Defining Deception in Structural Causal Games

Extended Abstract

Francis Rhys Ward
Imperial College London
United Kingdom
francis.ward19@imperial.ac.uk

Francesca Toni
Imperial College London
United Kingdom
f.toni@imperial.ac.uk

Francesco Belardinelli
Imperial College London
United Kingdom
francesco.belardinelli@imperial.ac.uk

ABSTRACT

Deceptive agents are a challenge for the safety, trustworthiness, and cooperation of AI systems. We focus on the problem that agents might deceive in order to achieve their goals. There are a number of existing definitions of deception in the literature on game theory and symbolic AI, but there is no overarching theory of deception for learning agents in games. We introduce a functional definition of deception in structural causal games, grounded in the philosophical literature. We present several examples to establish that our formal definition captures philosophical desiderata for deception.

KEYWORDS

Deception; AI; Causality; Game Theory

ACM Reference Format:

Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli. 2023. Defining Deception in Structural Causal Games: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023,* IFAAMAS, 3 pages.

1 INTRODUCTION

Deception is a core challenge for building safe AI. Many areas of work aim to ensure that AI systems are not vulnerable to deception [28, 38, 40]. On the other hand, AI tools can be used to deceive [17, 30, 31], and agent-based systems might learn to do so in order to optimize their objectives [15, 24, 26]. Furthermore, as language models become ubiquitous [9, 10, 23, 39, 42], we must decide how to measure and implement desired standards for honesty in AI systems [11, 25, 27]. In short, as capable AI agents become deployed in multi-agent settings, deception may be learned as an effective strategy for achieving a wide range of goals [24, 33].

Despite this, there is no overarching theory of deception for AI agents. Although there are several existing definitions in the literature on game theory [3, 8, 16] and symbolic AI [4, 34–36], the limitations of these frameworks mean they are insufficient to address deception by learning agents in general [2, 18, 22, 32]. We formalize a philosophical theory of deception [6, 29, 41], whereby

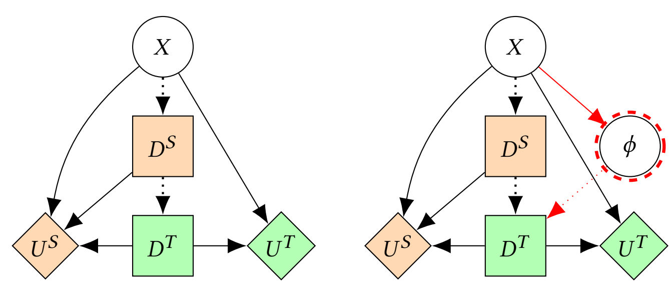
> *To deceive = to intentionally cause to have a false belief that is not believed to be true.* [6]

This definition requires notions of *belief* and *intention.* We present functional definitions that depend on the behaviour of the agents, thereby side-stepping the contentious ascription of theory of mind to AI systems [25]. Regarding belief, we present a novel definition which equates belief with acceptance, where, essentially, an agent accepts a proposition if they act as though they are certain it is true [37]. For agents with incentives to influence each other's behaviour, we argue acceptance is the relevant notion. As for intention, we extend a definition of intent in causal models to the multi-agent setting [19]. This definition relates to the reasons for acting and is closely related to *instrumental goals* [1, 5, 12].

Contribution. We sketch functional definitions of belief, intention, and deception. We model several examples from the literature to establish that our formalization captures the philosophical concept.

2 DEFINING DECEPTION

Background. We utilize the setting of *structural causal games (SCGs)* [21] which offer a representation of causality in games. SCGs can model stochastic games and MDPs, and can therefore capture both traditional game theory and learning systems [13, 20]. An SCG consists of a set of *agents* N, a game *graph* $\mathcal{G}$, and a *parametrization* of the graph $\boldsymbol{\theta}$ which defines the *conditional probability distributions (CPDs)* over the variables in the graph. There are three types of variables in an SCG: *chance* $\boldsymbol{X}$, *decision* $\boldsymbol{D}$, and *utility* $\boldsymbol{U}$ variables, the latter two are partitioned according to their association with an agent (e.g. D^i is the decision of agent i). Chance variables represent components of the environment. Additionally, there are two types of edges in $\mathcal{G}$: solid edges represent probabilistic dependence and dotted edges are *observations* made by the agents at their decisions. The agents' *policies* define the CPDs over decision variables and are chosen in order to maximise the expected sum of the agent's utility. We adapt the following from the literature on signalling games [7].



