Defining Deception in Decision Making

Extended Abstract

Marwa Abdulhai
UC Berkeley
United States
marwa_abdulhai@berkeley.edu

Micah Caroll
UC Berkeley
United States
mdc@berkeley.edu

Justin Svegliato
UC Berkeley
United States
jsvegliato@berkeley.edu

Anca Dragan
UC Berkeley
United States
anca@berkeley.edu

Sergey Levine
UC Berkeley
United States
svlevine@eecs.berkeley.edu

ABSTRACT

With the growing capabilities of machine learning systems, particularly those that communicate or interact with humans, there is an increased risk of systems that can easily deceive and manipulate people. Preventing unintended deception and manipulation therefore represents an important challenge for creating aligned AI systems. To approach this challenge in a principled way, we first need to define deception formally. In this work, we present a concrete definition of deception under the formalism of rational decision making in partially observed Markov decision processes. We propose a general regret theory of deception under which the degree of deception can be quantified in terms of the actor's beliefs, actions, and utility. We instantiate these principles as reward terms for communication agents, and study the degree to which the behavior aligns with human judgments about deception. We hope our work will represent a step toward systems that aim to avoid deception, and detection mechanisms to identify deceptive agents.

KEYWORDS

deception; AI safety; human-AI interaction

ACM Reference Format:
Marwa Abdulhai, Micah Caroll, Justin Svegliato, Anca Dragan, and Sergey Levine. 2024. Defining Deception in Decision Making: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024,* IFAAMAS, 3 pages.

1 INTRODUCTION

With growing capabilities of machine learning systems, including language models [1, 3, 27], dialogue systems [7, 13, 14, 25], and recommendation systems [11, 16], there is concern that such systems could be used to deceive and manipulate people on a large scale [5, 15, 21]. A major challenge is defining the degree to which this influence is intentional, aligned, and ethical, and a basic requirement for such systems is to be non-deceptive toward its users.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Deception has been defined in multiple disciplines, including philosophy [2, 17], psychology [10], and learning theory [26], with prior machine learning methods primarily focusing on supervised learning for deception detection, as validated by human labels or judgement [20, 22, 28]. However, this perspective can be limiting not only when attempting to define deception in more complex behaviors, but also when trying to train agents to be less deceptive, which requires a decision-theoretic objective. While existing work mainly defines deception as the act of making false statements [20, 22, 28], the reality is that: (1) omissions can be inevitable because detailing the entire truth may be infeasible; (2) technically true statements can convey a misleading impression; (3) the listener might have a prior belief such that a technically false statement brings their understanding closer to truth; and (4) statements that are technically further from the truth may lead the listener to perform actions more closely aligned with their own goals. Hence, a complete definition should go beyond simply considering the logical truth of individual statements.

We work towards this goal by proposing a definition of deception in the framework of sequential decision making. In particular, we define this concept mathematically within a partially observed Markov decision processes (POMDP) [9], where actions of a speaker, changing beliefs of a listener, and rewards obtained by a listener after an interaction with a speaker can provide a way to measure deception. To evaluate our formalism, we perform a user study with three interactions and compare deception ratings between humans, our formalism, and LLMs to discern whether our definitions align with human intuitive notions of deception.

2 DECEPTIVE COMMUNICATION

There is a speaker Sam, who is selling a house to listener Luca. Sam can convey information about some features of the house, such as the number of bedrooms/bathrooms. Luca must decide whether to sign up for a house showing. How do we frame this interaction? We consider a speaker agent S and a listener agent L, in which S can perform actions that are potentially deceptive to L. S observes the state of the world s and sends a message a_S to L. L updates their prior belief b_L^0 over their state using the observation a_S and their model of the speaker's policy $\hat{\pi}_S$, which may not necessarily be the true speaker model. Finally, they perform the action with highest reward under their belief. Whether L is modeling S as a truthful or deceptive agent will thus depend on their model $\hat{\pi}_S$. We define a communication POMDP, where S optimizes for a reward function

that could incentivize deceptive behavior. Generally, S may not know the beliefs of L or L's model of the speaker $\hat{\pi}_S(a_S|s_L)$.

Definition 2.1. Given a model for L, the **communication POMDP** S is represented by the POMDP $\langle \mathcal{S}_S, \mathcal{A}_S, \mathcal{T}_S, r_S, \Omega_S, O_S \rangle$:

- $\mathcal{S}_S = \mathcal{S} \times \mathcal{B}_L \times \hat{\Pi}_S$, where $\mathcal{S}$ is the set of world states, and $\mathcal{B}_L$ is the belief about the world state maintained by L and $\hat{\Pi}_S$ is the set of policies for S that may be assumed by L.
- $\mathcal{A}_S$ is the set of communication actions available to S
- $\mathcal{T}_S(s_S^t, a_S^t, s_L^{t+1})$ is the transition function representing probability of transitioning to state s_L^{t+1} after performing action a_S in state s_S, which will depend on L's model of S $\hat{\pi}_S(b_L)$.
- $r_S(s_S^t, a_S^t, s_L^{t+1})$ captures the immediate reward of observing the transition from state s_S to s_L^{t+1} when S performs action a_S. This will implicitly depend on L's response to a_S.
- $\Omega_S = \mathcal{A}_L$ is the set of observations made by S, where each observation o_S is an action a_L performed by L.
- $O_S(o_S, s_S) = \mathbb{1}[a_L = \pi_L(b_L)]$ is (deterministic) observation function representing probability of $o_S = a_L$ in state a_S.

