OPEN ACCESS

KING'S College LONDON

King's Research Portal

Document Version
Publisher's PDF, also known as Version of record

Link to publication record in King's Research Portal

Citation for published version (APA):
Sarkadi, S. (2020). Argumentation-based Dialogue Games for Modelling Deception. In F. Castagna, F. Mosca, J. Mumford, S. Sarkadi, & A. Xydis (Eds.), *Online Handbook of Argumentation for AI* (Vol. 1, pp. 38-42) https://arxiv.org/pdf/2006.12020.pdf#chapter.8

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 08. Dec. 2021

Argumentation-based Dialogue Games for Modelling Deception

Ștefan Sarkadi

Department of Informatics, King's College London, UK

Abstract

Machines of the future might either be endowed with or might develop mechanisms to argue with other agents. We consider the contexts in which these types of machines also develop reasons to act dishonestly by attempting to deceive their interlocutors. Using the argumentation dialogue games approach, this work aims to explore how deceptive machines might be engineered in order to mitigate or neutralise their malicious behaviour. Argumentation dialogue games can be a powerful approach for the modelling of deception given that it offers an *explainable* way of representing the components necessary for deception such as the knowledge of the agents, their ability to perform actions (to communicate arguments), their ability to reason defeasibly about the world, and most importantly, their ability to reason defeasibly about each others' minds. This paper presents three different hybrid agent-based models derived from argumentation that (i) have been successfully used and that (ii) can be used in future work to model machine deception.

1 Introduction

The ability of machines to deceive autonomously is increasingly drawing the interest of the AI community, as well as the interest of the philosophical and digital humanities communities. This has also been enhanced by the emergence of the *post-truth* technological era driven by the popularisation of the term *fake news* [Lazer et al., 2018].

Currently, most of the approaches in AI merely focus on using machine learning capabilities to either (i) generate fake information [Yao et al., 2017], or (ii) detect fake news from big data [Conroy et al., 2015], or (iii) by enhancing machine learning using techniques such as argument mining to detect deception [Cocarascu and Toni, 2016]. However, as pointed out in [Sarkadi, 2018] and in [Sarkadi, 2020b], these data-oriented approaches fail to account for several critical components of deception, namely the intention of the agents to deceive, their *Theory of Mind* of their targets, and the reasoning behind their deceptive acts. Apart from these explainability issues (see [Miller, 2018]) which machine learning approaches face, there is also the issue of representational and design accuracy. Deception and deception detection require *social intelligence*. That is the ability of agents to reason about other minds in order to influence other agents through communication [Castelfranchi, 1998]. In this context, communication represents social actions which influence belief changes, and by extension behavioural changes, in other social agents. Is the AI used behind fake news and deepfakes truly autonomous? Obviously it is not, since it is unable to act autonomously, as well as unable to reason about what information should be fed to whom in order to deceive. These so called types of "deceptive" AI merely act as tools in the hands of truly socially intelligent agents, namely the humans that have the intent and the communicative capabilities to act dishonestly. How do we then model socially-intelligent autonomous deceptive agents that truly behave dishonestly?

In this paper we present a novel argumentation-based dialogue game method to model and study autonomous deceptive agents. This method is extensively described in [Sarkadi, 2020a] and aims to address the problem of modelling deceptive machines from the socially-intelligent agent-based perspective.

2 Method

To model deception, we mainly focus on interactions between two agents, the *Deceiver*, and its target, which we have called the *Interrogator*. The aim of the Deceiver is, obviously, to deceive the Interrogator, whereas the aim of the Interrogator is to find out the desired truth. We have adopted the following definition of deception:

Deception *The intention of a deceptive agent, to make or cause another target agent to believe something is true that the deceiver believes is false, with the aim of achieving an ulterior goal or desire.*

The method we have used to address deception as defined above is the application of opponent modelling to dialogue argumentation games. We have used *belief-desire-intention* BDI-like architectures to model the cognitive properties of the agents that play these dialogue games [McBurney and Parsons, 2009]. Giving BDI agents a communication protocol along with a reasoning mechanism enables them to think pragmatically about their beliefs, their desires, and their intentions in order to perform speech acts [Rao et al., 1995]. In argumentation, these speech acts can represent arguments, as well as argument chains and argument systems. Apart form performing speech acts, our agents are able to reason about their opponent's mind. In other words, they have a Theory of Mind (ToM) that enables mind-reading. ToM is the ability of an agent to reason about the mental attitudes (beliefs, desires, and intentions) of another agent [Goldman, 2012]. According to [Isaac and Bridewell, 2017], mind-reading is a crucial ability for a machine to have in order to be able to deceive or detect deception.

3 Discussion

In this section we present three models that we have built using the method presented in the previous section. All of these models represent deception according to our adopted definition where the Deceiver aims to achieve an ulterior goal. Once this ulterior goal is achieved, deception becomes successful.

