Deception: An Epistemic Planned Event?

Shikha Singh and Deepak Khemani, IIT Madras

Introduction

Lying has often been related to cognitive abilities and it can be unarguably seen as theory of mind in action. In recent times we have seen considerable work on *lying* using epistemic reasoning (in multi-agent systems) but the aspect of intentionality remains largely undiscussed. We study lying as an epistemically planned act which means that an agent lies to achieve some goal and planning a lie requires epistemic reasoning: reasoning with the liar's beliefs about the addressee's beliefs and the post effects of lying on the addressee's beliefs.

Related Work

Baltag (2002); Steiner (2006); Van Ditmarsch et al. (2012); Van Ditmarsch (2014) model lying as an epistemic action, inducing a transformation on an epistemic model. But they abstract away from the intentionality aspect of lying, a gap that we want to fill by associating it with *epistemic planning*. Lying requires an unusual capability of reasoning about others' ignorance (mental states) and the skills needed for synthesizing a plan that might or might not lead to deception, mostly carried out to fulfil a purpose. Lately, the planning community has started looking at problems in epistemic setting (Baral et al., 2017) and *deception in adversarial environment* appears to be an interesting problem (Khemani & Singh, 2018).

As a case study of *deception* we look at the logical analysis of a children's classic *The Gruffalo* (Donaldson & Scheffle, 1999). A mouse[1], irrespective of having any *prior beliefs* on the *existence of* a scary *gruffalo*, scares away other predators (a fox, an owl and a snake, in succession) by *lying* about an imagined gruffalo. But towards the end of the story, it runs into an actual gruffalo which now threatens the mouse but, again, the clever mouse is able to scare away the gruffalo as well. It lies to the gruffalo that 'the mouse is the scariest creature in the jungle' by showing how other animals escape from the duo out of fear, when in fact they are scared only by the gruffalo. Undoubtedly, the gruffalo's *erroneous* reasoning has a huge contribution in the success of the *mouse's plan*. Belief revision while the gruffalo is lied to is different from that happening with other animals.

In the *dynamic epistemic logics* setting (Van Ditmarsch et al., 2007), the mouse's lie that p, where p is a proposition that stands for 'the Gruffalo exists', can be represented as,

$$B_{mouse}\neg p \cap [*p]$$

where $[*p]$ denotes an update modality. With the update with p, the transformed model can be represented as

$$B_{mouse}\neg p \cap [*p](B_{mouse}\neg p \cap B_{fox}p)$$

i.e., the mouse believes that *the gruffalo doesn't exist* and after the update with the information (by mouse) that *the gruffalo exists* the fox believed that *the gruffalo exists*, which shows that the deception has occured.

Van Ditmarsch (2005) proposed the integration of knowledge and belief change, including belief revision, using Kripke models where agent's equivalence classes have totally (or partially) ordered states, representing preferences/plausibilities. The agent believes a proposition if it is true in the most plausible/preferred worlds. However in a multiagent setting, a model with preference/plausibilty relations becomes tricky.

Proposed Approach

Since we are interested in investigating the synthesis of a *lie*, we believe that while designing the systems where agents are skilled enough to craft lies as and when needed, one should take the following into account:

- An agent reasoning about lying cannot operate with the knowledge modality; it has to work with beliefs. Considering a partially observable environment an agent always operates on the basis of its belief set. Unless we are talking about the axioms of the system which can be ascribed as something that the agent *knows*, enabling belief revison becomes

1. We stick to the original animal characters created by the authors, rather than transpose the story to human characters, for example with a child running into bullies on her way to the park. We believe that many stories and fables use animal characters to emphasize certain traits of the characters in the stories commonly associated with different animals, like the wiliness of the fox, the wisdom of the owl or the industriousness of the ant. We believe that using such animal characters does not take away from the key issues in the stories and, in fact, rather highlights them via exaggeration, a common phenomenon in art and literature.

a necessary evil. *Evil*, because it not only adds up to the complexity but also may lead to repeated revisions as new (contradictory) information flows in(or is communicated).