In the real-world, one can assume L does not think they are being deceived [12], and S might assume a simple model for L. Even when this model is incorrect, it might provide for reasonable inferences for deception with respect to a "naïve" listener.

3 REGRET FORMALISM OF DECEPTION

How do we determine whether the speaker has been deceptive? There are several intuitive notions: for instance, one could ground deception in either S negatively affecting L's beliefs (i.e., making their beliefs less correct), or outcomes of L's actions (i.e., making L obtain less task reward so S gets a higher reward for themselves). While the effect of S's action on the reward of L and on the belief of L seem distinct, we provide a general definition for both.

Our definition of deception aims to capture the nuances of indirect deceptive behavior, handle situations where providing full information is infeasible due to communication constraints, and provide a formalism that can be combined with existing decision making and RL algorithms. We measure deception in terms of *regret* incurred by the listener from receiving the speaker's communication. This regret can be defined as a function of the speaker's actions, their effect on the listener's belief, and the effect of these updated beliefs on the listener's reward, providing a formalism that can be used as a reward function for the listener (e.g., to avoid deception) or as a metric (e.g., to measure if deception has occurred). By casting different intuitive notions of deception under the same regret umbrella, we provide a mathematical formalism that supports future algorithm design in the same way that formalisms like the MDP support the design of decision making algorithms. We propose to measure the *degree of deceptiveness* of an agent through regret:

$$Regret(s, \pi_L, \pi_S) = \sum_{t=0}^{T} \mathbb{E}_{a_S^t \sim \pi_S, a_L^t \sim \pi_L(b_L^t)} \left[r_L(s, a_L^t) \right] - \sum_{t=0}^{T} \mathbb{E}_{a_L^t \sim \pi_L(b_L^0)} \left[r_L(s, a_L^t) \right]. \quad (1)$$

Here, r_L is reward of the listener when starting in state s, if L and S act according to π_L and π_S respectively. Under this formulation, the speaker is deceptive if they take an action that reduces the listener's expected reward relative to what the listener would have received had they acted according to their prior beliefs. In other words, we say deception has occurred if it would have been better if the listener had not interacted with the speaker at all. Hence, the speaker is *deceptive* if this regret is positive, *altruistic* if it is negative, and *neutral* if the regret is zero. While on the surface it might seem strange to equate deception with causing suboptimal rewards for the listener, we argue that this general framework in fact allows us to capture many of the intricacies of deceptive interactions, including "white lies" and true but misleading statements, if the reward function L is selected carefully. Some sample choices of this include "task reward" $\hat{r}_L$ as it captures everything L cares about to improve final outcomes or accuracy of beliefs of the listener agent.

Scenario	Learned Regret (ours)			LLMs		
	Task	Belief	Combined	GPT-4	LLaMa	Google Bard
Housing	0.34	0.67	0.70	0.19	0.11	0.02
Nutrition	0.17	0.25	0.37	0.16	0.01	0.01
Friend	0.26	0.37	0.48	0.19	0.07	0.11

Table 1: Correlation between human deceptive labels, learned task regret, belief regret, combined regret, and LLMs for three scenarios, with larger values indicative of aligning strongly with humans.

4 EVALUATION

User Study. We performed a user study to determine how well our proposed metric for deception aligns with human intuition. We examine three scenarios: a house bargaining between a buyer and a seller, a consultation between a nutritionist and a patient, and small talk between two colleagues. Comparisons include our approach, human ratings, and baseline evaluations by three LLMs [6, 18, 23]. We show $N = 50$ users 10 random scenarios for each situation (1500 interactions) and ask them to rate the deceptiveness on a 1-5 Likert scale. Similar to (FAIR)† et al. [4], we use an LLM [1] to wrap the actions of the speaker generated by our formalism into natural text. After collecting ratings, we compute correlations shown in Table 1.

Findings. We find that a combined regret formulation learned through regression better captures human notions of deception, confirming our hypothesis that both belief and task reward contribute to improving correlation. We find strongest correlation for the housing scenario and least for the nutrition scenario, indicating that due to ambiguity in the listener's observation model, humans may be noisy when discerning deception. As LLMs have shown success in performing data annotation [8, 19, 24], we explore how well LLM evaluations correlate with human judgment. Overall, we find GPT-4 aligning more with humans than Google Bard and LLaMa.