The ulterior goal of the Deceiver can even be as simple as desiring that the target believes something is true, when the Deceiver believes it is false. If there is another ulterior goal, that in order to be met requires deception, then the causation of a false belief becomes the subgoal. This is the case in [Panisson et al., 2018], where we have implemented in an agent-oriented programming language a car-dealing agent that deceives in order to cause its target to buy a car the dealer agent desires to sell. The implemented agent can also decide to lie or bullshit. However, we mention the fact that lying or bullshitting is different from deception as they do not require a ToM to cause a false belief.

Another form of ulterior goal is in the case of interrogation games, where the Deceiver needs to cause the Interrogator into accepting the story that emerges from their dialogue. In [Sarkadi et al., 2019a], we have presented an argumentation dialogue game model for generating credible stories. In this model, both Deceiver and Interrogator use the same Toulmin-like reasoning technique to generate simple or complex arguments (that represent stories or narratives) using the ToM of their opponent. The ToM of the opponent contains simple arguments, as well as argument attacks and argument backings, that the opponent knows.

Estimating the success of deception given the communication of an argument can be problematic. The dynamics of deceptive attempts can be influenced by the uncertainty of certain social factors such as: the trust of the interlocutor, the degree of confidence in one's own ToM of the interlocutor, and one's degree of communicative skill. In [Sarkadi et al., 2019b] we have continued our work from [Panisson et al., 2018]. We have

used BDI architectures to model, evaluate and implement in an Agent-oriented Programming Language deceptive interactions under factors of social uncertainty. This model aims to integrate components of two major theories of deception, namely *Interpersonal Deception Theory* [Buller and Burgoon, 1996] and *Information Manipulation Theory 2* [McCornack et al., 2014]. By modelling information manipulation as well as uncertainty of the social factors, we have enabled the Interrogator to consider its own degree of trust in the Deceiver in order to reason about what is being communicated. Given the levels of trust of the Interrogator, we have also enabled the Deceiver to estimate its success at deception by taking into account the trust of its target, the uncertainty of its ToM of the target, as well as its communicative skill. A critical result from the evaluation of the model shows that agents with **strong skeptical attitudes** are prone to **unintended deception**.

Properties

The three models that we have developed in [Panisson et al., 2018], [Sarkadi et al., 2019a], and [Sarkadi et al., 2019b] present different desirable properties that are useful for the study of machine deception (See Table 1). Below we describe each of these properties:

1. **Explainability** should be a crucial property of argumentation-based models of deception. We should be able to evaluate deceptive mind games and say whether deception takes places and if it does we need to explain why and under which conditions it does. An explainable model should be able to inform us if deception can be prevented or mitigated in different contexts.

2. **Unintended Deception** happens when the Deceiver does not attempt deception, but the consequences of its communicative acts result in its potential target to be deceived. It is important for models of deception to be able to represent such unintended consequences as they are critical for accountability. We need to be able to tell if an agent that has the ability to deceive should be held responsible for its actions or not.

3. **Uncertainty** in communication should be considered when modelling deception. This is especially important for modelling an agent that estimates its likelihood of success, as well as modelling agents with different degrees of trust in each other. While most of the times trust should be a default attitude towards others [Levine, 2014], in cases of potential deception this is not the case.

4. **Storytelling** is the ability of an agent to communicate arguments in such a way as to describe to another agent a meaningful chain of events. The ability to build narratives is an emerging topic in AI. Deceptive agents can use this ability to their own benefit, e.g. deliver a fictitious story that compels a jury into absolving them of a crime. Therefore, it should be desirable for models of deception to consider or represent such mechanisms.

5. **Deception Detection** is desirable to be represented in a model. While some models represent and explain why deception is successful, they do not represent how and why deception might be detected. It is also important to distinguish between an agent that has the ability to detect deception and one that has the tendency to believe or not what a Deceiver is communicating, which is the case in some models. Representing deception detection could be also useful in showing how a Deceiver might act knowing that its target is able to detect its deceptive intents, as well as how its target might detect them.

6. **Implementation** is to be desired, but not necessary for modelling deception, or any other social phenomenon. However, demonstrating an implementation of the model helps others to use it for studying different multi-agent system setups and scenarios of social interactions. Implementation also improves the **transparency** of a model,

increasing the model's accessibility through its code.

Model Properties	1	2	3	4	5	6
[Panisson et al., 2018]	✓	-	-	-	-	✓
[Sarkadi et al., 2019a]	✓	-	-	✓	✓	-
[Sarkadi et al., 2019b]	✓	✓	✓	-	-	✓

Table 1: Comparison of our models in terms of their respective desirable properties for the study of machine deception.