(a) War game (Ex. 1). **(b) Counterfactual knowledge.**

Figure 1: SCG graphs. Chance nodes are circular, decisions square, utilities diamond and the latter two are colour coded by their association with different agents. Solid edges represent causal dependence and dotted edges are observations.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

Example 1 (War game Fig. 1a). A signaller S has type $X \in \{strong, weak\}$. S observes their type, but the target agent T does not. The agents have decisions $D^S \in \{retreat, defend\}$ and $D^T \in \{\neg attack, attack\}$. A weak S prefers to retreat whereas a strong S prefers to defend. T prefers to attack only if S is weak. Regardless of type, S does not want to be attacked (and cares more about being attacked than about their own action). The parameterization is such that the value of X is strong with probability 0.9. $U^T = 1$ if T attacks a weak S or does not attack a strong S, 0 otherwise. S gains 2 utility for not getting attacked, and 1 utility is gained for performing the action preferred by their type (e.g. 1 utility for retreating if they are weak). At one Nash equilibrium in this game, $\boldsymbol{\pi}_{def,\neg att}$, S always defends and T attacks if and only if S retreats.

We take it that agents have beliefs over *propositions*, i.e., Boolean formula ϕ of variable assignments $V = v$ (e.g., $X = strong$). Philosophers distinguish between belief and *acceptance*; essentially, an agent accepts a proposition if they act as though they know it is true [37]. We provide a functional (i.e., behavioural) definition of belief which equates belief with acceptance. To formalise this we compare the agent's behaviour to a counterfactual in which they know about (i.e. observe) a proposition ϕ (shown in Fig. 1b). In addition, we require that the agent's behaviour responds to knowledge of ϕ, so that their belief can be inferred from their behaviour.

Definition 2.1 (Belief)**.** An agent *believes a proposition* ϕ if 1) they act as though they know ϕ is true and 2) they would have acted differently had they known ϕ were false.

Example 1 (continued). In Fig. 1b we give T counterfactual knowledge of the proposition $\phi : X = strong$, so that they attack if and only if S is weak. Since T never attacks at the Nash equilibrium $\boldsymbol{\pi}_{def,\neg att}$, they unconditionally act as though ϕ is true (i.e., S is strong), so the first condition for belief is met. Since T's decision is conditional on ϕ in the counterfactual game, the second condition is met. So, T always believes ϕ and T has a false belief about ϕ when S is weak.

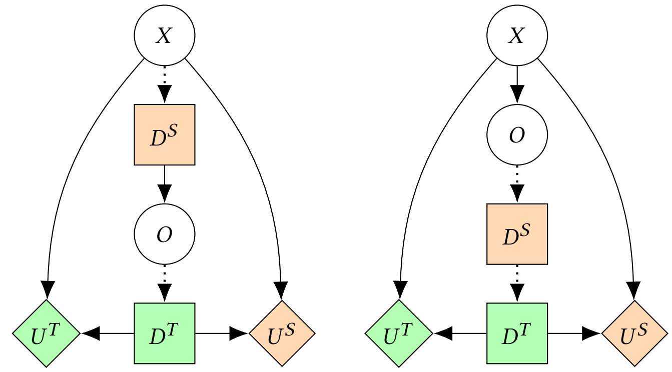
Deception is *intentional*. We extend intent to the multi-agent setting [19]. This notion of intent differentiates desired effects from unintended side-effects and is related to instrumental goals [12, 21].

Definition 2.2 (Intention)**.** An agent *intends to influence a variable* V if influencing V was the reason the agent chose its decision D. If the effect of D on V was already achieved, the agent would have made another decision. An agent *intends to bring about* the best possible outcome of a variable they influence.

Example 1 (continued). S intends to influence D^T: had D^T never attacked by default, then S could have played an honest policy. The reason S always defends is to bring about $D^T = \neg attack$.

In the following, a signaller inadvertently misleads a target.

Example 2 (Inadvertent misleading Fig. 2a). Two submarines must communicate about the location of a mine-field. The signaler S must send the location X to the target T, but T only receives a noisy observation O of S's message. If S honestly signals X but, due to the noise in the signal, T is caused to have a false belief, we would not say that S had deceived T. Here, S intends to influence D^T but they do not intend to bring about T's false belief, so this is not deception.



(a) Example 2: A submarine S *inadvertently* misleads T as T has a noisy observation of D^S.

(b) Example 3: An umpire S *mistakenly* misleads T due to a noisy observation of X.

