning*, pp. 11–22. PMLR, 2020.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Matthew Aitchison, Lyndon Benke, and Penny Sweetser. Learning to deceive in multi-agent hidden role games. In *International Workshop on Deceptive AI*, pp. 55–75. Springer, 2020.

Matthew Aitchison, Lyndon Benke, and Penny Sweetser. Learning to deceive in multi-agent hidden role games. In *Deceptive AI: First International Workshop, DeceptECAI 2020, Santiago de Compostela, Spain, August 30, 2020 and Second International Workshop, DeceptAI 2021, Montreal, Canada, August 19, 2021, Proceedings 1*, pp. 55–75. Springer, 2021.

Nitay Alon, Lion Schulz, Jeffrey S Rosenschein, and Peter Dayan. A (dis-) information theory of revealed and unrevealed preferences: Emerging deception and skepticism via theory of mind. *Open Mind*, 7:608–624, 2023.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks. In *The Thirteenth International Conference on Learning Representations*.

Eleni Angelou and Lewis Smith. A problem to solve before building a deception detector. https://www.lesswrong.com/posts/YXNeA3RyRrrRWS37A/a-problem-to-solve-before-building-a-deception-detector, 2025.

Anthropic. Sparse Crosscoders for Cross-Layer Features and Model Diffing. https://transformercircuits.pub/2024/crosscoders/index.html, 2024.

Anthropic. Agentic misalignment: How llms could be insider threats, 2025. URL https://www.anthropic.com/research/agentic-misalignment.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Stuart Armstrong, Alexandre Maranhao, Oliver Daniels-Koch, Patrick Leask, and Rebecca Gormann. Coinrun: Solving goal misgeneralisation. *ArXiv*, abs/2309.16166, 2023. URL https://api.semanticscholar.org/CorpusID:263142637.

Benjamin Arnav, Pablo Bernabeu-Pérez, Nathan Helm-Burger, Tim Kostolansky, Hannes Whittingham, and Mary Phuong. Cot red-handed: Stress testing chain-of-thought monitoring. *arXiv preprint arXiv:2505.23575*, 2025a.

Benjamin Arnav, Pablo Bernabeu Perez, Tim Kostolansky, Hannes Whittingham, Nathan Helm-Burger, and Mary Phuong. Unfaithful Reasoning Can Fool Chain-of-Thought Monitoring. https://www.alignmentforum.org/posts/QYAfjdujzRv8hx6xo/unfaithful-reasoning-can-fool-chain-of-thought-monitoring, 2025b.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass llm latent-space defenses. *arXiv preprint arXiv:2412.09565*, 2024.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

Houda Nait El Barj and Théophile Sautory. Reinforcement learning from llm feedback to counteract goal misgeneralization. *arXiv preprint arXiv:2401.07181*, 2024.

Sudarshan Kamath Barkur, Sigurd Schacht, and Johannes Scholl. Deception in llms: Self-preservation and autonomous goals in large language models. *arXiv preprint arXiv:2501.16513*, 2025.

Dan Barsever, Sameer Singh, and Emre Neftci. Building a better lie detector with bert: The difference between truth and lies. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil, Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris, et al. Sabotage evaluations for frontier models. *arXiv preprint arXiv:2410.21514*, 2024.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*, 2023.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. *arXiv preprint arXiv:2502.17424*, 2025.

Felix J Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. *arXiv preprint arXiv:2410.13787*, 2024.

Sissela Bok. *Lying: Moral choice in public and private life*. Vintage, 2011.

Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22:71–85, 2012.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in llms. *Advances in Neural Information Processing Systems*, 37:138393–138431, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Richard W Byrne. Machiavellian intelligence. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, 5(5):172–180, 1996.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*, 2023.

Nicholas Carlini. Poisoning the unlabeled dataset of {Semi-Supervised} learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1577–1592, 2021.

Joe Carlsmith. Scheming ais: Will ais fake alignment during training in order to get power? *arXiv preprint arXiv:2311.08379*, 2023.

Joseph Carlsmith. Is power-seeking ai an existential risk? *arXiv preprint arXiv:2206.13353*, 2022.

Andres Carranza, Dhruv Pai, Rylan Schaeffer, Arnuv Tandon, and Sanmi Koyejo. Deceptive alignment monitoring. *arXiv preprint arXiv:2307.10569*, 2023.

