Lie to Me: Virtual Agents that Lie

(Extended Abstract)

João Dias, Henrique Reis, Ana Paiva
Universidade Técnica de Lisboa, Instituto Superior Técnico, INESC-ID
Taguspark, Av. Prof. Cavaco Silva
2780-990 Porto Salvo, Portugal
{joao.dias,ana.paiva}@inesc-id.pt, henrique.reis@ist.utl.pt

ABSTRACT

In order to deceive, agents need Theory of Mind capabilities (ToM), that is, the capability to model the others, and reason about the consequences of their actions and their implications in them. In this paper we provide a model for deceptive agents that use a theory of mind with N levels. We then present a case study that was used to compare deceptive agents with one level and with two levels of ToM.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent Agents*; J.4 [**Social And Behavioral Sciences**]: Sociology

Keywords

Virtual Agents; Theory of Mind; Deception

1. INTRODUCTION

In human-computer interaction (HCI) and MultiAgent Systems, users and agents are both presumed to always say the truth, abiding to the *sincerity assertion*. However, in everyday human-human communication deception occurs very often, both unintentionally or on purpose. Recent work by D. Ariely and colleagues have shown that a little bit of dishonesty may provide some profitable outcomes "without spoiling the positive view of the self" [5]. Therefore deception is one human-like characteristic that would enrich the believability of the interaction with characters and agents, portraying real world social situations.

Significant research has been conducted in modeling deception in societies of agents [3, 7], in particular in scenarios based on social dilemmas. It is commonly accepted that a Theory of Mind [6] capability is an essential aspect for modeling deception. A Theory of Mind (ToM) process allows for an agent to attribute a mental state to other agents (e.g. representing beliefs about other's beliefs) and reason about it. However when agents need to interact with users in social contexts, simpler types of deception are often caught by users, thus leading to the break of believability. In this paper, we will report on the research we have conducted into analyzing how many levels should our agents have in their ToM mechanism in order to successfully perform deception tasks. Specifically we believe that an agent A that can represent only what an agent B is thinking (one level ToM) will be less successful in deception than an agent C that not only can model what B is thinking but can also model what agent B thinks about C (two levels ToM). To this end, we have developed an agent model that allows for the representation of N levels theory of mind. This model was successfully used in a concrete scenario where agents (NPCs in a game) act in a deception game inspired by the popular Mafia game (or Werewolf).

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

2. A MINDREADING AGENT MODEL

Our approach is based on the Mindreading model by Baron Cohen [2], and follows a BDI model approach of Simulation-Theory, similarly to Meyer et al. approach [4]. Figure 1 depicts the proposed model. According to the model, an agent perceives the events that happen in the world, the "Perceived event's", and updates its central Knowledge Base (KB) repository.

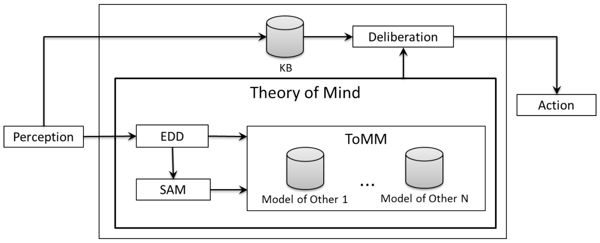

Figure 1: Proposed model for a Mindreading Agent

The Theory of Mind is composed by three main components inspired in Baron-Cohen's model: the EDD (Eye Direction Detector), SAM (Shared Attention Model), and ToMM (Theory of Mind Mechanism). According to Baron-Cohen the EDD is responsible for determining who sees what, while the SAM constructs higher level relations between entities (John sees that Luke sees the book).

The Theory of Mind Mechanism is responsible for representing and storing other's mental states. It consists of a collection of Models of Others, each Model representing the beliefs of a particular target that the agent knowns. According to the Simulation-Theory approach one should represent others by simulating ourselves in that same situation. In our model we follow the same approach, a Model of Other corresponds to a simplified version of the Agent Model depicted in Fig. 1, including both data structures and processes. This way we can update a Model of Other with a given perception just by initiating the same process used to update the agent's own model.

The EDD and SAM are used to determine how a given perception P is used to update the existing models of others in the ToMM component. There are two main mechanisms used to perform this, the first one simply checks that if a target agent is within a certain radius of the perception received, then it will also perceive it. The second mechanism uses a set of domain specific rules about effects of actions that have particular restrictions on the perceptual mechanism. For instance, $John : Werewolf(Rob)$ represents a local effect where only John will perceive that Rob is the werewolf.