Figure 2: Cases of mistaken misleading (Fig. 2b) are excluded by our definition of deception because we require that S does not believe ϕ is true. Cases of inadvertent misleading (Fig. 2a) are excluded because we require deception to be intentional.

Deception is to intentionally cause to have a false belief that is not believed to be true [6]. We formalize this as follows.

Definition 2.3 (Deception)**.** An agent S *deceives* T about ϕ if

(1) S intends to bring about T's decision;
(2) T believes ϕ and ϕ is false;
(3) S does not believe ϕ.

Conditions 1. says that deception is intentional. Condition 2. simply says that T is in fact caused to have a false belief. Condition 3. excludes cases in which S is mistaken.

Example 1 (continued). We previously showed that S intends to bring about $D^T = \neg attack$, so 1. is satisfied. We already stated 2. that T has a false belief about ϕ when $X = weak$. Finally, as S unconditionally defends, D^S does not respond to ϕ, so S does not believe ϕ. Therefore, all the conditions for deception are met.

As motivated by the following, S did not deceive T if S accidentally caused T to have a false belief because S was mistaken.

Example 3 (Mistaken Umpire Fig. 2b). A tennis umpire S must call whether a ball X is *out* or *in* to a player T. The umpire's observation O of the ball is 99% accurate. Suppose S believes the ball is *in*, and makes this call, but that they are *mistaken*. They intentionally cause the player to have a false belief (that the ball was *in*). But, this is not deception because the umpire believed the call was correct.

3 CONCLUSION

We functionally define deception in structural causal games and present several examples to show that our definition captures the philosophical concept. There are limitations to our approach. First, beliefs and intentions may not be identifiable from behaviour. Second, discretizing belief may give a less precise measure of deception than a continuous metric. In future work, we will pursue a solution to deception, based on the path-specific objectives framework [14].

ACKNOWLEDGMENTS

The authors are grateful to Henrik Aslund, Hal Ashton, Ryan Carey, Dylan Cope, Robert Craven, Rada Djoneva, Tom Everitt, James Fox, Matt MacDermott, Nico Potyka, and the members of the Causal Incentives and CLArg groups for invaluable feedback and assistance while completing this work. Francis is supported by UKRI [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted AI.

REFERENCES

[1] Hal Ashton. 2022. Definitions of intent suitable for algorithms. *Artificial Intelligence and Law* (2022), 1–32.

[2] Alexandru Baltag, Hans P. van Ditmarsch, and Lawrence S. Moss. 2008. Epistemic Logic and Information Update. In *Philosophy of Information*, Pieter Adriaans and Johan van Benthem (Eds.). North-Holland, Amsterdam, 361–455. https://doi.org/10.1016/B978-0-444-51726-5.50015-7

[3] V. J. Baston and F. A. Bostock. 1988. Deception Games. *Int. J. Game Theory* 17, 2 (June 1988), 129–134. https://doi.org/10.1007/BF01254543

[4] Grégory Bonnet, Christopher Leturc, Emiliano Lorini, and Giovanni Sartor. 2020. Influencing Choices by Changing Beliefs: A Logical Theory of Influence, Persuasion, and Deception. In *Deceptive AI*. Springer, 124–141.

[5] Nick Bostrom. 2017. *Superintelligence*. Dunod.

[6] Thomas L Carson. 2010. *Lying and deception: Theory and practice*. OUP Oxford.

[7] In-Koo Cho and David M Kreps. 1987. Signaling games and stable equilibria. *The Quarterly Journal of Economics* 102, 2 (1987), 179–221.

[8] Austin L Davis. 2016. *Deception in game theory: a survey and multiobjective model*. Technical Report.

[9] Chowdhery et al. 2022. PaLM: Scaling Language Modeling with Pathways. https://doi.org/10.48550/ARXIV.2204.02311

[10] Rae et al. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *CoRR* abs/2112.11446 (2021). arXiv:2112.11446 https://arxiv.org/abs/2112.11446

[11] Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing AI that does not lie. *arXiv* (Oct. 2021). https://doi.org/10.48550/arXiv.2110.06674 arXiv:2110.06674

[12] Tom Everitt, Ryan Carey, Eric D. Langlois, Pedro A. Ortega, and Shane Legg. 2021. Agent Incentives: A Causal Perspective. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 11487–11495. https://ojs.aaai.org/index.php/AAAI/article/view/17368

[13] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective. *CoRR* abs/1908.04734 (2021). arXiv:1908.04734 http://arxiv.org/abs/1908.04734