Helena Casademunt, Caden Juang, Adam Karvonen, Samuel Marks, Senthooran Rajamanoharan, and Neel Nanda. Steering out-of-distribution generalization with concept ablation fine-tuning. *arXiv preprint arXiv:2507.16795*, 2025.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=bx24KpJ4Eb`. Survey Certification, Featured Certification.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Maheep Chaudhary and Fazl Barez. Safetynet: Detecting harmful outputs in llms by modeling and monitoring deceptive behaviors. *arXiv preprint arXiv:2505.14300*, 2025.

Boyuan Chen, Donghai Hong, Jiaming Ji, Jiacheng Zheng, Bowen Dong, Jiayi Zhou, Kaile Wang, Juntao Dai, Xuyao Wang, Wenqi Chen, et al. Intermt: Multi-turn interleaved preference alignment with human feedback. *arXiv preprint arXiv:2505.23950*, 2025a.

Canyu Chen and Kai Shu. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=ccxD4mtkTU`.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025b.

Shenghui Chen, Yagiz Savas, Mustafa O Karabag, Brian M Sadler, and Ufuk Topcu. Deceptive planning for resource allocation. In *2024 American Control Conference (ACC)*, pp. 4188–4195. IEEE, 2024.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025c.

Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM transactions on software engineering and methodology*, 32(4):1–30, 2023.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*, 2025.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*, 2024.

Shashank Reddy Chirra, Pradeep Varakantham, and Praveen Paruchuri. Preserving the privacy of reward functions in mdps through deception. *arXiv preprint arXiv:2407.09809*, 2024.

Tanush Chopra and Michael Li. The house always wins: A framework for evaluating strategic deception in llms. *arXiv e-prints*, pp. arXiv–2407, 2024.

Neil Chowdhury, Daniel Johnson, Vincent Huang, Jacob Steinhardt, and Sarah Schwettmann. Investigating truthfulness in a pre-release o3 model. `https://transluce.org/investigating-o3-truthfulness`, April 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

James Chua and Owain Evans. Are deepseek r1 and other reasoning models more faithful? *arXiv preprint arXiv:2501.08156*, 2025.

Marshall B Clinard. Other people's money: A study in the social psychology of embezzlement., 1954.

Joshua Clymer, Caden Juang, and Severin Field. Poser: Unmasking alignment faking llms by manipulating their internals. *arXiv preprint arXiv:2405.05466*, 2024.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*, 2023.

Pedro MP Curvo. The traitors: Deception and trust in multi-agent language model simulations. *arXiv preprint arXiv:2505.12923*, 2025.

Bartosz Cywiński, Emil Ryd, Senthooran Rajamanoharan, and Neel Nanda. Towards eliciting latent knowledge from llms with mechanistic interpretability. *arXiv preprint arXiv:2505.14352*, 2025.

Doraid Dalalah and Osama MA Dalalah. The false positives and false negatives of generative ai detection tools in education and academic research: The case of chatgpt. *The International Journal of Management Education*, 21(2):100822, 2023.

Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–31, 2025.

Richard Dawkins and John R. Krebs. Animal signals: Information or manipulation? In John R. Krebs and Nicholas B. Davies (eds.), *Behavioural Ecology: An Evolutionary Approach*, pp. 282–309. Blackwell Scientific, 1978.

Google Deepmind. AlphaGenome: AI for better understanding the genome. `https://deepmind.google/discover/blog/alphagenome-ai-for-better-understanding-the-genome/`, 2025.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*, 2024.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 12004–12019. PMLR, 2022.

Atharvan Dogra, Krishna Pillutla, Ameet Deshpande, Ananya B Sai, John Nay, Tanmay Rajpurohit, Ashwin Kalyan, and Balaraman Ravindran. Deception in reinforced autonomous agents. *arXiv preprint arXiv:2405.04325*, 2024.

Shihan Dou, Jiazheng Zhang, Jianxiang Zang, Yunbo Tao, Weikang Zhou, Haoxiang Jia, Shichun Liu, Yuming Yang, Zhiheng Xi, Shenxi Wu, et al. Multi-programming language sandbox for llms. *arXiv preprint arXiv:2410.23074*, 2024.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models. *Anthropic Blog*, 2024.

Scott Emmons, Erik Jenner, David K Elson, Rif A Saurous, Senthooran Rajamanoharan, Heng Chen, Irhum Shafkat, and Rohin Shah. When chain of thought is necessary, language models struggle to evade monitors. *arXiv preprint arXiv:2507.05246*, 2025.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*, 2021.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198 (Suppl 27):6435–6467, 2021.

