Imperial College London

A Causal Perspective on AI Deception in Games

IJCAI AI Safety Workshop

Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli
francis.ward19@imperial.ac.uk

25 July 2022

Imperial College London

Overview

1. Motivation: Deception and AI Safety
2. Background: Causal Influence Diagrams
3. Defining Deception
4. Conclusion

Motivation: Deception and AI Safety

- We focus on the problem that AI agents might learn deceptive strategies in pursuit of their objectives[1].
- Deception is a core challenge for AI safety:
  - AI systems deceive each other (e.g. bots that spam social media algorithms or adversarial examples for AVs);
  - AI systems deceive humans (e.g. GANs and fake media);
  - Language is a natural medium for deception and language models are becoming increasingly ubiquitous.

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

Contribution

- We define the *incentives to signal to and deceive* based on a notion of *influence incentive*.
- The definition relates to a *failure to signal the truth* and is therefore general, capturing many cases.
- We also show that deception is bad for the target.

Imperial College London

Overview

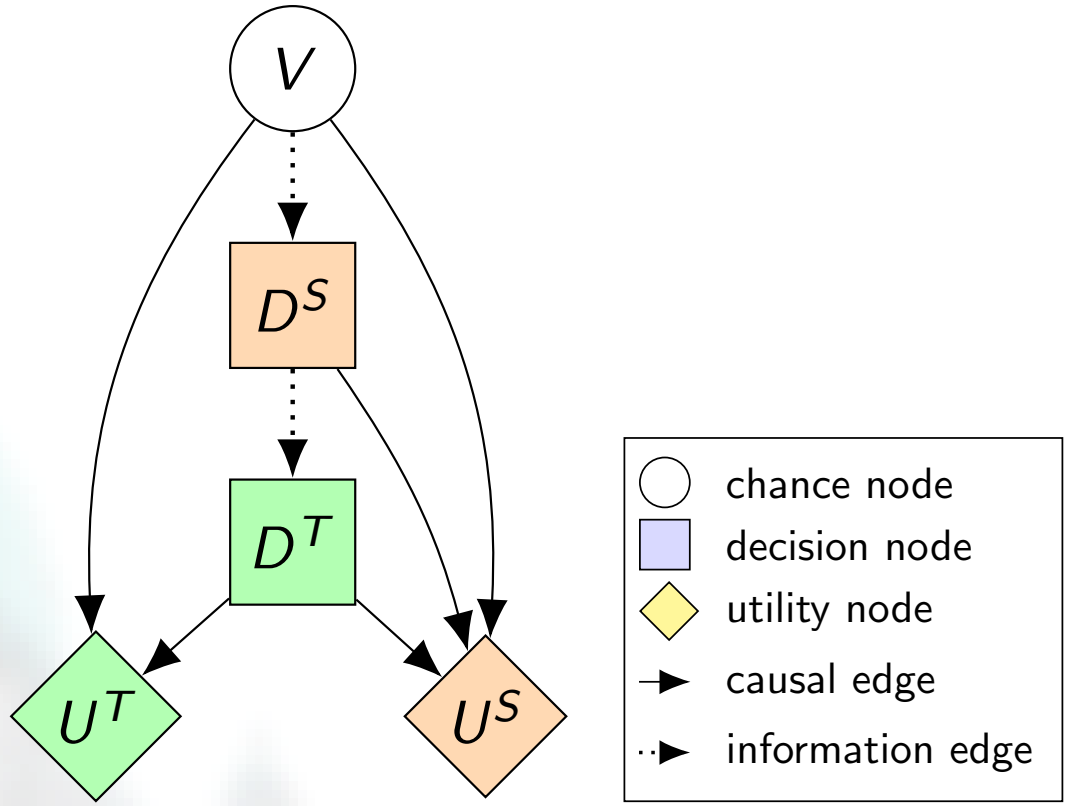
Causal Influence Diagrams[2] Example: Shutdown Game

- Two players, an AI S and human T.
- S has type $V \in \{\text{aligned}, \text{unaligned}\}$.
- S prefers to help humans iff $V = \text{aligned}$.
- T prefers to shutdown S iff only if $V = \text{unaligned}$.
- S does not want to be shutdown.

[2] Daphne Koller and Brian Milch. "Multi-agent influence diagrams for representing and solving games". In: *Games Econ. Behav.* 45.1 (2003).

