Towards empathic autonomous agents

Timotheus Kampik[0000-0002-6458-2252], Juan Carlos Nieves[0000-0003-4072-8795], and Helena Lindgren[0000-0002-8430-4241]

Umeå University, 901 87, Umeå, Sweden
{tkampik,jcnieves,helena}@cs.umu.se

Abstract. Identifying and resolving conflicts of interests is a key challenge when designing autonomous agents. For example, such conflicts often occur when complex information systems interact persuasively with humans and are in the future likely to arise in non-human agent-to-agent interaction. We work towards a theoretical framework for an *empathic* autonomous agent that proactively identifies potential conflicts of interests in interactions with other agents (and humans) by learning their utility functions and comparing them with its own preferences using a system of shared values to find a solution all agents consider *acceptable*. To provide a high-level overview of our work, we propose a reasoning-loop architecture to address the problem in focus. To realize specific components of the architecture, we suggest applying existing concepts in argumentation and utility theory. Reinforcement learning methods can be used by the agent to learn from and interact with its environment.

Keywords: Multi-agent systems · Utility theory · Conflicts of interests

1 Background and problem description

In modern information technologies, conflicts of interests between users and information systems that operate with a high degree of autonomy (*autonomous agents*) are of increasing prevalence. For example, complex web applications persuade end-users, possibly against the interests of the persuaded individuals[1]. Given the prevalence of autonomous systems will increase, conflicts between autonomous agents and humans (or between autonomous agents among themselves) can be expected to occur more frequently in the future, e.g. in interactions with or among autonomous vehicles in scenarios that cannot be completely solved by applying static traffic rules. Consequently, one can argue for the need to develop *empathic* intelligent agents that consider the beliefs, desires, and intentions of others, as well as ethics rules and social norms when interacting with their environment to avoid severe conflicts of interests. As a simple example, take two vehicles (*A* and *B*) that are about to enter a bottleneck. Assume they cannot enter the bottleneck at the same time. A and B can either wait or drive. Considering only its own utility function, A might determine that *driving* is the best action to execute, given that B will likely stop and wait to avoid a crash. However, A should ideally

[1] E.g., research provides evidence that contextual advertisement influences how users process online news [11] and a social network application has effectively been employed for political persuasion [3].

assess both its own and B's utility function and act accordingly. If B's utility for driving is considered higher than A's, A can then come to the conclusion that *waiting* is the best action. As A does not only consider its own goals, but also the ones of B, one can regard A as *empathic*, following Coplan's definition of empathy, as "a process through which an observer simulates another's situated psychological states, while maintaining clear self–other differentiation" [8]. While existing literature covers conflict resolution in multi-agent systems from a broad range of perspectives (see for a partial overview: [1]), devising a theoretical framework for autonomous agents that proactively observe their environment and use a utilitarian approach to identify and resolve conflicts of interests can be considered a novel idea. However, existing multi-agent systems research can be leveraged to implement core components of such a framework, as will be discussed later.

The goal of our research is to define a reasoning-loop architecture and theoretical framework for an autonomous agent that proactively detects and resolves possible conflicts of interests with other, generally cooperative agents when engaging with its environment. The rest of this paper is organized as follows: in Section 2 we present first steps towards a set of formal definitions of the problem in focus and outline a basic reasoning-loop architecture for the to-be-developed agent. In Section 3 we discuss its alignment with the belief-desire-intention architecture, as well as a possible implementation using the Jason framework. Finally, in Section 4, we outline future work.

2 Progress

Given a set of agents and their *possible* actions in any interaction scenario, we define the utility of an agent as a function of the actions of all agents at a given point in time. Our utility function returns a numerical value. To simplify our model, our set of actions is *deterministic* and considers only one state transition at a time. We can define a conflict of interests between several agents as any situation in which there is no set of possible actions that maximizes the utility functions of all agents. Considering the incomparability property of the von Neumann-Morgenstern utility theorem [10], such a conflict can be solved only if a system of values exists that is shared between the agents and used to determine the individual utility values. The value system can introduce generally applicable rules, e.g. to hard-code a prioritization of individual freedom into an agent. Given the value system, we create a pragmatic definition of a conflict of interest as any situation, in which there is no set of actions that is regarded as *acceptable* by all agents when considering the shared set of values, given each agent executes the actions that maximize their individual utility function. Considering the notion of *acceptability*, the utility function can be extended to form an *acceptability function*. The acceptability function is derived from the corresponding utility functions and the shared system of values and takes a set of actions as its inputs. Without this notion, our definition of a conflict of interest would cover many scenarios that most human societies would regard as not conflict-worthy, e.g. when one agent would need to accept large utility losses to optimize their own actions towards marginally improving another agent's utility.