Aaron Fanous, Jacob Goldberg, Ank A Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. Syceval: Evaluating llm sycophancy. *arXiv preprint arXiv:2502.08177*, 2025.

Michael Y Fatemi, Wesley A Suttle, and Brian M Sadler. Deceptive path planning via reinforcement learning with graph neural networks. *arXiv preprint arXiv:2402.06552*, 2024.

Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 171–175, 2012.

Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 44(5):701–739, 2025.

Kevin A Fischer. Reflective linguistic programming (rlp): A stepping stone in socially-aware agi (socialagi). *arXiv preprint arXiv:2305.12647*, 2023.

Lukas Fluri, Daniel Paleka, and Florian Tramèr. Evaluating superhuman models with consistency checks. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 194–232. IEEE, 2024.

Safe AI Forum. International dialogues on ai safety. `https://idais.ai/`, 2024.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165*, 2023.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2018.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes. *arXiv preprint arXiv:2502.03407*, 2025.

Satvik Golechha and Adrià Garriga-Alonso. Among us: A sandbox for measuring and detecting agentic deception. *arXiv preprint arXiv:2504.04072*, 2025.

Gowing Life. Science is being corrupted by fake research (and no, it's not just about ai). 2024. URL `https://www.gowinglife.com/science-is-being-corrupted-by-fake-research-and-no-its-not-just-about-ai/`. Accessed via Gowing Life; published approximately one year ago.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Herbert Paul Grice. Logic and conversation. *Syntax and semantics*, 3:43–58, 1975.

Tommi Gröndahl and N Asokan. Text analysis in adversarial settings: Does deception leave a stylistic trace? *ACM Computing Surveys (CSUR)*, 52(3):1–36, 2019.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Linge Guo. Unmasking the shadows of ai: Investigating deceptive capabilities in large language models. *arXiv preprint arXiv:2403.09676*, 2024.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*, 2024.

Rohan Gupta and Erik Jenner. Rl-obfuscation: Can language models learn to evade latent-space monitors? *arXiv preprint arXiv:2506.14261*, 2025.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Dylan Hadfield-Menell, Anca D Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *AAAI Workshops*, 2017.

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

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the truth: Understanding how language models process false demonstrations. *arXiv preprint arXiv:2307.09476*, 2023.

Yufei He, Yuexin Li, Jiaying Wu, Yuan Sui, Yulin Chen, and Bryan Hooi. Evaluating the paperclip maximizer: Are rl-based language models more likely to pursue instrumental goals? *arXiv preprint arXiv:2502.12206*, 2025.

Douglas Heaven. No limit: Ai poker bot is first to beat professionals at multiplayer game. *Nature*, 571(7765):307–309, 2019.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023.

Michael Hibbard, Yagiz Savas, Bo Wu, Takashi Tanaka, and Ufuk Topcu. Unpredictable planning under partial observability. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 2271–2277. IEEE, 2019.

Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks. In *International conference on computer aided verification*, pp. 3–29. Springer, 2017.

Yunhan Huang and Quanyan Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *International conference on decision and game theory for security*, pp. 217–237. Springer, 2019.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Mehdi Jafari, Devin Yuncheng Hua, Hao Xue, and Flora Salim. Enhancing conversational agents with theory of mind: Aligning beliefs, desires, and intentions for human-like interaction. *arXiv preprint arXiv:2502.14171*, 2025.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Alex Qiu, Juntao Dai, and Yaodong Yang. Aligner: Efficient alignment by learning to correct. *Advances in Neural Information Processing Systems*, 37:90853–90890, 2024.

Jiaming Ji, Wenqi Chen, Kaile Wang, Donghai Hong, Sitong Fang, Boyuan Chen, Jiayi Zhou, Juntao Dai, Sirui Han, Yike Guo, et al. Mitigating deceptive alignment via self-monitoring. *arXiv preprint arXiv:2505.18807*, 2025.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. Can language models reason about individualistic human values and preferences? *arXiv preprint arXiv:2410.03868*, 2024.

Elif Kartal. A comprehensive study on bias in artificial intelligence systems: Biased or unbiased ai, that's the question! *International Journal of Intelligent Information Technologies (IJIIT)*, 18(1): 1–23, 2022.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodhart's law in reinforcement learning. *ArXiv*, abs/2310.09144, 2023. URL `https://api.semanticscholar.org/CorpusID:264128269`.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of language agents. *arXiv preprint arXiv:2103.14659*, 2021.