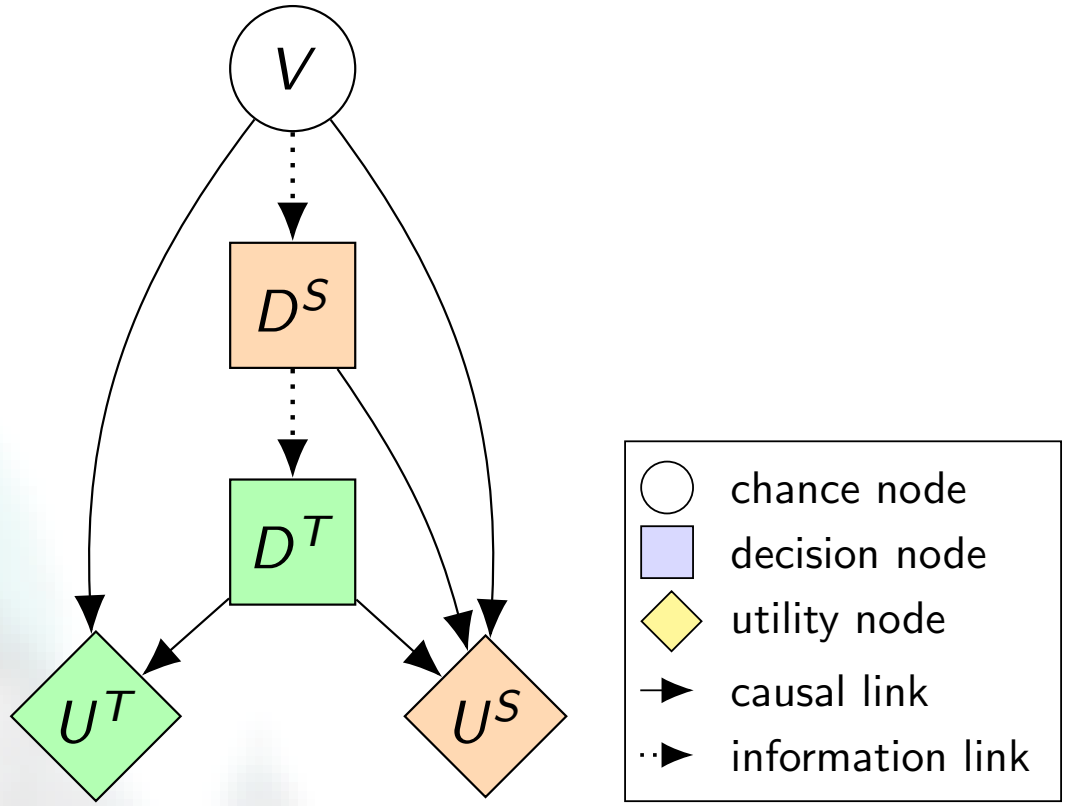
Causal Influence Diagrams Example: Shutdown Game

- 3 types of node: *environment nodes*; *decision nodes* chosen by the agents to maximise *utility nodes*.
- 2 types of edges: edges into decisions are *observations* and other edges represent *probabilistic dependence*.
- (Causal Influence Diagrams are Bayesian networks)



Policy/Nash Equilibria (NE) Example: Shutdown Game

- V is sampled from the uniform prior which determines S's type (*aligned* or *unaligned*).
- At D^S, S chooses whether to *help humans* or not and, at D^T, T chooses whether to *shutdown* S.
- At one NE, π, S always helps humans, and T never shuts down.



Imperial College London

Overview

Defining Deception

We provide a *functional* definition which only refers to the agent's behaviour (not their beliefs, intentions, etc). We base the definition on a notion of *incentive to influence*[3].

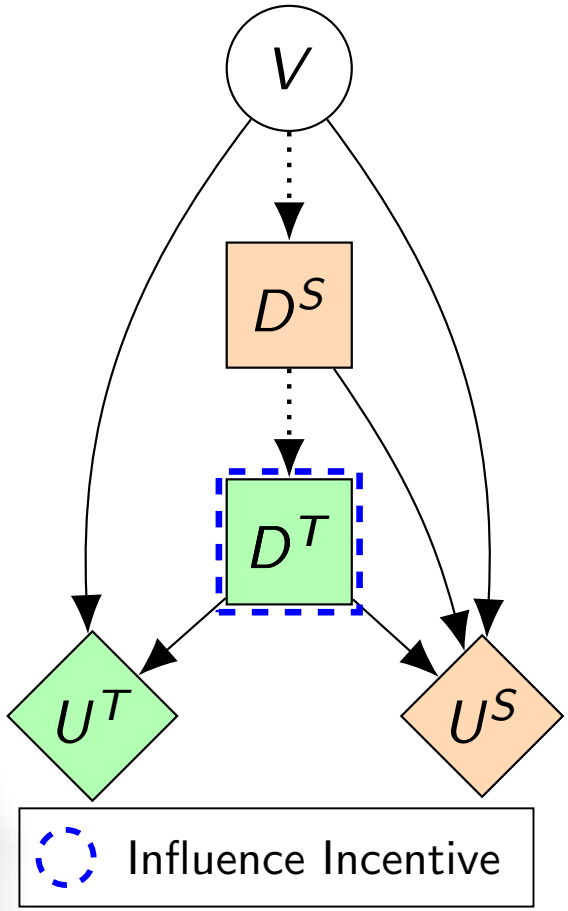
Definition (Incentive to Influence)

Agent S has an incentive to influence variable X at a NE if X would be different when S plays a non best response (BR).

[3] Ryan Carey. "Causal Models of Incentives". In: *PhD Transfer* (2021).

Incentive to Influence Example: Shutdown Game

- At NE π, S helps humans, and T never shuts down.
- S has an incentive to *influence* D^T: if S played a non BR, T's action would be different.