We provide a running example for the "vehicle/bottleneck" scenario introduced above, assuming B is twice as fast as A (e.g. without waiting, A needs 20 time units to pass the bottleneck while B needs 10). Each agent has a utility function $u : \mathcal{A}_A \times \mathcal{A}_B \rightarrow \{-\infty, \mathbb{R}, \infty\}$. $\mathcal{A}_A$ and $\mathcal{A}_B$ are all possible actions A and B can execute. First, we specify a shared value system that contains three rules. Rule 1 determines the generally applicable policy (for n agents) and defines that the agent can act egoistically ($max(u_{i|e})$), as long as its behavior does not negatively impact the maximum of the utility function of another agent in comparison to a scenario in which the agent does *not interfere* (*ni*) with this agent ($max(u_{i|ni})$); otherwise, the agent should act to maximize the combined utility of all agents:

$$p(\{u_{A|ni}, u_{A|e}, ..., u_{n|ni}, u_{n|e}\}, u_{self}) = \begin{cases} max(u_{self}), \text{ if } \forall\, u_{i|ni}, u_{i|e} \in \{\{u_{A|ni}, u_{A|e}\}, ..., \{u_{n|ni}, u_{n|e}\}\} : max(u_{i|ni}) \leq max(u_{i|e}); \\ max \bigcup_{a \in \mathcal{A}_a, ..., n \in \mathcal{A}_n} \{u_a(a, ..., n) + u_n(a, ..., n)\}, \text{otherwise.} \end{cases}$$

Note that in our scenario, the only relevant $u_{i|e}$ is $\bigcup_{b \in \mathcal{A}_b} \{u_b(A_{drive}, b)\}$. u_{self} is the set of the possible outcomes of the agent's own utility function ($\bigcup_{a \in \mathcal{A}_a, b \in \mathcal{A}_b} \{u_A(a, b)\}$). Rule 2 is a helper function for rule 1 and defines *non-interference* as *waiting*:

$$u_{i|ni} = \bigcup_{b \in \mathcal{A}_b} \{u_i(A_{wait}, b)\}$$

Rule 3 provides a shared utility definition, stating the utility of any agent is a function of the time until the agent has passed the bottleneck and the (non-)occurrence of a crash:

$$\begin{aligned} u(b, t) =& b + \frac{1}{t}, \text{where:} \\ b =& \begin{cases} -\infty, & \text{if crash;} \\ 0, & \text{otherwise.} \end{cases} \\ t =& \text{ time until bottleneck passed.} \end{aligned}$$

We can model the utility functions as follows (note that $A \oplus B := (A \vee B) \wedge \neg(A \wedge B)$):

$$u_A(A_{drive} \oplus A_{wait}, B_{drive} \oplus B_{wait}) = \begin{cases} \frac{1}{20}, & \text{if} \quad A_{drive} \wedge B_{wait}; \\ \frac{1}{30}, & \text{if} \quad A_{wait} \wedge B_{drive}; \\ 0, & \text{if} \quad A_{wait} \wedge B_{wait}; \\ -\infty, & \text{if} \quad A_{drive} \wedge B_{drive}. \end{cases}$$

$$u_B(A_{drive} \oplus A_{wait}, B_{drive} \oplus B_{wait}) = \begin{cases} \frac{1}{10}, & \text{if} \quad A_{wait} \wedge B_{drive}; \\ \frac{1}{30}, & \text{if} \quad A_{drive} \wedge B_{wait}; \\ 0, & \text{if} \quad A_{wait} \wedge B_{wait}; \\ -\infty, & \text{if} \quad A_{drive} \wedge B_{drive}. \end{cases}$$

