Management of obfuscation-based decision making in a coalition[*]

Nicolas Cointe[0000−0002−7332−1476], Amineh Ghorbani[0000−0002−9985−8239], and Caspar Chorus[0000−0002−6380−4853]

Delft University of Technology, Faculty of Technology, Policy and Management,
Delft, The Netherland
{nicolas.cointe,A.Ghorbani,C.G.Chorus}@tudelft.nl

Abstract. In artificial agents' societies, various plans and goals recognition algorithms provide hypotheses about an agent's motivations based on their observed behavior. As these techniques are useful to detect maleficent agents, they might also be used to threaten privacy and jeopardize their strategy. This paper proposes to introduce in the decision process of the agents an evaluation of the information given to an observer, and raise several questions about the impact of obfuscation on the collective and social dimension of a society of artificial agent.

Keywords: Obfuscation · Transparency · Multiagent systems.

Introduction

Plan and goals recognition is often presented as a necessity to identify agents considered as threats or accused of misconduct, and eventually react in an appropriate way. This approach is obviously well accepted in several application domains such as computer network security [4] or software user assistance [6]. But in many systems, especially where privacy and safety of the human users is involved, being able to identify the goals of the others might be considered as intrusive and illegitimate.

This paper briefly exposes in Section 1 a way to introduce a measurement of the information given to an observer through a behavior, to let the agents choose the best option when they may execute several possible plans (or part of plans) to achieve a goal. Then, Section 2 is dedicated to the discussion about the impact of a decision process based on obfuscation for a society of agents in terms of trust, coalition stability, and betrayal avoidance.

1 Evaluation of the information given to an observer

An observed ordered set of actions executed by an agent is called an *observed behavior* and noted $b_{ag_i} = \{a_n, \cdots, a_{n+x}\}$ where ag_i is the observed agent. To

[*] This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 724431)

identify the goals and plans used by an agent with an observation and interpretation of its behavior, the Plan Recognition (PR)[8] methodology proposes a way to infer the most likely goal according with a shared plan library [7,10]. To do so, the observer compares the behavior with a plan library defined by Kabanaza et al. as a tuple $L = \langle A, G, I, R\rangle$ with A a set of actions, G a set of goals, $I \subseteq G$ a set of intendable goals and R a set of goal-decompostion rules defined as $g \rightarrow \tau$ meaning that the goal g can be accomplished by achieving each element of the partially ordered string τ over the alphabet $(A \cup G)$ represented as a pair $[\beta, C]$. $\beta \in (A \cup G)*$ represents a string of goal and action symbols and C is a set of ordering constraints. A constraint (i, j) means that the i^{th} symbol of β must be achieved or executed before the j^{th} symbol of β.

This definition is only a conceptual view of the observer and is used to evaluate the similarity between an observed behavior and a set of plans, even if the observed agent is not really using this plan library (for instance, if the observed agent is a human being) or if the actions are not entirely observable (e.g. if they are internal actions or if agents only have a local perception). We denote $\mathcal{A}$ the set of the agents and b_a a sequence of actions executed by an agent $a \in \mathcal{A}$ (called behavior of a). An *explanation* of an observed behavior b_a is defined [4] as a minimal forest of plan trees to allow the assignment of each observation to a specific action in the plans.

Considering these concepts, a *Plan recognition problem* PR [10] is defined by a tuple $\langle L, b\rangle$. A PR solving algorithm takes the plan library L and an observed behavior b, and returns a set of explanations consistent with the observations. Many efficient algorithms have been proposed in the literature to solve PR problems, such as ELEXIR[3], PHATT[4], YAPPR[5], DOPLAR[7] and SLIM[9].

As an observer agent may use these techniques to analyze a behavior, an observed agent may take in consideration the information provided through the observations in order to obfuscate [2] her goals or at the opposite, make it as transparent as possible. Obfuscation is here a manner to have more control and select who you want (or should) share information with, and must be distinguished from deception as it does not broadcast fake information or leads to delude observers, but cares about (and minimizes) the given information.

We introduce the set $\mathcal{E}_{b_{ag_i}}$ of the explanations of the behavior b_{ag_i} according with the shared plan library L and produced by an explanation function EF such as $EF : I \times b_{ag_i} \rightarrow \mathcal{E}_{b_{ag_i}}$ where $e_j \in \mathcal{E}_{b_{ag_i}}$ is set of pairs of an observed action $a_{ag_i,t}$ associated with an intendable goal $g \in I$. An intuitive, but inefficient approach to evaluate the set of the possible explanations require to explore the space of all the possible combinations of pairs. The size of this space is $\mathcal{O}(|I|^{|b_{ag_i}|})$, that makes any naive implementation suffers from a combinatorial explosion. Many efficient algorithm are building incrementally the explanation set and updating it during the observation [1], and tends to eliminate the most unlikely hypothesis at each step [7]. We define then the *Most Probable Goal* (or MPG) as the intendable goal the most associated with actions in the set $\mathcal{E}_{b_{ag_i}}$ of all the possible explanations for the behavior b_{ag_i}. As a naive measure of the *likelihood* of this plan, we may consider the proportion of action-goal pairs containing the MPG in $\mathcal{E}_{b_{ag_i}}$. An *onlooker* agent is defined as an agent that infers and updates, for each action executed by an agent ag_i and perceived through her perception function, such belief as `isTheMPGof(`ig `,` ag_i `)`. This predicate mean that $ig \in I$ is considered at this moment as the Most Probable Goal of the agent ag_i according with the reasoning process defined in this section.