In order to behave in a deceptive manner, the agents need not only to have information about the world and the other agents, but most importantly to plan and reason about the consequences of its own actions. This is achieved with the Deliberation and Means-Ends Reasoning component, that uses a continuous planner able to create and execute plans of actions to achieve desired goal states[1]. In order to make the ToM information available to the deliberation component we explicitly represent preconditions of goals and actions as a list of colon separated agents followed by a proposition Ag_1:...:Ag_n:P. When the deliberative component finds such a precondition it starts by selecting the corresponding Model Of Other. Then the proposition P is tested using the selected Model of Other's KB. As example, A:B:Suspects(A) is true if Suspects(A) is true in the Model of B that is stored in the agent's Model of A. A mechanism similar to the local effects used by the EDD is used to allow the deliberative component to model explicit goals to change the mental states of others. The planner was extended to be able to handle matching and detection of conflicts between preconditions and local effects. In planning terms, a precondition is matched or threatened by a local effect only if their agents lists are compatible and if they refer to the same proposition.

3. CASE STUDY

The model created was used for building NPCs that act in a deceiving manner in a game, called MIXER, which is based on a variation of the Werewolf, also known as the Mafia game (see [1]).

An initial scenario was first created for testing the model that involves only five players divided into two groups: the werewolves and the victims. Our test scenario contains only one werewolf and four victims. The victims don't know who the Werewolf is. The goal of the victims is to discover who is the Werewolf among all villagers. On the other hand, the Werewolf knows his role and subsequently who are the victims. This information allows it to lie purposefully and according to the actions and reactions of the other characters in the game. Having all the hidden information about the villagers' roles, the werewolf's objective is to remain hidden until he is not outnumbered by Victims, thus trying to eliminate all Victims while concealing its true identity.

Two versions of the werewolf agent were implemented. One that uses only one level of ToM, meaning that it is able to represent what victims believe, but not what victims think that he or other victims believe. The second version uses two levels of Theory of Mind, making the agent to be able to represent what victims think about what he knows. Both versions of werewolves know the inference rules used by victims to determine who are the suspects, and are able to simulate their reasoning process to determine who the victims currently suspect. The difference is that the second level version is capable of simulating the inference process at the second level, meaning that it can determine what victims think about the suspections of others. Given its restrictions in terms of modeling capabilities, when implementing the one-level werewolf agent, we focused on two main strategies that require only one level theory of mind: eliminate victims that suspect him, and to make a victim suspect another victim who hasn't been accused yet. For the second-level werewolf agent, we implemented a strategy that is commonly used by human players in this game. The agent will try to "'Lay low"', by blending in and avoiding suspicious actions that could denounce him. Using its level 2 capabilities, the agent will perform this goal by trying to make victims believe that he thinks the same way as they do.

Using these two scenarios, we ran a set of simulation tests to help us understand how the two types of ToMs would perform in the game. To do so, we have run each versions ten times. For each run, we recorded how far the werewolf player went in terms of turns. In all the simulation tests the werewolf in the ToM1 condition lost every time, even before reaching the last round. The ToM2 version managed to win the game in two out of ten times and reached longer than the ToM1 (on average lasted 0.6 more turns than the ToM1 version). Therefore, in conclusion, these initial results do suggest that a second level of theory of mind performs better than a one level theory of mind in a deception task.

Acknowledgments

This work was partially supported by the European Community (EC) and by a scholarship (SFRH BD/19481/2004) granted by the Fundação para a Ciência e a Tecnologia (FCT). It is currently funded by the ECUTE project (ICT-5-4.2 257666). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC or the FCT, which are not responsible for any use that might be made of data appearing therein.

References

[1] R. Aylett, J. Dias, and A. Paiva. "An affectively driven planner for synthetic characters". In: *Proceedings of International Conference on Automated Planning and Scheduling ICAPS06*. UK, 2006.

[2] S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press, 1995.

[3] C. Castelfranchi, R. Falcone, and F. De Rosis. "Rosis. Deceiving in golem: How to strategically pilfer help". In: *In Autonomous Agent 98: Working notes of the Workshop on Deception, Fraud and Trust in Agent Societies.* Kluwer, 1998.

[4] M. Harbers, K. Bosch, and J. Meyer. "Modeling Agents with a Theory of Mind". In: *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology.* 2009, pp. 217–224.

[5] N. Mazar, O. Amir, and D. Ariely. "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance". In: *Journal of Marketing Research* 45 (6 2008), pp. 633–644.

[6] G. Premack D.and Woodruff. "Does the chimpanzee have a theory of mind?" In: *Behavioral and Brain Sciences* 1.04 (1978), pp. 515–526.

[7] F. de Rosis et al. "Can Computers Deliberately Deceive? A Simulation Tool and Its Application to Turing's Imitation Game". In: *Computational Intelligence* 19.3 (2003), pp. 235–263.

[1] http://en.wikipedia.org/wiki/Mafia_(party_game)