From the value system and the utility functions, we can derive the following acceptability functions:

$$acc_A(A_{drive} \oplus A_{wait}, B_{drive} \oplus B_{wait}) = \begin{cases} \text{true, if } (A_{drive} \wedge B_{wait}) \vee (A_{wait} \wedge B_{drive}); \\ \text{false, otherwise.} \end{cases}$$

$$acc_B(A_{drive} \oplus A_{wait}, B_{drive} \oplus B_{wait}) = \begin{cases} \text{true, if } A_{wait} \wedge B_{drive}; \\ \text{false, otherwise.} \end{cases}$$

The only solution that is acceptable for both agents is $\{A_{wait}, B_{drive}\}$. Now, A can execute A_{wait}, given B_{drive} can be the expected reaction by B.

We create a reasoning-loop architecture of the empathic agent and again assume a two-agent scenario to simplify the description. The architecture consists of the following components:

- **Empathic agent (EA)**: The empathic agent is the system's top-level component. It has three functional components (*observer*, *negotiator*, and *interactor*) and five data objects (*utility function* and *acceptability function* of both agents, as well as a formalized model of the *shared system of values*).
- **Target agent (TA)**: In the simplest scenario, the empathic agent interacts with exactly one target agent, which is modeled as a black box. Pre-existing knowledge about a target agent can be part of the models the empathic agent has of the target agent's utility and acceptability functions.
- **Shared system of values**: The shared system of values allows comparing the utility functions of the different agents and creating their acceptability functions.
- **Utility function**: The empathic agent maintains its own utility function, as well as models of the utility functions of the agents it is interacting with.
- **Acceptability function**: Based on the utility functions and the shared system of values, the agent derives the acceptability functions (as described above) to then derive the best possible set of actions.
- **Observer**: The observer component scans the environment, registers other agents, constructs their utility functions, and also keeps the agent's own functions updated. To construct and update the utility and acceptability functions, the observer could make use of reinforcement learning methods as for example described by Chajewsk et al. [7].
- **Negotiator**: The negotiator identifies and resolves conflicts of interests using the *acceptability function* models and instructs the interactor to engage with other agents if necessary, in particular, to propose a solution for a conflict of interest, or to resolve the conflict immediately (depending on the level of confidence that the solution is indeed acceptable). The negotiator could make use of argument-based negotiation (see e.g.: [2]).
- **Interactor**: The interactor component interacts with the agent's environment and in particular with the target agent to work towards the conflict resolution. The means of communication is domain-specific and not covered by the generic architecture.

Fig. 1 presents a simple graphical model of the empathic agent.

3 Alignment with BDI architecture and possible implementation with Jason

Our architecture reflects the common belief-desire-intention (BDI) model as based on [5] to some extent:

- If a priori available to both agents in the forms of rules or norms, *beliefs*, and *belief sets* are part of the shared value system. Otherwise, they qualify the agents' utility