In order to integrate the impact of a decision for a potential onlooker, we provide here a new process to add in the context of a plan some additional information and introduce the predicates $maximizeEntropy(c, t, O, b, h_{c,t,b})$ and $minimizeEntropy(c, t, O, b, h_{c,t,b})$ with c a choice, such as $c \in A \cup G$, t the moment when the choice is supposed to be executed, O the set of options, or possible choices, such as $O \subseteq A \cup G$ and $c \in O$, b the past behavior of the agent until t, $h_{c,t,b}$ the shannon entropy of the likelihood of all the intendable goals if the choice c is executed from t and observed in addition with the behavior b.

The predicate $maximizeEntropy(c, t, O, b, h_{c,t,b})$ is true iff $\nexists c' \in O$ such as $h_{c',t,b} > h_{c,t,b}$. Respectively, the predicate $minimizeEntropy(c, t, O, b, h_{c,t,b})$ is true iff $\nexists c' \in O$ such as $h_{c',t,b} < h_{c,t,b}$. Obfuscation-based decision makers maximizes entropy for all the intendable goals of I. Transparency-based decision maker minimizes the entropy to maximize the likelihood of their intended goal. Here, obfuscation-based decision making differs from deception-based agents as it does not tend to behave such that $P(b_a|ig') > P(b_a|ig)$ with $ig \in \mathcal{I}$ the selected intendable goal the agent actually wants to achieve and $ig' \in \mathcal{I}$ another intendable goal of the library L.

A proof of concept has been implemented to illustrate these concepts both in a realistic application or in randomly-generated plan libraries, to illustrate and compare the results of these strategies[1].

2 Obfuscation and cooperation

Let us imagine an application where agents are allowed to cooperate, able to hamper the success of the others' plans, and have a motivation to do so (for instance in a cooperative zero-sum game). If the target is engaged in a cooperation, a plan failure may have an impact on the coalition's efficiency. In this case, to hide the goals and selected plans to the opponents seems a major concern. As a counterpart, being able to explain your goals and plans to the coalition might be a requirement to both prevent betrayals from malicious agents and coordinate actions (for instance : resources sharing, avoid deadlocks and so on). To verify the likelihood of a declared explanation and build a trust-based relationship, agents may observe the other members of the coalition and verify that the given explanation is in the set of the possible explanations of the observed behavior. The design of such framework will be the next step in our research.

We also have to explore the potential benefit for an observer if an agent is identified as a member of a coalition. If the observer assumes that a set of agents are probably trying to achieve a common goal, and eventually using coordination

[1] Download it, and find more information on `www.nicolascointe.eu/projects/`

or delegation mechanisms to collectively execute a plan, the aggregation and analysis of their behaviors may be different from the analysis of their behaviors independently. Then, obfuscating their goals is no longer only a personal concern, but should also be considered as a duty towards the coalition.

Finally, if keeping her goals secret is an absolute priority, it might be a valuable option to quit a coalition, and eventually switch to another one, if the behavior of the current one is not considered obfuscated enough. In an open and decentralized system, this social behavior will lead to an obfuscation-based self-organized society.

3 Conclusion

We have presented a mechanism to embed a classic plan recognition technique into a BDI agent, not only to evaluate the most probable goal of the others, but also to anticipate their evaluation of the decision maker's behavior and ensure that a decision minimizes or maximizes the given information to the others. In the last section, we exposed a set of propositions and questions to explore the impact of obfuscation-based trust mechanisms or information sharing in agents' communities, and the potential benefit of a set of agents' behavior analysis from an observer perspective to identify the goals of a coalition.

References

1. Dorit Avrahami-Zilberbrand and Gal A Kaminka. Fast and complete symbolic plan recognition. In *IJCAI*, pages 653–658, 2005.
2. Caspar G Chorus. How to keep your av on the right track? an obfuscation-based model of decision-making by autonomous agents. *hEART 2018*, 2018.
3. Christopher W Geib. Delaying commitment in plan recognition using combinatory categorial grammars. In *IJCAI*, pages 1702–1707, 2009.
4. Christopher W Geib and Robert P Goldman. Partial observability and probabilistic plan/goal recognition. In *Proceedings of the International workshop on modeling other agents from observations (MOO-05)*, volume 8, pages 1–6, 2005.
5. Christopher W Geib, John Maraist, and Robert P Goldman. A new probabilistic plan recognition algorithm based on string rewriting. In *ICAPS*, pages 91–98, 2008.
6. Eric Horvitz, Jack Breese, David Heckerman, David Hovel, and Koos Rommelse. The lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 256–265. Morgan Kaufmann Publishers Inc., 1998.
7. Froduald Kabanza, Julien Filion, Abder Rezak Benaskeur, and Hengameh Irandoust. Controlling the hypothesis space in probabilistic plan recognition. In *IJCAI*, pages 2306–2312, 2013.
8. Henry A Kautz and James F Allen. Generalized plan recognition. In *AAAI*, volume 86, page 5, 1986.
9. Reuth Mirsky et al. Slim: Semi-lazy inference mechanism for plan recognition. *arXiv preprint arXiv:1703.00838*, 2017.
10. Reuth Mirsky, Roni Stern, Kobi Gal, and Meir Kalech. Sequential plan recognition: An iterative approach to disambiguating between hypotheses. *Artificial Intelligence*, 260:51–73, 2018.