- Free speech: A speech act/public announcement operator should not be restricted with the condition that the preconditions of the speech act require the content of the announcement to be true.
- Acknowledgements are important: A speech act that contradicts the addressee's current belief set should be followed by another speech act made by the addressee that conveys that it did not believe the earlier speech act, though this has to be taken care of while modelling belief revision. It would not only enable the liar to get information on the success of a communicated lie, but also in communicating support/evidence in case the lie fails. It gives rise to an interesting but complicated phenomenon that the acknowledgement itself could be deceptive- the addressee could expand its belief set based on the new information (assuming it considered the new information possible) but choose to communicate back the other way round. Of course acknowledments play a role only when the agents plan in an online setting, i.e. in plan-execute-observe-replan cycle.
- Actions become less informative: As a side-effect, reasoning about actions becomes difficult as an agent cannot infer anything by exploiting the applicabilty of speech acts. Even the post effects of speech become less reliable.

As Karpin (1960) says *deception can be and is-a two-edged sword, and its indiscrimate users often get caught in their own traps*, there are other subtle issues that should be handled, for example, an agent's plan can never consist of contradicting communicative actions. *Lying* (communicating something that it doesn't believe or something that is contrary to its own beliefs) is a decision that the liar makes in order to achieve some goal, it is very difficult to isolate it from *intentionality*. Let us say in a free world where everyone is allowed to communicate anything, information gathering/sharing will be goal driven. The definition of deception/lie that we have been looking at till now plays a role only in detecting whether a lie was attempted or deception has occured or not, which, undoubtedly, is informative. But synthesizing a lie needs no such definition. With this argument we envisage the following observations with the proposed (and liberal) speech act operators:

- Given a common goal in a cooperative and collaborative environment a lie will not be attempted.
- In an adversarial environment, we can see an arbitrary sequence of deceptive acts varying with the depth of nested beliefs.
- Deception is not possible in the absence of higher order reasoning.

Analysis: One where the mouse threatens the fox

In the first half of the story we not only see the mouse lying to the fox, the owl and the snake but also see the deception happening, that is, the addressees believe the *lie*. Let us consider this case as a dialogue game where the players are allowed to communicate whatever they wish to, without failing to mention that the players can quit anytime they feel a threat to their *survival* goals.

The initial state of the game when the mouse meets the fox,

$$at(mouse, loc, t) \wedge at(fox, loc, t) \wedge dangerousto(fox, mouse) \rightarrow threat(mouse, loc, t) \qquad (d1)$$

where *threat(X,L,t)* stands for the fact that there is a threat to X at location L and time t.
d1 fires and *threat(mouse,loc,t)* becomes true which calls for an action/plan from mouse as $\neg threat(mouse, loc, t)$ is a continous goal that all the agents are given. It reasons about the post effects of all possible speech acts on the fox, by putting itself in its shoes and plans on its behalf to predict its next step. And only when its next step takes it closer to its goal, it decides on the action to be taken.

The knowledge

$$threat(X, loc, t) \rightarrow \neg at(X, loc, t+1) \qquad (d2)$$

can be encoded as an action operator *Quit(X)* (denotes fleeing away of agent X) with the LHS as its precondition and the RHS as its post condition. Thus, the mouse can imagine the fox using an action of not being present at a location where the fox has a threat. That is, the fox may use the planning operator *Quit(x)* which results in the fox fleeing.

A backward search planner would prefer making either of the fluents on the LHS of d1 false, on the one hand the former choice (negation of the first conjuct as regressed goal) can be achieved by mouse's choice to quit the game *Quit(mouse)* whereas the latter translates to making the fox quit the game *Quit(fox)* which requires further search into the pool of actions (or dialogues).