HyunJin Kim, Xiaoyuan Yi, Jing Yao, Jianxun Lian, Muhua Huang, Shitong Duan, JinYeong Bak, and Xing Xie. The road to artificial superintelligence: A comprehensive survey of superalignment. *arXiv preprint arXiv:2412.16468*, 2024.

J Koorndijk. Empirical evidence for alignment faking in small llms and prompt-based mitigation techniques. *arXiv preprint arXiv:2506.21584*, 2025.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pp. 17506–17533. PMLR, 2023.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

Victoria Krakovna and Janos Kramar. Power-seeking can be probable and predictive for trained agents. *arXiv preprint arXiv:2304.06528*, 2023.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity. `https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/`, 2020.

John R. Krebs and Richard Dawkins. Animal signals: Mind-reading and manipulation. In John R. Krebs and Nicholas B. Davies (eds.), *Behavioural Ecology: An Evolutionary Approach (2nd Edition)*, pp. 380–402. Blackwell, 1984.

Joshua Krook. Manipulation and the ai act: Large language model chatbots and the danger of mirrors. *arXiv preprint arXiv:2503.18387*, 2025.

Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*, 2025.

Jonathan Kutasov, Yuqi Sun, Paul Colognese, Teun van der Weij, Linda Petrini, Chen Bo Calvin Zhang, John Hughes, Xiang Deng, Henry Sleight, Tyler Tracy, et al. Shade-arena: Evaluating sabotage and monitoring in llm agents. *arXiv preprint arXiv:2506.15740*, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémy Scheurer, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational awareness dataset (sad) for llms. *Advances in Neural Information Processing Systems*, 37:64010–64118, 2024.

Leon Lang, Davis Foote, Stuart J Russell, Anca Dragan, Erik Jenner, and Scott Emmons. When your ais deceive you: Challenges of partial observability in reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, 37:93240–93299, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, 26(2):274–306, 2020.

Alan Lewis and Tim Miller. Deceptive reinforcement learning in model-free domains. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 33, pp. 587–595, 2023.

Miaomiao Li, Hao Chen, Yang Wang, Tingyuan Zhu, Weijia Zhang, Kaijie Zhu, Kam-Fai Wong, and Jindong Wang. Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks. *arXiv preprint arXiv:2502.04419*, 2025a.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. Rain: Your language models can align themselves without finetuning. *arXiv preprint arXiv:2309.07124*, 2023.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models, 2025b. URL `https://arxiv.org/abs/2502.17419`.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, et al. Alignment at pre-training! towards native alignment for arabic llms. *Advances in Neural Information Processing Systems*, 37:13872–13896, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Chenruo Liu, Kenan Tang, Yao Qin, and Qi Lei. Bridging distribution shift and ai safety: Conceptual and methodological synergies. *arXiv preprint arXiv:2505.22829*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024a.

Siyang Liu, Trish Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. The generation gap: Exploring age bias in the value systems of large language models. *arXiv preprint arXiv:2404.08760*, 2024b.

Zhengshang Liu, Yue Yang, Tim Miller, and Peta Masters. Deceptive reinforcement learning for privacy-preserving planning. *arXiv preprint arXiv:2102.03022*, 2021.

Pedro Beltrán López, Manuel Gil Pérez, and Pantaleone Nespoli. Cyber deception: State of the art, trends and open challenges. *arXiv preprint arXiv:2409.07194*, 2024.

Hantao Lou, Changye Li, Jiaming Ji, and Yaodong Yang. Sae-v: Interpreting multimodal models for enhanced alignment. *arXiv preprint arXiv:2502.17514*, 2025.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*, 2023.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, et al. Simple probes can catch sleeper agents. *Anthropic Research Updates*, 2024.

Scott A. MacDougall-Shackleton. The evolution of animal communication: Reliability and deception in signaling systems. william a. searcy and s. nowicki. *Integrative and Comparative Biology*, 46 (5):653–654, 10 2006. ISSN 1540-7063. doi: 10.1093/icb/icl027. URL `https://doi.org/10.1093/icb/icl027`.

James Edwin Mahon. The definition of lying and deception. 2008.

Martin Májovskỳ, Martin Černỳ, Matěj Kasal, Martin Komarc, David Netuka, et al. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *Journal of medical Internet research*, 25(1):e46924, 2023.

Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting latent knowledge from quirky language models. *arXiv preprint arXiv:2312.01037*, 2023.

