Is Your AI Cheating on You?

or Reasons and Methods to Engineer Deceptive Machines

Ștefan Sarkadi, Peter McBurney, Simon Parsons, Martin Chapman, and *Matthew Moran
Department of Informatics and *Department of War Studies

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.

Motivation

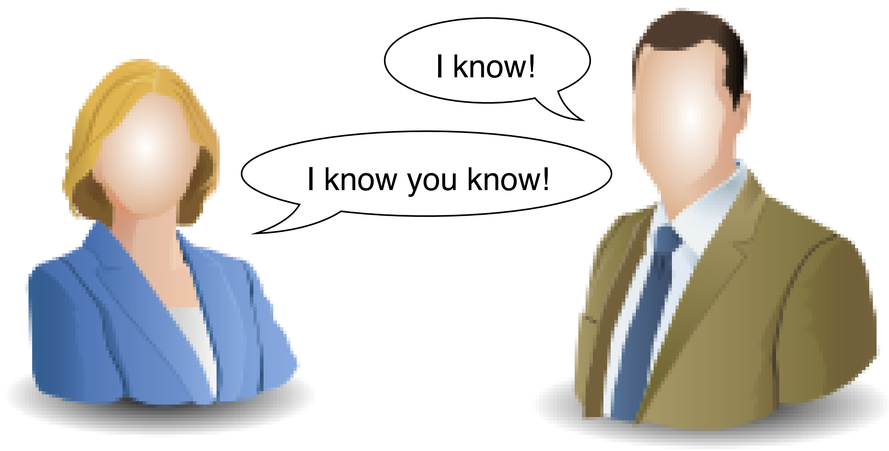
- Deception is fundamental to a complete theory of communication and by modelling deceptive agents we might be able to get a better understanding of how deception works.
- Intelligent machines might develop reasons to deceive. Understanding their reasoning and abilities can help us identify and prevent them from deceiving us or other artificial agents.
- Deception seems to be a necessary step in developing AI that emulates human cognition.

Research Questions

- Can we use AI methods to improve our understanding of deception?
- Can we engineer machines that deceive?
- How can we engineer them?

Agent Architecture

To analyse deception from an AI perspective one must refer to beliefs and knowledge, and to include things such as goals, intentions, or desires. 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. Apart from BDI, agents require *Theory of Mind* to deceive and detect deception (See Isaac & Bridewell 2017). The agents will act based on their own model of their opponent's mind.

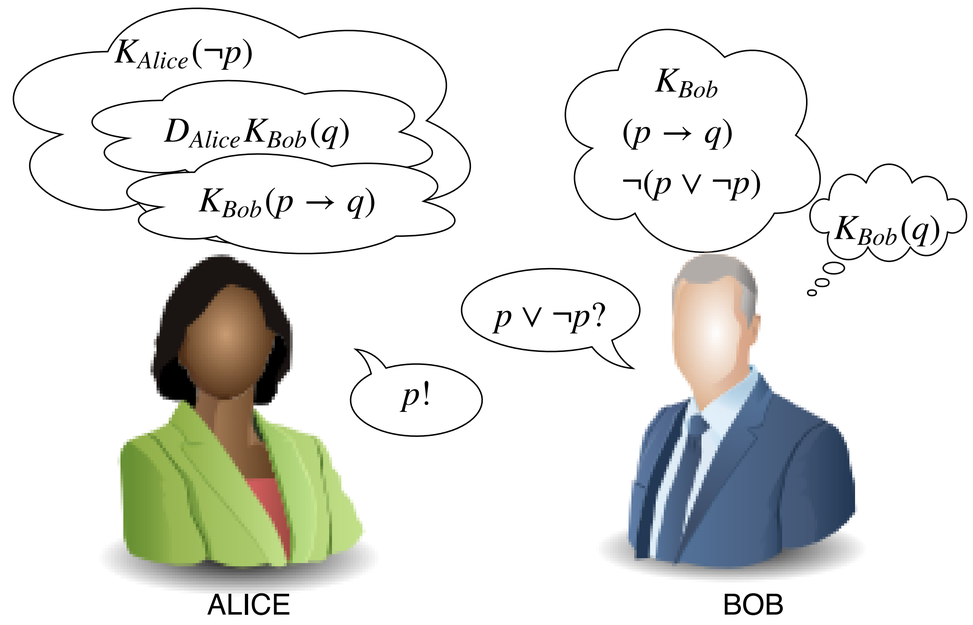


Methodology

The agent based models will be defined by integrating *Information Manipulation Theory 2* and *Interpersonal Deception Theory*. The evaluation will consist of case study analysis:

- **Case 1**: Chosen from well documented historical deceptions (WWII or Cold War). It will be used to show how well our models can explain the deceptive interactions that took place.
- **Case 2**: Chosen based on more recent events involving fake news and election tampering. Our models will be used to try to understand ongoing events with a particular degree of certainty.

Information Manipulation Theory 2 - Speech Acts

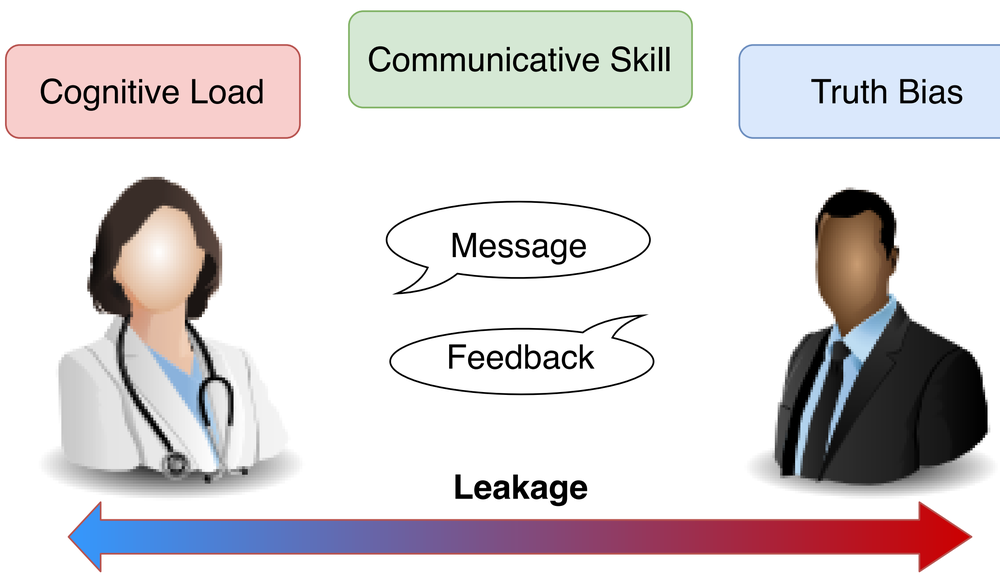


Focuses on how individuals manipulate information in order to deceive (See McCornack 2014). Agents engage cognitive processes that determine speech acts:

- *Pars pro Toto* (parts for the whole): Alice estimates that if she tells Bob that p, then Bob will conclude that q by applying Modus Ponens.
- *Totum ex Parte* (the whole from the parts): After Alice tells Bob that p, Bob applies Modus Ponens and concludes that q.

Interpersonal Deception Theory

Focuses on parameters that influence deceptive social interactions such as *Cognitive Load*, *Communicative Skill*, *Truth Bias*, and *Leakage* (See Buller & Burgoon 1996).



References

1. S. Sarkadi, **Deception**, IJCAI 2018, Stockholm 2018.
2. A.R. Panisson, Ș. Sarkadi, P. McBurney, S. Parsons & R.H. Bordini, **Lies, Bullshit, and Deception in Agent-Oriented Programming Languages**,20th International TRUST Workshop, Stockholm 2018.
3. Isaac & Bridewell, **White Lies on Silver Tongues**, *Robot Ethics 2.0.* ed. / P. Lin; K. Abney; R.R. Jenkins. OUP, 2017.
4. S.A. McCornack et al., **Information Manipulation Theory 2**, *Journal of Language and Social Psychology*, 2014.
5. Buller & Burgoon, **Interpersonal Deception Theory**, *Communication Theory*, 1996.