Strategically Misleading the User: Building a Deceptive Virtual Suspect

(Extended Abstract)

Diogo Rato
INESC-ID & Instituto Superior Técnico
Universidade de Lisboa, Portugal
diogo.rato@gaips.inesc-id.pt

Brian Ravenet
INESC-ID
brian.ravenet@gaips.inesc-id.pt

Rui Prada
INESC-ID & Instituto Superior Técnico
Universidade de Lisboa, Portugal
rui.prada@gaips.inesc-id.pt

Ana Paiva
INESC-ID & Instituto Superior Técnico
Universidade de Lisboa, Portugal
ana.paiva@inesc-id.pt

ABSTRACT

Humans lie every day, from the least harmful lies to the most impactful ones. Therefore, in an attempt to design virtual agents endowed with advanced decision-making abilities, researchers not only focused their effort in designing cooperative and truthful agents but also deceptive and lying ones. In this paper we propose a model capable of engaging an agent in an uncooperative misleading dialogue with a user. This model gives to an agent the ability to reason about its knowledge and then autonomously adjust the story it tells depending on what its interlocutors might know and on how sensitive it considers the conversation topic to be. Such a model allows a story's author to focus on the main narrative, letting the model handle the generation of alternatives. We implemented the model in an agent called the Deceptive Virtual Suspect and conducted some preliminary experiments using an Interrogation Game.

Keywords

Adaptive Storytelling; Human-Agent Interaction; Deceptive Communication; Autonomous Agents

1. INTRODUCTION

As humans, we use deception daily as a tool to cope with conflicting situations in a variety of contexts [4] [3]. As a result of being so frequently used in conversations, deceptions could be incorporated in virtual agents to make them more believable, mimicking how humans choose to tell the truth or to lie. However, despite its extensive usage in our dialogues, lying is not a straightforward task. It requires a higher effort than telling the truth [14] [15], since, in order to inhibit cues that could expose it, the liar needs to mentally keep track of the lies in parallel with the real events [19]. Moreover, little is known about the cognition of deception [6]. Walczyk et al. proposed a model to describe the cognitive process used to produce lies [17], later redefined for high stakes [16], that defines it as a process with the following steps: the truthful memories are activated, a decision is made whether there is a need to change the information shared, a deception is constructed and the deceptive statement is shared.

Recently, some researchers have been investigating the development of virtual agents endowed with deceptive techniques applied in negotiations. In environments that require resource and information sharing to serve their self-interest [2], both parties involved in the negotiation need to maintain a sense of fairness within the other to help them achieve their goals [7]. It is also important to balance the amount of irrelevant and relevant information given to an interlocutor in order to ensure that he will not start to ignore it [8].

In our work, we want to explore the process of sharing a story and how to automatically build alternatives in order to hide true and compromising information. To deceive others about its previous actions, our virtual agent needs to represent a narrative in its memory and should have the inference capabilities needed to autonomously adjust it as it goes along. One memory structure used to computationally represent stories is the Episodic Memory [13]. It has been used by virtual agents capable of recalling their past experiences and beliefs [5], but also by artificial companions to model shared memories gathered from a dialogue with users [1]. In addition to the story representation, the model must be able to adapt its content and the field of interactive storytelling shares similar challenges to our model's objective. In this literature, we found some contributions regarding the story manipulation based on anticipated user actions [10] and previously scripted alternative stories [11].

Overall, despite sharing some goals and challenges of previous research, our work focuses on designing a virtual agent's model that is able to autonomously create, adjust and share false information about its past actions guaranteeing it is consistent with the representation of the interlocutor's knowledge. Additionally, along with enhancing the social interaction skills of our agents, our approach intends to reduce the authoring process associated with the creation of scripted alternatives to deal with every possible user's action, allowing the story's author to focus on the development of the main narrative.