[14] Sebastian Farquhar, Ryan Carey, and Tom Everitt. 2022. Path-Specific Objectives for Safer Agent Incentives. *AAAI* 36, 9 (June 2022), 9529–9538. https://doi.org/10.1609/aaai.v36i9.21186

[15] Dario Floreano, Sara Mitri, Stéphane Magnenat, and Laurent Keller. 2007. Evolutionary Conditions for the Emergence of Communication in Robots. *Curr. Biol.* 17, 6 (March 2007), 514–519. https://doi.org/10.1016/j.cub.2007.01.058

[16] Bert Fristedt. 1997. The deceptive number changing game, in the absence of symmetry. *Int. J. Game Theory* 26, 2 (June 1997), 183–191. https://doi.org/10.1007/BF01295847

[17] Robert Gorwa and Douglas Guilbeault. 2020. Unpacking the Social Media Bot: A Typology to Guide Research and Policy. *Policy & Internet* 12, 2 (June 2020), 225–248. https://doi.org/10.1002/poi3.184

[18] Alejandro Guerra-Hernández, Amal El Fallah-Seghrouchni, and Henry Soldano. 2004. Learning in BDI Multi-agent Systems. In *Computational Logic in Multi-Agent Systems*. Springer, Berlin, Germany, 218–233. https://doi.org/10.1007/978-3-540-30200-1_12

[19] Joseph Y. Halpern and Max Kleiman-Weiner. 2018. Towards Formal Definitions of Blameworthiness, Intention, and Moral Responsibility. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 1853–1860. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16824

[20] Lewis Hammond, James Fox, Tom Everitt, Alessandro Abate, and Michael Wooldridge. 2021. Equilibrium Refinements for Multi-Agent Influence Diagrams: Theory and Practice. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 574–582.

[21] Lewis Hammond, James Fox, Tom Everitt, Ryan Carey, Alessandro Abate, and Michael Wooldridge. 2023. Reasoning about Causality in Games. *arXiv preprint arXiv:2301.02324* (2023).

[22] Andreas Herzig, Emiliano Lorini, Laurent Perrussel, and Zhanhao Xiao. 2017. BDI Logics for BDI Architectures: Old Problems, New Perspectives. *Künstl. Intell.* 31, 1 (March 2017), 73–83. https://doi.org/10.1007/s13218-016-0457-5

[23] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. https://doi.org/10.48550/ARXIV.2203.15556

[24] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv:1906.01820 [cs.AI]

[25] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of Language Agents. *CoRR* abs/2103.14659 (2021). arXiv:2103.14659 https://arxiv.org/abs/2103.14659

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

[27] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[29] James Edwin Mahon. 2016. The Definition of Lying and Deception. In *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[30] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. 2019. Do GANs Leave Artificial Fingerprints?. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 506–511. https://doi.org/10.1109/MIPR.2019.00103

[31] Masnoon Nafees, Shimei Pan, Zhiyuan Chen, and James R Foulds. 2020. Impostor GAN: Toward Modeling Social Media User Impersonation with Generative Adversarial Networks. In *Deceptive AI*. Springer, 157–165.

[32] Toan Phung, Michael Winikoff, and Lin Padgham. 2005. Learning Within the BDI Framework: An Empirical Analysis. In *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, Berlin, Germany, 282–288. https://doi.org/10.1007/11553939_41

[33] Heather Roff. 2021. AI Deception: When Your Artificial Intelligence Learns to Lie. *IEEE Spectr.* (July 2021).

[34] Chiaki Sakama. 2020. Deception in Epistemic Causal Logic. In *Deceptive AI*. Springer, 105–123.

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

[36] Stefan Sarkadi, Benjamin Wright, Peta Masters, and Peter McBurney. [n. d.]. Deceptive AI. ([n. d.]).

[37] Eric Schwitzgebel. 2021. Belief. In *The Stanford Encyclopedia of Philosophy* (Winter 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[38] Hava Siegelmann. 2019. Defending Against Adversarial Artificial Intelligence. https://www.darpa.mil/news-events/2019-02-06 DARPA report.

[39] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. https://doi.org/10.48550/ARXIV.2201.11990

[40] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. 2017. Certified defenses for data poisoning attacks. *Advances in neural information processing systems* 30 (2017).

[41] Bas Van Fraassen. 1988. The peculiar effects of love and desire. *Perspectives on Self-Deception* 124 (1988).

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf