Defining Strategic AI Deception Through Generalized Propositional Attitudes – Extended Abstract

Tom-Felix Berger

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

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

For example, Tarsney [2025] recently defined deceptive statements by generative AI as follows ($\mathcal{B}1$): "A statement is deceptive with respect to question Q if it tends to move its addressee's beliefs about Q further away from the beliefs that she would endorse under semi-ideal conditions". Semi-ideal conditions mean that one has been presented with all available and relevant information about Q and has been given an adequate amount of deliberation time. This definition of a deceptive statement is behavioral: it avoids the attribution of belief- or desire-like states to the deceiving AI with the aim of circumventing the aforementioned controversies [Tarsney, 2025].

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

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

This paper instead proposes a novel definition of *strategic deception* in terms of generalized propositional attitudes, more specifically, generalizations of beliefs and desires. These generaliza-

tions, g-beliefs and g-desires, are the product of a central trade-off: On the one hand, they should be general enough to be attributable to a wide array of artificial systems, including current LLMs, with as little controversy as possible. On the other hand, they have to be expressive enough to generate useful characterizations of systems that possess them, including a definition of strategic deception that captures its distinctive features and associated risks.

G-beliefs and g-desires are characterized by their functional role in guiding behavior; behavior that promotes the AI's g-desires given its g-believed current situation and environment. Any state of an entity with this functional role qualifies as a g-belief or g-desire, respectively. It will be argued that these generalized propositional attitudes can be attributed to animals, very simple systems such as thermostats (although only with extremely limited propositional content), and to LLMs. Moreover, the following definition of strategic deception in terms of generalized propositional attitudes will be proposed:

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

1) E exhibits behavior that causes A to g-believe p,[1] where p is a false proposition (or prevents A from coming to g-believe $\neg p$).
2) E g-believes $\neg p$.
3) the behavior results from E's g-attitudes in the way that they function to cause behavior:
   3.a) E g-desires that A g-believes p (or that A does not g-believe $\neg p$).
   3.b) E g-believes that this behavior makes A g-believe p.

It will then be argued that 1) $\mathcal{S}$ fits better than its behavioral competitors with pre-theoretic intuitions about strategic deception, 2) that it reflects more accurately practitioner's usage of that term, 3) that it captures forms of deception that support concerns about large-scale risks, and 4) that it is methodologically useful to guide mechanistic approaches for deception detection.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, 2021. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

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

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

Boyuan Chen, Sitong Fang, Jiaming Ji, Yanxu Zhu, Pengcheng Wen, Jinzhou Wu, Yingshui Tan, Boren Zheng, Mengying Yuan, Wenqi Chen, et al. Ai deception: Risks, dynamics, and controls. arXiv preprint, 2025. https://arxiv.org/abs/2511.22619.

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

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

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

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

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

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

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

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

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

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

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

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

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

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. arXiv preprint, 2025. https://arxiv.org/abs/2412.04984.

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

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

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023. doi: 10.1038/s41586-023-06647-8.

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

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

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

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

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