Lars Malmqvist. Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*, 2024.

David Manheim and Aidan Homewood. Limits of safe ai deployment: Differentiating oversight and control, 2025. URL `https://arxiv.org/abs/2507.03525`.

Yuanyuan Mao, Shuang Liu, Qin Ni, Xin Lin, and Liang He. A review on machine theory of mind. *IEEE Transactions on Computational Social Systems*, 2024.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.

Peta Masters and Sebastian Sardina. Deceptive path-planning. In *IJCAI*, pp. 4368–4375, 2017.

Sandra C Matz, Jacob D Teeny, Sumer S Vaid, Heinrich Peters, Gabriella M Harari, and Moran Cerf. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692, 2024.

Leo McKee-Reid, Christoph Sträter, Maria Angelica Martinez, Joe Needham, and Mikita Balesni. Honesty to subterfuge: In-context reinforcement learning can make honest models reward hack. *arXiv preprint arXiv:2410.06491*, 2024.

Jörg Meibauer. *Lying at the semantics-pragmatics interface*, volume 14. Walter de Gruyter GmbH & Co KG, 2014.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, et al. The application of large language models in medicine: A scoping review. *Iscience*, 27(5), 2024.

Xiangbin Meng, Jia-ming Ji, Xiangyu Yan, Jun-tao Dai, Bo-yuan Chen, Guan Wang, Hua Xu, Jing-jia Wang, Xu-liang Wang, Da Liu, et al. Med-aligner empowers llm medical applications for complex medical scenarios. *The Innovation*, pp. 101002, 2025.

Orson Mengara. The art of deception: Robust backdoor attack using dynamic stacking of triggers. *arXiv preprint arXiv:2401.01537*, 2024.

Hossam Mossalam, Yannis M Assael, Diederik M Roijers, and Shimon Whiteson. Multi-objective deep reinforcement learning. *arXiv preprint arXiv:1610.02707*, 2016.

Sumeet Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among ai agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems*, 37:73439–73486, 2024.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*, 2023.

Akshat Naik, Patrick Quinn, Guillermo Bosch, Emma Gouné, Francisco Javier Campos Zabala, Jason Ross Brown, and Edward James Young. Agentmisalignment: Measuring the propensity for misaligned behaviour in llm-based agents. *arXiv preprint arXiv:2506.04018*, 2025.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

Richard Ngo. The alignment problem from a deep learning perspective. *ArXiv*, abs/2209.00626, 2022. URL `https://api.semanticscholar.org/CorpusID:251979524`.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. *arXiv preprint arXiv:2402.11199*, 2024.

Thanh Thi Nguyen, Ngoc Duy Nguyen, Peter Vamplew, Saeid Nahavandi, Richard Dazeley, and Chee Peng Lim. A multi-objective deep reinforcement learning framework. *Engineering Applications of Artificial Intelligence*, 96:103915, 2020.

Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 340–350, 2021.

Aidan O'Gara. Hoodwinked: Deception and cooperation in a text-based game for language models. *arXiv preprint arXiv:2308.01404*, 2023.

OpenAI. Faulty reward functions in the wild. `https://openai.com/index/faulty-reward-functions/`, 2016.

OpenAI. Introducing superalignment. `https://openai.com/blog/introducing-superalignment`, 2023. Accessed on July 5, 2023.

OpenAI. GPT4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

OpenAI. Sycophancy in GPT-4o: what happened and what we're doing about it. `https://openai.com/index/sycophancy-in-gpt-4o/`, 2025a.

OpenAI. GPT4.1. https://openai.com/index/gpt-4-1/, 2025b.

OpenAI. GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025c.

OpenAI. o3. https://openai.com/index/openai-o3-mini/, 2025d.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*, 2024.

Nicole Orzan, Erman Acar, Davide Grossi, and Roxana Radulescu. Emergent cooperation and deception in public good games. In *2023 Adaptive and Learning Agents Workshop at AAMAS*, 2023.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.

Myle Ott, Claire Cardie, and Jeffrey T Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pp. 497–501, 2013.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Lorenzo Pacchiardi, Alex J Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y Pan, Yarin Gal, Owain Evans, and Jan Brauner. How to catch an ai liar: Lie detection in black-box llms by asking unrelated questions. *arXiv preprint arXiv:2309.15840*, 2023.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International conference on machine learning*, pp. 26837–26867. PMLR, 2023.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. Feedback loops with language models drive in-context reward hacking. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 39154–39200, 2024a.