Future Work

Modelling deception using dialogue games for argumentation offers explanatory and representational power, especially if we want to show properties such as the ones expressed by the models we introduced and described here. However, none of the models presented expresses the full spectrum of these properties. By no means should this discourage the continuation of our method for the study of deception. This paper has classified *a posteriori* the models according to the properties they have managed to express. The models have not been designed and built starting from an *a priori* knowledge of this classification. Having defined this classification should help us continue using our method towards building more expressive models of machine deception.

A problem that future work should aim to overcome is the introduction of an environment in the socio-dynamical representations of deception using dialogue games. Our models have mainly focused on the direct interaction between the Deceiver and its target, but unfortunately they have failed to take into account how a Deceiver might use the environment for manipulating the beliefs of its target. We believe this could be a viable research path worth pursuing in the modelling of machine deception.

4 Conclusion

In this paper we have presented an argumentation-based dialogue game method for modelling deception in AI that is extensively described in [Sarkadi, 2020a]. We have introduced two critical components for representing deception, namely BDI agent architectures and Theory of Mind. Our method relies on these components for representing social interactions between two agents, the Deceiver and the Interrogator. We have described and compared three models that we have built using the presented method. We have also compared the models according to their results and to several desirable properties introduced in the paper, namely *explainability, unintended deception, uncertainty, storytelling, deception detection, and implementation.*

Acknowledgements

The research described by this short paper could not have been possible without their influence and/or contribution: R.H. Bordini, M. Chapman, P. McBurney, F. Mosca, A.R. Panisson and S. Parsons.

References

[Buller and Burgoon, 1996] Buller, D. B. and Burgoon, J. K. (1996). Interpersonal Deception Theory. *Communication Theory*, 6(3):203–242.

[Castelfranchi, 1998] Castelfranchi, C. (1998). Modelling social action for ai agents. *Artificial intelligence*, 103(1-2):157–182.

[Cocarascu and Toni, 2016] Cocarascu, O. and Toni, F. (2016). Detecting deceptive reviews using argumentation. In *Proceedings of the 1st International Workshop on AI for Privacy and Security*, page 9. ACM.

[Conroy et al., 2015] Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

[Goldman, 2012] Goldman, A. I. (2012). Theory of mind. In *The Oxford Handbook of Philosophy of*

Cognitive Science, volume 1. Oxford Handbooks Online, 2012 edition.

[Isaac and Bridewell, 2017] Isaac, A. and Bridewell, W. (2017). *White lies on silver tongues: Why robots need to deceive (and how)*. Oxford University Press.

[Lazer et al., 2018] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

[Levine, 2014] Levine, T. R. (2014). Truth-Default Theory (TDT). *Journal of Language and Social Psychology*, 33(4):378–392.

[McBurney and Parsons, 2009] McBurney, P. and Parsons, S. (2009). Dialogue games for agent argumentation. In Simari, G. and Rahwan, I., editors, *Argumentation in Artificial Intelligence*, pages 261–280. Springer US.

[McCornack et al., 2014] McCornack, S. A., Morrison, K., Paik, J. E., Wisner, A. M., and Zhu, X. (2014). Information manipulation theory 2: a propositional theory of deceptive discourse production. *Journal of Language and Social Psychology*, 33(4):348–377.

[Miller, 2018] Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

[Panisson et al., 2018] Panisson, A. R., Sarkadi, S., McBurney, P., Parsons, S., and Bordini, R. H. (2018). Lies, bullshit, and deception in agent-oriented programming languages. In *20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018*, pages 50–61. CEUR-WS.

[Rao et al., 1995] Rao, A. S., Georgeff, M. P., et al. (1995). BDI agents: from theory to practice. In *ICMAS*, volume 95, pages 312–319.

[Sarkadi, 2018] Sarkadi, S. (2018). Deception. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5781–5782. AAAI Press.

[Sarkadi, 2020a] Sarkadi, S. (2020a). *Deception*. PhD thesis, King's College London.

[Sarkadi, 2020b] Sarkadi, S. (2020b). Deceptive autonomous agents. Cranfield Online Research Data (CORD).

[Sarkadi et al., 2019a] Sarkadi, S., McBurney, P., and Parsons, S. (2019a). Deceptive storytelling in artificial dialogue games. In *Proceedings of the 2019 AAAI Spring Symposium Series on Story-Enabled Intelligence*.

[Sarkadi et al., 2019b] Sarkadi, Ş., Panisson, A. R., Bordini, R. H., McBurney, P., Parsons, S., and Chapman, M. (2019b). Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4):287–302.

[Yao et al., 2017] Yao, Y., Viswanath, B., Cryan, J., Zheng, H., and Zhao, B. Y. (2017). Automated crowdturfing attacks and defenses in online review systems. *arXiv preprint arXiv:1708.08151*.