Deception

Ștefan Sarkadi
King's College London
stefan.sarkadi@kcl.ac.uk

Abstract

Recent events that revolve around fake news indicate that humans are more susceptible than ever to mental manipulation by powerful technological tools. In the future these tools may become autonomous. One crucial property of autonomous agents is their potential ability to deceive. From this research we hope to understand the potential risks and benefits of deceptive artificial agents. The method we propose to study deceptive agents is by making them interact with agents that detect deception and analyse what emerges from these interactions given multiple setups such as formalisations of scenarios inspired from historical cases of deception.

1 Deception and Thinking Machines

There is some related work in the AI literature that focuses on the issue of deception. Bridewell and Isaac define a model of dynamic belief attribution for deception in [Bridewell and Isaac, 2011]. Jones models self-deception using epistemic logic in [Jones, 2015]. Lambert defines a cognitive model of deception based on human-computer interaction in [Lambert, 1987]. Multiple studies have been done by Sakama and Caminada on formalising dishonesty in [Sakama, 2011], [Sakama *et al.*, 2014] such as deductive and abductive dishonesty, lies, bullshit, and deception. In [Sakama and Caminada, 2010] and [Sakama, 2015] they formalise multiple types of deception.

Deceptive machines first appeared as concepts in Turing's *Imitation Game*. Today it is reasonable to imagine machines that exploit human masses to extract rewards, e.g. deceiving humans into voting for an entity that the machines consider to be a necessary requirement for their success in attaining an ulterior goal. Such autonomous systems might emerge from complex social interactions that they will be able to eventually manipulate according to their will. For example, the system might use crowdturfing attacks to generate fake reviews as demonstrated in [Yao *et al.*, 2017]. Such attacks can be used to maximise the profits of entities that can afford them financially or by entities that consider that legal risks are too low and choose to employ them in spite of the established rules.

There are three main reasons for modelling artificial agents. *Firstly*, deception is fundamental to a complete theory of communication and by modelling deceptive agents we might be able to get a better understanding of how deception works. *Secondly*, intelligent machines might develop reasons to deceive. Understanding their reasoning and abilities can help us identify and prevent them from deceiving us or other artificial agents [Sakama, 2011] [Castelfranchi and Tan, 2001]. *Thirdly*, deception seems to be a necessary step in developing AI that emulates human cognition.

2 Approach and Foundations

Two main paradigms seem to be emerging within the AI community: (i) a *model-driven* paradigm and (ii) a *data-driven* paradigm. The model-driven paradigm stands for building an AI that reasons using models which contain beliefs and knowledge about the world and about other agents, in order to interpret evidence (data) and to act according to these models. The data-driven paradigm stands for building AIs that reason based on available evidence (data) without using such models.

To analyse deception from an AI perspective one must refer to beliefs and knowledge, and to include things such as goals, intentions, or desires. Security and Intelligence analysts often confront themselves with the problem of reading the intentions of and accessing the knowledge of their potential adversaries. This problem imposes cognitive limitations on building strategies to deter or counter malicious activities [Heuer, 1999]. We consider that a BDI agent architecture is able to capture the issue of other agents' intentions and we decided to use BDI as a basis for defining agents. Therefore, we choose a model-driven approach to study deceptive interactions between agents and see what might emerge from these interactions given multiple setups/scenarios inspired by carefully chosen historical and intelligence analysis cases of deception. A data-driven approach seems counter-intuitive to use in the study of deception, because such an approach would limit itself to the analysis of agents' behaviour.

Studying deception requires one to look inside systems known as black boxes, which in the case of deceptive interactions are the minds of the deceivers and their targets. Reducing the study to the inputs and outputs of these black boxes will most likely tell us nothing about things such as hidden intentions or goals. For example, imagine if, for whatever

reason, you wish to deceive your friend into thinking that you like tea, but in reality you like coffee and hate tea. You could drink tea every day in front of your friend. Your friend will start to think (rationally) that you like tea. The only data your friend has about you is the behaviour you exhibit, which is that you drink tea. Thus, the most likely explanation for you drinking tea that your friend can come up with is that you like tea. One can generate deceptive data to influence the results of a data-driven analysis. However, if your friend had access to your deceptive intentions, then your friend would have interpreted your behaviour entirely different and would have gotten the bigger picture.