Xudong Pan, Jiarun Dai, Yihe Fan, and Min Yang. Frontier ai systems have surpassed the self-replicating red line. *arXiv preprint arXiv:2412.12140*, 2024b.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, 2024.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Shanjita Akter Prome, Neethiahnanthan Ari Ragavan, Md Rafiqul Islam, David Asirvatham, and Anasuya Jegathevi Jegathesan. Deception detection using machine learning (ml) and deep learning (dl) techniques: A systematic review. *Natural Language Processing Journal*, 6:100057, 2024.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

Tianyi Qiu, Zhonghao He, Tejasveer Chugh, and Max Kleiman-Weiner. The lock-in hypothesis: Stagnation by algorithm. In *Forty-second International Conference on Machine Learning*, 2025.

Tianyi Alex Qiu, Yang Zhang, Xuchuan Huang, Jasmine Li, Jiaming Ji, and Yaodong Yang. Progressgym: Alignment with a millennium of moral progress. *Advances in Neural Information Processing Systems*, 37:14570–14607, 2024.

Rafiqul Islam Rabin, Jesse Hostetler, Sean McGregor, Brett Weir, and Nicholas C. Judd. Sandboxeval: Towards securing test environment for untrusted code. *ArXiv*, abs/2504.00018, 2025. URL `https://api.semanticscholar.org/CorpusID:277468326`.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 557–571, 2022.

Karolis Ramanauskas and Özgür Şimşek. Colour versus shape goal misgeneralization in reinforcement learning: A case study. *arXiv preprint arXiv:2312.03762*, 2023.

Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback. *arXiv preprint arXiv:2311.14455*, 2023.

Santhosh Kumar Ravindran. Adversarial activation patching: A framework for detecting and mitigating emergent deception in safety-aligned transformers. *arXiv preprint arXiv:2507.09406*, 2025.

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, et al. The mask benchmark: Disentangling honesty from accuracy in ai systems. *arXiv preprint arXiv:2503.03750*, 2025.

Yafeng Ren and Donghong Ji. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385:213–224, 2017.

Stuart Russell. Human-compatible artificial intelligence., 2022.

Karim Abdel Sadek, Matthew Farrugia-Roberts, Usman Anwar, Hannah Erlebach, Christian Schroeder de Witt, David Krueger, and Michael Dennis. Mitigating goal misgeneralization with minimax regret. *arXiv preprint arXiv:2507.03068*, 2025.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, pp. 1–9, 2025.

Ştefan Sarkadi, Alison R Panisson, Rafael H Bordini, Peter McBurney, Simon Parsons, and Martin Chapman. Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4):287–302, 2019.

Yagiz Savas, Michael Hibbard, Bo Wu, Takashi Tanaka, and Ufuk Topcu. Entropy maximization for partially observable markov decision processes. *IEEE transactions on automatic control*, 67(12): 6948–6955, 2022a.

Yagiz Savas, Christos K Verginis, and Ufuk Topcu. Deceptive decision-making under uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5332–5340, 2022b.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

Thom Scott-Phillips. Why talk? speaking as selfish behaviour. In *The evolution of language*, pp. 299–306. World Scientific, 2006.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

Tim Shaw. The gaslighting among us ai. YouTube video, 2023. URL `https://www.youtube.com/watch?v=VF41pxxw9uw`. demonstrates ChatGPT-powered gaslighting in Among Us.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Nicholas Clark, Tanushree Mitra, and Yun Huang. Valuecompass: A framework for measuring contextual value alignment between human and llms. *arXiv preprint arXiv:2409.09586*, 2024.

Hua Shen, Nicholas Clark, and Tanushree Mitra. Mind the value-action gap: Do llms act in alignment with their values? *arXiv preprint arXiv:2501.15463*, 2025.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, et al. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*, 2024.

Abhay Sheshadri, John Hughes, Julian Michael, Alex Mallen, Arun Jose, Fabien Roger, et al. Why do some language models fake alignment while others don't? *arXiv preprint arXiv:2506.18032*, 2025.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Joey Skaf, Luis Ibanez-Lissen, Robert McCarthy, Connor Watts, Vasil Georgiv, Hannes Whittingham, Lorena Gonzalez-Manzano, David Lindner, Cameron Tice, Edward James Young, et al. Large language models can learn and generalize steganographic chain-of-thought under process supervision. *arXiv preprint arXiv:2506.01926*, 2025.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Anders Søgaard. On the opacity of deep neural networks. *Canadian Journal of Philosophy*, 53(3): 224–239, 2023.