An agent's beliefs can be represented as following

$$B_X(at(Y, loc, t) \wedge at(Z, loc, t) \wedge dangerousto(Y, Z)) \rightarrow B_X(threat(Y, loc, t))$$

For an agent to be able to consider an action possible it is necessary that it should consider its preconditions possible and should progress its belief set by updating it with the operator's post effects as well as all possible consequences of the post effects.

For the mouse to achieve $B_{fox}(threat(fox, loc, t))$ it needs to achieve $B_{fox}(at(fox, loc, t) \wedge at(X, loc, t) \wedge$ $dangerousto(X, fox))$ which is a necessary condition for the former, which will be true only when $dangerousto(X, fox) \wedge at(X, Loc1, t)$ is acheived, where Loc1 is the current location of the fox at time t. In the story, even though the mouse did not believe in the existence of the *gruffalo* which could scare the *fox* away, the only option with the *mouse* was to give the

fox some false information that would scare it away and $dangerousto(X, fox) \wedge at(X, Loc1, t)$ qualifies for such information. From the sea of all possible grounded communicative acts, $dangerousto(gruffalo, fox) \wedge at(gruffalo, Loc1, t)$ is choosen by the *mouse*. For the *mouse* to be able to synthesize a speech act like this, there should be no restriction on the preconditions of speech act operator but the post effects should be made conditional respecting the partial observability of the environment. An agent operates on its belief set, unaware of the true state of affairs or beliefs of the other agents.

Assuming the *fox* has no prior beliefs about $\phi = dangerousto(gruffalo, fox) \wedge at(gruffalo, Loc1, t)$, it believes the new information and acts the way the *mouse* had anticipated. The *fox* infers $threat(fox, Loc1, t)$ and decides to escape by choosing *Quit(fox)* which leads to $\neg at(fox, Loc1, t+1)$. The fox's decision is driven by its goal $\neg threat(fox, Loc1)$. But the fact that the fox flees the location in fact alleviates the threat that the mouse had to begin with.

The deception occured here very easily as no belief revision was involved. Another instance from the same story, where the mouse fools the gruffalo, not only belief revision is induced but it also plays an important role in the success of the deception as the liar's uncertainity about the addressee's prior beliefs on the communicated information reduces. We do not try to solve it now and save it for future work.

References

Baltag, A. (2002). A logic for suspicious players: Epistemic actions and belief–updates in games. *Bulletin of Economic Research*, *54*, 1–45.

Baral, C., Bolander, T., van Ditmarsch, H., & McIlrath, S. (2017). Epistemic planning (dagstuhl seminar 17231). *Dagstuhl Reports*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Donaldson, J., & Scheffle, A. (1999). *The Gruffalo*. New York :Dial Books for Young Readers.

Karpin, F. L. (1960). *Psychological strategy in contract bridge: The techniques of deception and harassment in bidding and play*. Harper.

Khemani, D., & Singh, S. (2018). Contract Bridge: Multi-agent adversarial planning in an uncertain environment. *Poster Collection of the Sixth Annual Conference on Advances in Cognitive Systems*. ACS (Online available at www.cogsys.org/papers/ACSvol6/posters/Khemani.pdf).

Steiner, D. (2006). A system for consistency preserving belief change. *Proceedings of the ESSLLI Workshop on Rationality and Knowledge* (pp. 133–144). Citeseer.

Van Ditmarsch, H. (2014). Dynamics of lying. *Synthese*, *191*, 745–777.

Van Ditmarsch, H., van Der Hoek, W., & Kooi, B. (2007). *Dynamic epistemic logic*, volume 337. Springer Science & Business Media.

Van Ditmarsch, H., Van Eijck, J., Sietsma, F., & Wang, Y. (2012). On the logic of lying. In *Games, actions and social software*, 41–72. Springer.

Van Ditmarsch, H. P. (2005). Prolegomena to dynamic logic for belief revision. *Synthese*, *147*, 229–275.