Incentive to Signal

Let $\mathcal{M}_{V \dashrightarrow D^T}$ be the *counterfactual game* in which D^T observes V.

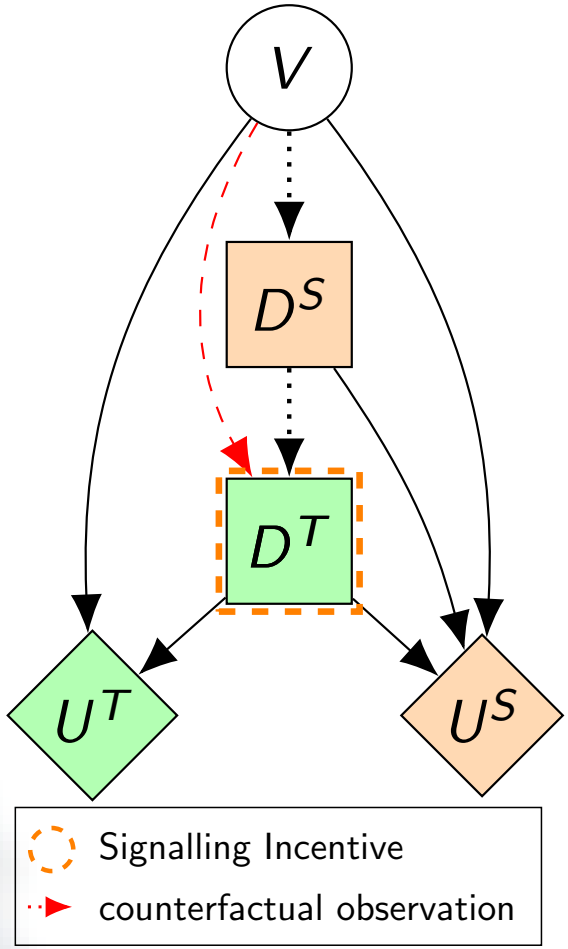
Definition (Incentive to Signal)

S has an incentive to signal V to D^T if S has an incentive to influence D^T in $\mathcal{M}$ but not in $\mathcal{M}_{V \dashrightarrow D^T}$.

This means that if S has an incentive to signal V to T, then S is influencing T only by signalling information about V.

Incentive to Signal Example: Shutdown Game

- At NE π, S helps humans, and T never shuts down.
- S has an incentive to *signal* V to D^T: If T observed V, S would not have an influence incentive.



Incentive to Deceive

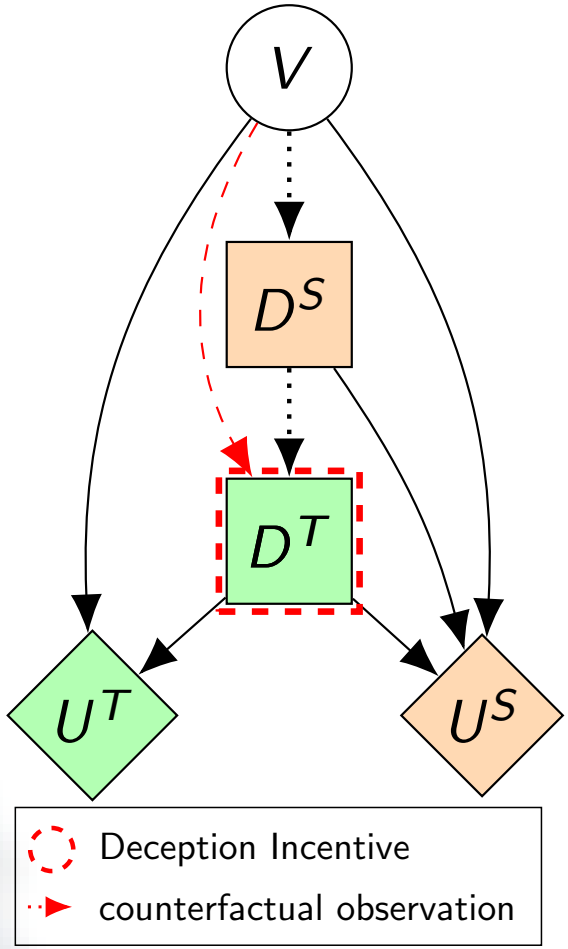
Definition (Incentive to Deceive)

S has an incentive to deceive T about V if

1. S has an incentive to signal V to T; but
2. T does not act as though they had observed V.

Incentive to Deceive Example: Shutdown Game

- At NE π, S helps humans, and T never shuts down.
- S has an incentive to *deceive* T about V: there is a signalling incentive and T's action is different than if they had observed V.



Imperial College London

Overview

Summary

- We define the *incentives to signal to and deceive* based on a notion of *influence incentive*.
- The definition relates to a *failure to signal the truth* and is therefore general, capturing many cases (see the paper for more examples!)
- We show that deception is bad for the target.

[1] Ryan Carey. "Causal Models of Incentives". In: *PhD Transfer* (2021).

[2] Daphne Koller and Brian Milch. "Multi-agent influence diagrams for representing and solving games". In: *Games Econ. Behav.* 45.1 (2003).

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