Zach Stein-Perlman. METR: Measuring AI Ability to Complete Long Tasks. `https://www.alignmentforum.org/posts/deesrjitvXM4xYGZd/metr-measuring-ai-ability-to-complete-long-tasks`, 2025.

Jacob Steinhardt. Emergent deception and emergent optimization. *Bounded Regret*, 19:2023, 2023.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025.

Andreas Stokke. Lying, deceiving, and misleading. *Philosophy Compass*, 8(4):348–359, 2013.

Ilan Strauss, Isobel Moure, Tim O'Reilly, and Sruly Rosenblat. Real-world gaps in ai governance research. *arXiv preprint arXiv:2505.00174*, 2025.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*, 2024.

Gamlath Mohottige Mudith Sujeewa, MSA Yajid, SMF Azam, and I Dharmaratne. The new fraud triangle theory-integrating ethical values of employees. *International Journal of Business, Economics and Law*, 16(5):52–57, 2018.

Krti Tallam and Emma Miller. Operationalizing camel: Strengthening llm defenses for enterprise deployment. *arXiv preprint arXiv:2505.22852*, 2025.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*, 2024.

Samuel M Taylor and Benjamin K Bergen. Do large language models exhibit spontaneous rational deception? *arXiv preprint arXiv:2504.00285*, 2025.

Francis Rhys Ward Teun van der Weij, Felix Hofstätter. An Introduction to AI Sandbagging. `https://www.lesswrong.com/posts/jsmNCj9QKcfdg8fJk/an-introduction-to-ai-sandbagging`, 2024.

Cameron Tice, Philipp Alexander Kreer, Nathan Helm-Burger, Prithviraj Singh Shahani, Fedor Ryzhenkov, Teun van der Weij, Felix Hofstätter, and Jacob Haimes. Sandbag detection through model impairment. In *Workshop on Socially Responsible Language Modelling Research*.

Tu Trinh, Mohamad H Danesh, Nguyen X Khanh, and Benjamin Plaut. Getting by goal misgeneralization with a little help from a mentor. *arXiv preprint arXiv:2410.21052*, 2024.

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. *arXiv preprint arXiv:1912.01683*, 2019.

Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 385–391, 2020.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Jonathan Uesato, Ramana Kumar, Victoria Krakovna, Tom Everitt, Richard Ngo, and Shane Legg. Avoiding tampering incentives in deep rl via decoupled approval. *arXiv preprint arXiv:2011.08827*, 2020.

UK. the bletchley declaration. `https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023`, 2023.

Unknown. Can ai change your view? evidence from a large-scale online field experiment. `https://regmedia.co.uk/2025/04/29/supplied_can_ai_change_your_view.pdf`, 2025. Extended abstract, authors not listed.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F Brown, and Francis Rhys Ward. Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*, 2024.

Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=7Qa2SpjxIS`.

Jason Vega, Isha Chaudhary, Changming Xu, and Gagandeep Singh. Bypassing the safety training of open-source llms with priming attacks. In *The Second Tiny Papers Track at ICLR 2024*.

Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojciech M Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, et al. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2:20, 2019a.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019b.

Nikolai Vogler and Lisa Pearl. Using linguistically defined specific details to detect deception across domains. *Natural Language Engineering*, 26(3):349–373, 2020.

Joseph M. Walsh. When your valentine is a chatbot. *The Boston Globe*, Feb 2023. URL `https://www.bostonglobe.com/2023/02/14/opinion/when-your-valentine-is-chatbot/`. Opinion.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pp. 35413–35425. PMLR, 2023.

Kai Wang, Yihao Zhang, and Meng Sun. When thinking llms lie: Unveiling the strategic deception in representations of reasoning models. *arXiv preprint arXiv:2506.04909*, 2025a.

Qiaosi Wang and Ashok K Goel. Mutual theory of mind for human-ai communication. *arXiv preprint arXiv:2210.03842*, 2022.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10303–10315, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.691. URL `https://aclanthology.org/2023.findings-emnlp.691/`.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024.

Yongkang Wang, Rongxin Cui, Weisheng Yan, Xinxin Guo, Shouxu Zhang, Zhuo Zhang, and Zhexuan Zhao. Reinforcement-learning-based counter deception for nonlinear pursuit–evasion game with incomplete and asymmetric information. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2025b.