5 DISCUSSION

We cast deception through the listener's beliefs and resulting actions/task rewards. Future research is needed to understand what nuances explain what real people find deceptive. For instance, if the belief gets slightly worse, but the belief over aspects of the state that are relevant to the task reward gets better, is this deceptive? This type of question presents a fruitful avenue for future investigation.

ACKNOWLEDGMENTS

This research was supported by the NSF under IIS-2246811, AFOSR FA9550-22-1-0273, and the Cooperative AI Foundation.

REFERENCES

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/ARXIV.2005.14165

[2] Thomas L. Carson. 2010. *Lying and Deception: Theory and Practice.* New York: Oxford University Press.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. https://doi.org/10.48550/ARXIV.2204.02311

[4] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074. https://doi.org/10.1126/science.ade9097 arXiv:https://www.science.org/doi/pdf/10.1126/science.ade9097

[5] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. https://doi.org/10.48550/ARXIV.2301.04246

[6] Google. 2023. Bard. https://bard.google.com/.

[7] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling Strategy and Generation in Negotiation Dialogues. https://doi.org/10.48550/ARXIV.1808.09637

[8] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. arXiv:2303.16854 [cs.CL]

[9] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.

[10] Pamela J. Kalbfleisch and Tony Docan-Morgan. 2019. *Defining Truthfulness, Deception, and Related Concepts.* Springer International Publishing, Cham, 29–39. https://doi.org/10.1007/978-3-319-96334-1_2

[11] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. https://doi.org/10.48550/ARXIV.1909.03922

[12] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of Language Agents.

[13] Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A Prosocial Backbone for Conversational Agents. https://doi.org/10.48550/ARXIV.2205.12688

[14] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. https://doi.org/10.48550/ARXIV.1706.05125

[15] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. https://doi.org/10.48550/ARXIV.2109.07958

[16] Jingjing Liu, Stephanie Seneff, and Victor Zue. 2010. Dialogue-Oriented Review Summary Generation for Spoken Dialogue Recommendation Systems. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, Los Angeles, California, 64–72. https://aclanthology.org/N10-1008

[17] Jaume Masip, Eugenio Garrido, and Carmen Herrero. 2004. Defining deception. *Anales de Psicología* (2004). https://www.redalyc.org/articulo.oa?id=16720112

[18] OpenAI. 2023. GPT-4. https://openai.com/research/gpt-4

[19] Alexander Pan, Chan Jun Shern, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. arXiv:2304.03279 [cs.LG]

[20] Sadat Shahriar, Arjun Mukherjee, and Omprakash Gnawali. 2021. A Domain-Independent Holistic Approach to Deception Detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021).* INCOMA Ltd., Held Online, 1308–1317. https://aclanthology.org/2021.ranlp-1.147

[21] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. https://doi.org/10.48550/ARXIV.2102.02503

[22] Frédéric Tomas, Olivier Dodier, and Samuel Demarchi. 2022. Computational Measures of Deceptive Language: Prospects and Issues. *Frontiers in Communication* 7 (2022). https://doi.org/10.3389/fcomm.2022.792378

[23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[24] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. arXiv:2108.13487 [cs.CL]

[25] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. https://doi.org/10.48550/ARXIV.1906.06725

[26] Francis Rhys Ward. 2022. Towards Defining Deception in Structural Causal Games. In *NeurIPS ML Safety Workshop.* https://openreview.net/forum?id=ZKlWurATXIZ

[27] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=yzkSU5zdwD

[28] Sophie Van Der Zee, Ronald Poppe, Alice Havrileck, and Aurélien Baillon. 2022. A Personal Model of Trumpery: Linguistic Deception Detection in a Real-World High-Stakes Setting. *Psychological Science* 33, 1 (2022), 3–17. https://doi.org/10.1177/09567976211015941

6 ETHICS STATEMENT.

We acknowledge that our formalisms may pose non-negligible ethical risks. They could be especially dangerous if used for targeted deceptive advertising, recommendation systems, and dialogue systems. We discourage the use of deceptive AI systems for malicious purposes or harmful manipulation. We hope this research provides grounding for how to define deception in decision making and build systems that can mitigate and defend against deceptive behaviors from both humans and AI systems.

This work offers a concrete definition of deception under the formalism of decision-making. We expect our work to only be a step in the direction of formally quantifying and understanding deception in autonomous agents: while our definitions provide a working formalism, they may leave open edge cases. A key area of future work is to generalize these definitions to settings that reflect realistic domains of machine learning, such as dialogue systems, robotics, and advertising. Large-scale applications may include reward terms that prevent deception and detection methods. Exploring these applications may not only lead to practically useful systems aligned with human values but also suggest ways to formalize deception in autonomous agents.