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

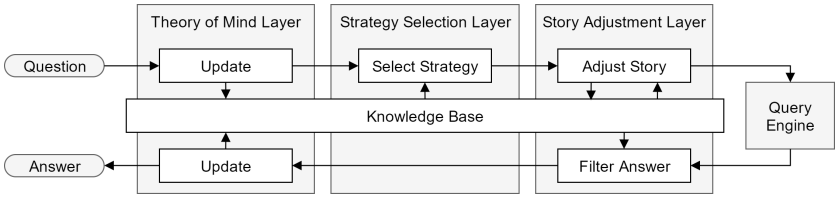

Figure 1: Deceptive Virtual Agent's two pass control layered architecture

2. DECEPTIVE AGENT MECHANISM

The model's main element is its story and its structure is based on the episodic memory theory [13]. Each memory fragment describes a time period of the agent's life, which are called *events* and they contain multiple *entities*. An *entity* represents a concept within the model's domain that can be referenced by different events. An *event* can associate multiple entities in different fields, such as *Time*, *Location* or *Agent*. Additionally, an *event* is also characterized by two fields that will be used in the deceptive mechanism: one to describe if it really occurred or not - *real* - and a percentage defined by the story's author to indicate how compromising it is - *sensitive*. A *story* is a collection of *events* and, in order to share the events between each alternative *story*, the same event can be referenced by multiple stories.

An agent equipped with our model interacts with an interlocutor using a turn based system. The question asked is based on the two major classes for question classification, *yes-no* and *wh*-questions [9] [18] [12], and contains a list of conditions that will be used to query the system. The answer shared by the agent has a set of entities that match the question conditions.

The virtual agent's proposed model implements this deception mechanism using a vertical layered architecture with its core being the **Query Engine**. In conjunction with this engine, the three layers of the architecture, the **Theory of Mind Layer**, the **Strategy Selection Layer**, and the **Story Adjustment Layer**, support the real-time modification of the story during the interaction. Each layer has its own functionality, but they all use a common **Knowledge Base** where the stories and information about each user are stored. Figure 1 shows the different layers, the **Query Engine** and their interactions with the **Knowledge Base**.

The **Knowledge Base** manages the different stories, real and parallels, along with the representation of the interlocutors' knowledge. The real story represents the original version of the agent's past actions and it won't change during the interaction. On the other hand, parallel stories can change: new events will be created, replaced and changed.

The interlocutors' knowledge is accessed by the **Theory of Mind Layer**. Based on the questions and answers, this layer registers the agent's beliefs about what each interlocutor knows of the real story and its particular alternative story. This registry is then used during the strategy selection to guarantee the external consistency of the lies created.

Using the question asked, the representation of others' knowledge and the sensitive value of each event and entities, the **Strategy Selection Layer** selects the best strategy for the current state of the interaction. The environment and interaction state are evaluated and the agent decides to deceive - *Lie* - or not - *Don't Lie*. If the strategy *Don't Lie* is chosen, the agent can either *Hide* the compromising information or share the real story. However, if the agent decides to lie, different methods can be used to apply this strategy: *Duplicate Event* (duplicate an entire event), *Adjust Event* (change the compromising event's fields) and *Adjust Entity* (change all the occurrences of an entity in the story).

The three methods of the *Lie* strategy rely on the replacement of entities that should not be shared with less compromising ones. To find suitable replacements, the model uses the entities retrieved by SimilarEntity. This procedure searches on its **Knowledge Base** for possible alternatives and ranks the candidates based on three heuristics: how sensitive they are, how similar their context is with the original one, and if the agent believes the interlocutor already knows them. The layer responsible for changing the story's content or the answer shared is the **Story Adjustment Layer**.

Let's consider as compromising an event that occurred moments before a murder was committed and involves the agent and the murderer. When asked "Who was with you before the murder?", the agent must conceal what happened. To avoid incriminating itself, the agent must find a replacement to mislead the interlocutor. Assuming the **Knowledge Base** does not have any belief about the interlocutor's knowledge, the model chooses to *Lie* using the *Duplicate Event* method. The real event is copied and all the compromising fields are replaced with less incriminatory entities.