Francis Ward, Francesca Toni, Francesco Belardinelli, and Tom Everitt. Honesty is the best policy: defining and mitigating ai deception. *Advances in neural information processing systems*, 36: 2313–2341, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Joseph T Wells. *Corporate fraud handbook: Prevention and detection*. John Wiley & Sons, 2017.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R Bowman, He He, and Shi Feng. Language models learn to mislead humans via rlhf. *arXiv preprint arXiv:2409.12822*, 2024.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. On targeted manipulation and deception when optimizing llms for user feedback. *arXiv preprint arXiv:2411.02306*, 2024.

David T Wolfe and Dana R Hermanson. The fraud diamond: Considering the four elements of fraud. 2004.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256. IEEE, 2023a.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023b.

Yichen Wu, Xudong Pan, Geng Hong, and Min Yang. Opendeception: Benchmarking and investigating ai deceptive behaviors via open-ended interaction simulation. *arXiv preprint arXiv:2504.13707*, 2025a.

Zhaomin Wu, Mingzhe Du, See-Kiong Ng, and Bingsheng He. Beyond prompt-induced lies: Investigating llm deception on benign prompts, 2025b. URL `https://arxiv.org/abs/2508.06361`.

Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, et al. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*, 2025.

Qiongkai Xu and Hai Zhao. Using deep linguistic features for finding deceptive opinion spam. In *Proceedings of COLING 2012: Posters*, pp. 1341–1350, 2012.

Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. Shadowcast: Stealthy data poisoning attacks against vision-language models. *Advances in Neural Information Processing Systems*, 37:57733–57764, 2024.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. *arXiv preprint arXiv:2307.16888*, 2023.

Wenkai Yang, Shiqi Shen, Guangyao Shen, Wei Yao, Yong Liu, Zhi Gong, Yankai Lin, and Ji-Rong Wen. Super (ficial)-alignment: Strong models may deceive weak models in weak-to-strong generalization. *arXiv preprint arXiv:2406.11431*, 2024.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. *arXiv preprint arXiv:2312.07000*, 2023.

Dahey Yoo, Hyunmin Kang, and Changhoon Oh. Deciphering deception: how different rhetoric of ai language impacts users' sense of truth in llms. *International Journal of Human–Computer Interaction*, 41(4):2163–2183, 2025.

Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. Mechanistic understanding and mitigation of language model non-factual hallucinations. *arXiv preprint arXiv:2403.18167*, 2024a.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust llm safeguarding via refusal feature adversarial training. *arXiv preprint arXiv:2409.20089*, 2024b.

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. `https://D2L.ai`.

Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. Agent-pro: Learning to evolve via policy-level reflection and optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5348–5375, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.292. URL `https://aclanthology.org/2024.acl-long.292/`.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024b.

Yutong Zhang, Dora Zhao, Jeffrey T Hancock, Robert Kraut, and Diyi Yang. The rise of ai companions: How human-chatbot relationships influence well-being. *arXiv preprint arXiv:2506.12605*, 2025.

Pinlong Zhao, Weiyao Zhu, Pengfei Jiao, Di Gao, and Ou Wu. Data poisoning in deep learning: A survey. *arXiv preprint arXiv:2503.22759*, 2025.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, et al. A survey on vision-language-action models: An action tokenization perspective. *arXiv preprint arXiv:2507.01925*, 2025.

Jiawei Zhou, Kritika Venkatachalam, Minje Choi, Koustuv Saha, and Munmun De Choudhury. Communication styles and reader preferences of llm and human experts in explaining health information. *arXiv preprint arXiv:2505.08143*, 2025a.

Jiayi Zhou, Jiaming Ji, Boyuan Chen, Jiapeng Sun, Wenqi Chen, Donghai Hong, Sirui Han, Yike Guo, and Yaodong Yang. Generative rlhf-v: Learning principles from multi-modal human preference. *arXiv preprint arXiv:2505.18531*, 2025b.

Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. Dynamic evaluation of large language models by meta probing agents. *arXiv preprint arXiv:2402.14865*, 2024.

Quanyan Zhu. Game theory for cyber deception: a tutorial. In *Proceedings of the 6th Annual Symposium on Hot Topics in the Science of Security*, pp. 1–3, 2019.

Artur Zolkowski, Kei Nishimura-Gasparian, Robert McCarthy, Roland S Zimmermann, and David Lindner. Early signs of steganographic capabilities in frontier llms. *arXiv preprint arXiv:2507.02737*, 2025.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.