Proceedings, The Thirteenth AAAI Conference on
Artificial Intelligence and Interactive Digital Entertainment (AIIDE-17)

Intentional Agents for Epistemic Games

Markus Eger
Principles of Expressive Machines Lab
NC State University
Raleigh, NC, USA

Introduction

When humans observe other agents, one key aspect of that agent's behavior that they expect is intentionality, i.e. that the agent is working towards some goal and committed to achieving it (Dennett 1971). It is therefore desirable to develop agents that exhibit such behavior when they are supposed to interact with humans, especially if communication between the agent and the human is involved. However, as Cohen and Levesque (1990) have noted, intention can not be viewed in a vacuum, because it is tightly linked with an agent's beliefs about the world, and that effect is magnified when communication is involved. For my thesis, I am planning an agent framework that exhibits intentional behavior by modeling the agent's beliefs about the world and other agents' beliefs, in a theory of mind. My work not only focuses on having agents act with intentions, but also how they can communicate these intentions to other agents, and even deceive other agents by communicating intentions that they don't actually hold. Therefore, I will be focusing on what I call *epistemic games*, which are (turn-based) games in which the acquisition and exchange of knowledge is an intrinsic part of game play.

Background Information: Epistemic Games

While in almost all games players benefit from having a mental model of other players' beliefs[1], there is a certain class of games where it is almost inevitable, because they include actions that have the sole purpose of changing a player's beliefs as part of the game rules. For example, while having some model of an opponent's cards in poker can be beneficial, the rules of poker never talk about actions that would change a players belief explicitly. On the other hand, in game with communicative actions as part of the game mechanics, the only effect such actions have is to change the belief of the recipient of the communication. Two examples for such games are:

- In **Hanabi** (Bauza 2010), players cooperate to build fireworks represented by cards in five colors with ranks from 1 to 5. However, unlike in most other card games, players can not see the cards in their own hand, but only the cards in the other players' hands. A key component of game play is the ability of players to *give hints* to other players, but these hints have to follow restrictions outlined in the game rules.
- **One Night Ultimate Werewolf** (Alspach and Okui 2014), on the other hand, is a social deduction game, where players are randomly assigned to one of two teams, Werewolves and Villagers, and have to deduce which roles the other players belong to by talking to each other. To aid with this task, some players get special roles that can e.g. look at another player's role card, or secretly swap two role cards.

Approach and Current Status

As a general strategy for playing epistemic games, I propose to use an approach in which the AI agent has intentions, i.e. they form goals and are committed to achieving those goals as long as is feasible. The goals are obtained from a game-specific strategy, but once an agent has formed an intention, it will autonomously pursue it. Intentions are formed in accordance with the agent's beliefs, i.e. they will only work towards goals that they believe to be achievable, but they can also involve getting other agents to do or believe something. To do so, the agent uses its beliefs about the other agents' beliefs as the basis for a model of communication. Using its own goal-finding mechanism, it also predicts which goals other agents are likely to have or adopt when their beliefs change. The actual communicative act is then modeled by actions in epistemic games that have epistemic effects. For example, performing a "hint" action in Hanabi is equivalent to updating the recipients beliefs with the information about the cards the hint is about. In non-cooperative settings, like social deduction games, communicative actions may have non-deterministic epistemic effects, because the speaker does not know whether the recipient believes them or not. The model for beliefs and actions the agent uses must therefore be capable of representing such nuances. As an initial proof of concept, I have already successfully used this approach on Hanabi, and have started working on the more generalized version.

[1]For brevity I will use knowledge and belief interchangeably and not discuss the distinction further

Hanabi Agent

My first step in developing this approach to an intentional agent was to modify agents presented by Osawa (2015). The main reason for this was the observation that while the agents he presented play reasonable well with themselves, they perform actions that are not very intuitive to humans and therefore struggle when they are to cooperate with a human player. My approach added an explicit goal-direct component to the agent, as well as the capability to communicate these goals by taking Grice's maxims of communication into account. As a result, the agent I designed performed significantly better when it played with human players. This agent is described in more detail in an upcoming paper (Eger, Martens, and Alfaro Córdoba 2017)[2], together with a detailed description of the experiment we performed and the results we obtained. Furthermore, the implementation of the agent[3] and the data of all participants who consented to publication[4] has been made available on github, and will be presented as a tech demo at FDG 2017.

Macro-system for Baltag's Logic for Suspicious Players

While the Hanabi agent described above serves as a good first demonstration of an intentional agent in an epistemic game, extending it to other games is non-trivial because it uses a knowledge and action representation that is specific to Hanabi. Baltag's variant of Dynamic Epistemic Logic can be used to describe a wide range of actions that are used in epistemic games (Baltag 2002), including factual changes of the world and their appearance to agents, as well as changes of agent beliefs and observations. However, writing any non-trivial action in this logic is cumbersome. Therefore, I proposed a macro system for the logic, complete with an implementation that makes expressing commonly used actions easier. For example, an action telling an agent about all cards of a particular color that they have can be seen in Listing 1. An initial approach to this macro system was discussed in more detail in our presentation at OBT 2017 (Eger and Martens 2017a), while a paper about the finished version will be presented at AIIDE 2017 (Eger and Martens 2017b).

```
hintcolor(player: Players, col: Colors)
    tell (player):
      Each i in HandIndex:
        color(at(player, i)) == col
```