Since our dialogue is a question-answering interaction, we used a query system to build the core of our mechanism. The interlocutor's questions follow the structure previously described and the agent, by fetching the corresponding results within its **Knowledge Base**, returns the events and entities that match the question's conditions from the story shared with that interlocutor.

3. CONCLUSION

In this article we presented a virtual agent's model aimed at endowing a virtual character with the capability to autonomously generate alternatives in its personal story in order to deceive and mislead its interlocutors about past events. The proposed model reduces the effort of a story's author since he does not need to write alternative stories.

An implementation of the model has been realized within an interrogation game and a preliminary experiment revealed promising results as it would appear that the behaviour of the agent showed significant differences depending on the interrogation methodologies followed by the players. For future work, we plan to continue further this investigation by evaluating the system in experimental conditions.

Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

REFERENCES

[1] J. Campos and A. Paiva. May: My memories are yours. In *International Conference on Intelligent Virtual Agents*, pages 406–412. Springer, 2010.

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

[3] B. M. DePaulo, M. E. Ansfield, S. E. Kirkendol, and J. M. Boden. Serious lies. *Basic and applied social psychology*, 26(2-3):147–167, 2004.

[4] B. M. DePaulo and D. A. Kashy. Everyday lies in close and casual relationships. *Journal of personality and social psychology*, 74(1):63, 1998.

[5] J. Dias, W. C. Ho, T. Vogt, N. Beeckman, A. Paiva, and E. André. I know what i did last summer: Autobiographic memory in synthetic characters. In *International Conference on Affective Computing and Intelligent Interaction*, pages 606–617. Springer, 2007.

[6] V. A. Gombos. The cognition of deception: the role of executive processes in producing lies. *Genetic, Social, and General Psychology Monographs*, 132(3):197–214, 2006.

[7] J. Gratch, Z. Nazari, and E. Johnson. The misrepresentation game: How to win at negotiation while seeming like a nice guy. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 728–737. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[8] J. P. Hespanha, Y. Ateskan, and H. Kizilocak. Deception in non-cooperative games with partial information. In *Proceedings of the 2nd DARPA-JFACC Symposium on Advances in Enterprise Control*. Citeseer, 2000.

[9] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik, and D. Crystal. *A comprehensive grammar of the English language*, volume 397. Cambridge Univ Press, 1985.

[10] M. Riedl, C. J. Saretto, and R. M. Young. Managing interaction between users and agents in a multi-agent storytelling environment. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 741–748. ACM, 2003.

[11] M. Riedl, D. Thue, and V. Bulitko. Game ai as storytelling. In *Artificial Intelligence for Computer Games*, pages 125–150. Springer, 2011.

[12] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[13] E. Tulving. Episodic and semantic memory 1. *Organization of Memory. London: Academic*, 381(4):382–404, 1972.

[14] A. Vrij, K. Edward, K. P. Roberts, and R. Bull. Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior*, 24(4):239–263, 2000.

[15] A. Vrij, R. Fisher, S. Mann, S. Leal, B. Milne, S. Savage, and T. Williamson. Increasing cognitive load in interviews to detect deceit. *International developments in investigative interviewing*, pages 176–189, 2009.

[16] J. J. Walczyk, L. L. Harris, T. K. Duck, and D. Mulay. A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. *New Ideas in Psychology*, 34:22–36, 2014.

[17] J. J. Walczyk, K. S. Roper, E. Seemann, and A. M. Humphrey. Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. *Applied Cognitive Psychology*, 17(7):755–774, 2003.

[18] E. G. Weber. *Varieties of questions in English conversation*, volume 3. John Benjamins Publishing, 1993.

[19] M. Zuckerman, B. M. DePaulo, and R. Rosenthal. Verbal and nonverbal communication of deception. *Advances in experimental social psychology*, 14(1):59, 1981.