According to the main theories of deception [Buller and Burgoon, 1996] [McCornack *et al.*, 2014], there are two main actors in deceptive interactions, that we define as the agents of our models. These are the *Deceiver* and the *Interrogator*. Both of these actors have different goals from each other. The goal of the Deceiver is to make the Interrogator believe something that the Deceiver thinks is false (or true) -*to deceive*. The goal of the Interrogator is to detect if the Deceiver is trying to deceive or not. These dynamics can increase in complexity if we add unknown unknowns or uncertainty. In fact, most of the times the Interrogator is simply unaware of the fact that a Deceiver might want to deceive it (Interrogator) - we call this an unknown unknown. In other situations, the Interrogator is aware of the possibility that the Deceiver might want to deceive it, but the Interrogator is uncertain if this is really the case. A necessary requirement to deceive or detect deception for artificial agents is a mental model or a *Theory of Mind* (ToM) of the other agents [Isaac and Bridewell, 2017]. This is not the case with lying, which deception is commonly confused with. Lying is usually defined as: saying that something is true (or false) when in fact that something is false (or true). On the other hand, we define deception as the intended action (or actions) of a Deceiver to make an Interrogator believe something is true (or false) that the Deceiver believes to be false (or true). A Deceiver does not need to lie in order to make an Interrogator have false beliefs. A Deceiver can tell the truth and still succeed in deceiving the Interrogator. E.g. If the Interrogator believes the Deceiver is a liar, then the Interrogator will believe the opposite of what the Deceiver says.

We started building the models of deception from a simple formalisation of information manipulation described in [McCornack *et al.*, 2014]. The Deceiver wants the Interrogator to believe that q (this counts as a desire or as an ulterior goal). Deceiver knows that $\neg p$ (counts as a belief about the world) and it also knows that the Interrogator knows that $p \rightarrow q$ (counts as a belief about another agent or as a ToM of the target). The Deceiver employs a strategy that is known as *Pars pro Toto* (parts for the whole): Deceiver tells the Interrogator that p, so that the Interrogator will conclude that q by applying Modus Ponens. When the Interrogator receives the information that p, it employs *Totum ex Parte* (the whole from the parts): Interrogator applies Modus Ponens and concludes that q. Thus, the Deceiver manages to deceive the Interrogator. In a similar manner, artificial agents might be able to execute malicious intentions in order to reach an ulterior goal. We plan to formalise models of such scenarios and see what sort of dynamics emerge.

3 Conclusions

We investigate how artificial agents can deceive and detect deception by modelling deceptive interactions. The main contributions of this study are (i) understanding and classifying multiple types of deception through formalisation and multi-agent modelling of case studies, (ii) setting the foundations of AI that deceives and of AI that detects deception, and (iii) creating methodological tools to be used by Intelligence analysts to deal with cases of deception.

References

[Bridewell and Isaac, 2011] Will Bridewell and Alistair Isaac. Recognizing deception: A model of dynamic belief attribution. In *AAAI Fall Symposium: Advances in Cognitive Systems*, 2011.

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

[Castelfranchi and Tan, 2001] Cristiano Castelfranchi and Yao-Hua Tan. *Trust and deception in virtual societies*. Springer, 2001.

[Heuer, 1999] Richards J Heuer. *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, 1999.

[Isaac and Bridewell, 2017] Alistair Isaac and Will Bridewell. *White lies on silver tongues: Why robots need to deceive (and how)*. Oxford University Press, 11 2017.

[Jones, 2015] Andrew J.I. Jones. On The Logic of Self-deception. *South American Journal of Logic*, 1:387–400, 2015.

[Lambert, 1987] D.R. Lambert. A cognitive model for exposition of human deception and counterdeception. Technical report, DTIC Document, 1987.

[McCornack *et al.*, 2014] Steven A McCornack, Kelly Morrison, Jihyun Esther Paik, Amy M Wisner, and Xun Zhu. Information manipulation theory 2: a propositional theory of deceptive discourse production. *Journal of Language and Social Psychology*, 33(4):348–377, 2014.

[Sakama and Caminada, 2010] Chiaki Sakama and Martin Caminada. The many faces of deception. *Proceedings of the Thirty Years of Nonmonotonic Reasoning (NonMon@ 30)*, 2010.

[Sakama *et al.*, 2014] Chiaki Sakama, Martin Caminada, and Andreas Herzig. A formal account of dishonesty. *Logic Journal of the IGPL*, 23(2):259–294, 2014.

[Sakama, 2011] Chiaki Sakama. Dishonest reasoning by abduction. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1063, 2011.

[Sakama, 2015] Chiaki Sakama. A formal account of deception. In *2015 AAAI Fall Symposium Series*, 2015.

[Yao *et al.*, 2017] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y Zhao. Automated crowdturfing attacks and defenses in online review systems. *arXiv preprint arXiv:1708.08151*, 2017.