Listing 1: Hanabi "hint color" action

Future Work

My current focus is on applying the lessons learned from the Hanabi agent to build a generalized framework for intentional agents for epistemic games. This involves defining an encoding of such games, using the macro system described above, and implementing the actual agent. Another part of future work is determining a better evaluation of the agents. Because Hanabi is a cooperative game, the evaluation could use the score the players obtained to determine how well the agent plays with humans, but the same metric may not be applicable in social deduction games, where player enjoyment is likely more important than the agent's win rate.

The first step towards a generalized framework, and the one I am currently working on, is to provide an encoding of the game rules of One Night Ultimate Werewolf as actions in Dynamic Epistemic Logic, utilizing the macro system described above. Since the actual game rules do not limit what players can state, this encoding will only be an abstraction of the communicative actions performed in a typical game, providing players with the option to state truth or falsehoods about their faction affiliation or game actions that they performed. This abstraction can then subsequently be used by the agent to determine its actions. Similar to AI planning (Russell and Norvig 2009), this affords us with actions that have preconditions and effects that can be used to fulfill a goal. The agent then uses these actions to find action sequences that achieve its goals, if possible, and adopts them as intentions. In subsequent turns the currently held intention and any new information will be weighed to determine whether the agent should adopt a new goal, according to the strategy.

As mentioned earlier, evaluating this agent will be another challenge that needs to be addressed. While it will be interesting to see how this agent performs when playing against other agents of the same kind or different ones, the actual goal of my work is to have the agent play well with humans. Of course, such an implementation will not be able to capture all aspects of the game, such as the interpretation of non-verbal player reactions to new information, but one fundamental aspect of One Night Ultimate Werewolf is the deception of other players, and a good measure for the quality of an agent could be how believable the agent performed, i.e. how often the human player actually believes what the agent tells them, or at least considers it possible. To have a baseline to compare to, I want to utilize previous work on social deduction games (Dias et al. 2013) and adapt it to One Night Ultimate Werewolf. Additionally, agents that do not form and commit to intentions would be a reasonable comparison. However, the exact design of the experiment is still open.

Since my thesis is mainly about modeling epistemic actions and their use in intentional agents and conveying intentions, it is not strictly limited to games. While I focus on games, because they provide a convenient test platform with relatively clear performance measures, I am also interested in how my approach could apply to other domains, such as Intelligent Tutoring Systems (Nwana 1990). These systems are used to teach students about a concept which often involves giving hints to the student about what they should do next. One approach to hint generation is to use previous students' attempts at solving the problem at hand (Barnes and Stamper 2008), but my approach could be useful to generate hints for new problems, or situations for which not enough data is present, or simply to augment existing approaches.

[2] Accepted for publication at Computational Intelligence in Games. Preprint available here: http://yawgmoth.github.io/research/hanabiai

[3] https://github.com/yawgmoth/pyhanabi

[4] https://github.com/yawgmoth/HanabiData

Other domains where my approach could be applicable are user support, narrative generation or science communication.

Finally, I also want to address how my approach integrates with a larger-scale agent architecture. Since I am mostly concerned with the implementation of short-term goals and currently use a pre-defined long-term strategy, an important issue to address is how these longer-term strategies can be defined, or which approaches can be used in conjunction with mine.

References

Alspach, T., and Okui, A. 2014. One night ultimate werewolf. https://beziergames.com/collections/all-uw-titles/products/one-night-ultimate-werewolf.

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

Barnes, T., and Stamper, J. 2008. Toward automatic hint generation for logic proof tutoring using historical student data. In *International Conference on Intelligent Tutoring Systems*, 373–382. Springer.

Bauza, A. 2010. Hanabi.

Cohen, P. R., and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial intelligence* 42(2-3):213–261.

Dennett, D. C. 1971. Intentional systems. *The Journal of Philosophy* 68(4):87–106.

Dias, J.; Aylett, R.; Paiva, A.; and Reis, H. 2013. The great deceivers: Virtual agents and believable lies. In *CogSci*.

Eger, M., and Martens, C. 2017a. Practical specification of belief manipulation in games. In *Proceedings of the 13th AAAI International Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Eger, M., and Martens, C. 2017b. Practical specification of belief manipulation in games. In *Proceedings of the 13th AAAI International Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Eger, M.; Martens, C.; and Alfaro Córdoba, M. 2017. An intentional AI for hanabi. In *Computational Intelligence and Games (CIG), 2017 IEEE Conference on*. IEEE.

Nwana, H. S. 1990. Intelligent tutoring systems: an overview. *Artificial Intelligence Review* 4(4):251–277.

Osawa, H. 2015. Solving Hanabi: Estimating hands by opponent's actions in cooperative game with incomplete information. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Russell, S., and Norvig, P. 2009. Planning. In *AI a modern approach*. Prentice Hall. chapter 11, 375–416.