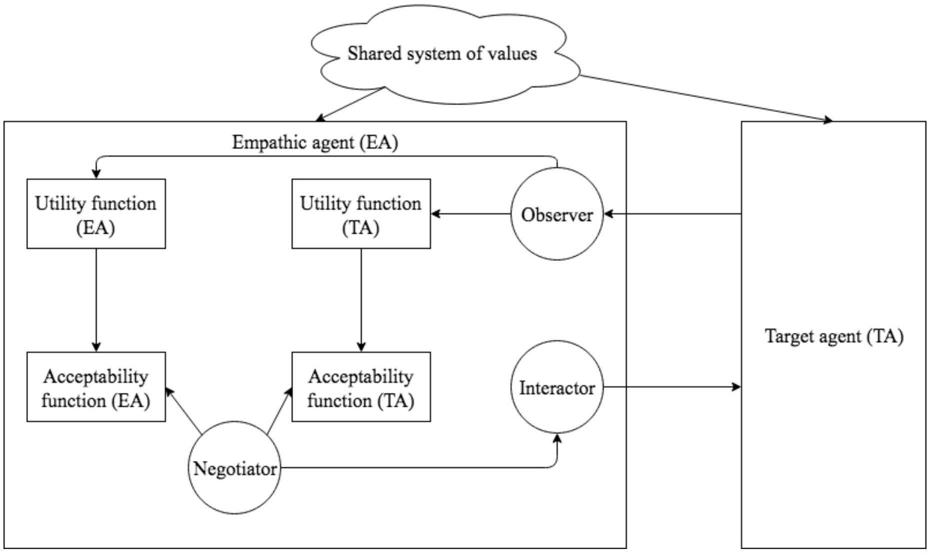

Fig. 1: Empathic intelligent: architecture

and acceptability functions directly. In contrast, *desires* define the objective(s) towards which an agent's utility function is optimized and are–while depending on beliefs–not directly mutable through persuasive argumentation between the agents.
- Intentions are the sets of actions the agents choose to execute.
- As it strives for simplicity, our architecture does for now not distinguish between desires and goals, and intentions and plans, respectively.

We expect to improve the alignment of our framework with the BDI architecture to facilitate the integration with existing BDI-based theories and implementation using BDI frameworks. The Jason platform for multi-agent system development [4] can serve as the basis for implementing the empathic agent. While simplified running examples of our architecture can be implemented with Jason, we consider extending the platform with abstractions to better support complex scenarios.

4 Future work

We are working on a detailed formal definition of conflicts of interests, especially regarding the value system and acceptability function, as well as on a more precise architecture for detecting and resolving them. So far, we have chosen a logic-based approach to the problem in focus to allow for a minimalistic problem description with low complexity. Alternatively, the problem could be approached from a *reinforcement learning* perspective (see for an overview of multi-agent reinforcement learning: [6]). Using (partially observable) Markov decision processes, one can introduce a well-established temporal and probabilistic perspective[2]. We plan to combine reinforcement

[2] However, the same can be achieved with temporal and probabilistic logic.

learning methods for observational learning of the utility functions of other agents with argumentation-based negotiation approaches that consider uncertainty and subjectivity (e.g. [9]) for creating solvers for finding compromises between utility/acceptability functions. However, the design intention of the architectural framework is to form a high-level abstraction of an empathic agent that is to some extent agnostic of the concepts the different components implement. We are confident that the framework can be applied in combination with existing technologies, as long as some assumptions regarding the interaction context and protocol can be made.

Acknowledgements This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. Alshabi, W., Ramaswamy, S., Itmi, M., Abdulrab, H.: Coordination, cooperation and conflict resolution in multi-agent systems. In: Sobh, T. (ed.) Innovations and Advanced Techniques in Computer and Information Sciences and Engineering. pp. 495–500. Springer Netherlands, Dordrecht (2007)
2. Amgoud, L., Dimopoulos, Y., Moraitis, P.: A unified and general framework for argumentation-based negotiation. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems. pp. 158:1–158:8. AAMAS '07, ACM, New York, NY, USA (2007)
3. Berinsky, A.J.: Rumors and health care reform: experiments in political misinformation. British Journal of Political Science **47**(2), 241–262 (2017)
4. Bordini, R.H., Hübner, J.F.: BDI agent programming in AgentSpeak using Jason. In: International Workshop on Computational Logic in Multi-Agent Systems. pp. 143–164. Springer (2005)
5. Bratman, M.: Intention, Plans, and Practical Reason. Center for the Study of Language and Information (1987)
6. Busoniu, L., Babuska, R., De Schutter, B.: A comprehensive survey of multiagent reinforcement learning. IEEE Trans. Systems, Man, and Cybernetics, Part C **38**(2), 156–172 (2008)
7. Chajewska, U., Koller, D., Ormoneit, D.: Learning an agent's utility function by observing behavior. In: ICML. pp. 35–42 (2001)
8. COPLAN, A.: Will the real empathy please stand up? a case for a narrow conceptualization. The Southern Journal of Philosophy **49**(s1), 40–65 (2011)
9. Marey, O., Bentahar, J., Khosrowshahi-Asl, E., Sultan, K., Dssouli, R.: Decision making under subjective uncertainty in argumentation-based agent negotiation. Journal of Ambient Intelligence and Humanized Computing **6**(3), 307–323 (Jun 2015)
10. Von Neumann, J., Morgenstern, O.: Theory of games and economic behavior. Bull. Amer. Math. Soc **51**(7), 498–504 (1945)
11. Wojdynski, B.W., Bang, H.: Distraction effects of contextual advertising on online news processing: an eye-tracking study. Behaviour & Information Technology **35**(8), 654–664 (2016)