Empathic Agents: A Hybrid Normative/Consequentialistic Approach

Doctoral Consortium

Timotheus Kampik
Dept. of Computing Science, Umeå University
Umeå, Sweden
tkampik@cs.umu.se

ABSTRACT

Complex information systems operate with increasing degrees of autonomy. Consequently, such systems should not only optimize for simple metrics (like clicks and views) that reflect the system provider's preferences but also consider norms or rules, as well as the preferences of other agents that are affected by the systems' actions. As a means to achieve such behavior, we propose the design and development of *empathic agents* that use a mixed rule/utility-based approach when deciding on how to act, considering both their own and others' utility functions. The agents make use of formal argumentation to reach an agreement on how to act in case of inconsistent beliefs. A promising domain for applying our empathic agents is *recommender systems*.

ACM Reference Format:
Timotheus Kampik. 2019. Empathic Agents: A Hybrid Normative/ Consequentialistic Approach. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Modern information systems are increasingly autonomous in their decision-making processes. Consequently, autonomous decision-making processes have a growing impact on individuals and society as a whole. However, these processes are often designed to primarily serve the interests of the system provider [6]. For example, recommender systems are typically optimized for simple metrics–like clicks and views–that are proxies for the interests of the system provider and largely ignore the impact on the end-users. In the recommender system community, it is a well-established challenge to provide recommendations that form "an optimal solution to meet the needs of both the users and the system designers" [13]. From a more generic perspective, one can hence derive the challenge of designing rational agents that make trade-offs between their own utility functions (or preferences) and the utility functions of other agents in their environment, while ideally also considering generally applicable norms; *i.e.*, these agents neither act in pure self-interest, nor in a fully collaborative mode. Further scenarios that face this type of challenge are, for example, healthcare scenarios, in which information systems make health-related decisions that can potentially be in conflict with the will of the affected individual(s) and traffic situations, in which static rules are insufficient for resolving conflicts between (potentially autonomous) traffic participants.

2 PROPOSED APPROACH

To address the problem introduced above, we propose developing agents that use a hybrid rule-based (*normative*) and utility-based (*consequentialistic*) approach when determining their actions and take into account not only their own utility but also the utility of other agents in their environment. We call these agents *empathic*, based on Coplan's definition of empathy as "a process through which an observer simulates another's situated psychological states, while maintaining clear self–other differentiation" [3]. The definition resembles the established psychological concept of *theory of mind* [9], which has recently gained the attention of artificial intelligence researchers (*e.g.* [12]).

Our research is intended to range from developing theoretical concepts, over the design of software frameworks for implementing the proposed concepts, to the empirical evaluation of the concepts in human-computer interaction studies. As a relevant domain for applied research contributions, we have identified *recommender systems*. The empathic agent approach could help solve the aforementioned challenge of mitigating conflicts of interests between users and system providers. A multi-agent approach, in which users and system providers are represented by agents and that uses the empathic agent concept as an agreement technology interface between these agents, could be employed to mitigate this problem[1].

We ask the following overall research questions:

- How can a basic definition of an *empathic agent* agreement algorithm be formalized? (**Q1**)
- How can the empathic agent algorithm be enhanced to deal with inconsistent beliefs and subjective information? (**Q2**)
- What software engineering abstractions can enable the application of the empathic agent concept? (**Q3**)[2]
- How can the empathic agent concept be applied to the recommender systems domain? (**Q4**)
- Does the empathic agent concept provide usability advantages that can be confirmed in an empirical human-computer interaction study? (**Q5**)

[1] As discussed by Sunstein, system providers could be incentivized to implement solutions that compromise between their own interests and the interests of the end-users–or society at large–by consumer choice (assuming consumers will prefer services that consider their long-term interests) or be forced to do so by government regulation[14].
[2] It is planned to provide such abstractions as open-source libraries/frameworks.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

3 PROGRESS

So far, we have worked towards research questions **Q1**, **Q2**, and **Q3**. We have established preliminary theoretical foundations for *empathic agents* that use a mixed rule-/utility-based approach to resolve conflicts of interests with other agents [7].

The empathic agent core framework can be summarized as follows:

- In an environment n *empathic* agents $\{A_0, ..., A_n\}$ interact. In the core framework, we assume the environment is fully observable. At a specific point in time each agent can execute a set of actions $Acts_i := \{Act_i^0, ..., Act_i^m\}$. We assume the agents act simultaneously.
- Each agent A_i has a utility function u_i that maps all possible action combination to a numeric value: $u_i := Acts_0 \times ... \times Acts_n \rightarrow \{null, -\infty, \mathbb{R}, \infty\}$. In addition to the utility functions, the agents refer to a set of acceptability rules. An acceptability rule Acc_i determines whether a set of actions is acceptable for an agent A_i: $Acc_i := Acts_0 \times ... \times Acts_n \rightarrow \{null, true, false\}$. Note that the *null* value as the output of utility and acceptability functions must only be used for action combinations that are practically impossible, *i.e.* when actions are mutually exclusive.
- A conflict of interests between the agents exists if there is no equilibrium strategy that is acceptable and maximizes the utility of all agents.
- If no conflict exists, the agents execute the actions that maximize their own utility function; otherwise, the agents find the acceptable equilibrium strategy that provides maximal own utility. If no such strategy exists, the agents execute the actions that maximize shared utility[3]. In some cases, empathic agents will consider multiple action sets as feasible. Given all agents implement the same deterministic empathic agent algorithm, the agents can then execute the first set in the list of possible action sets the algorithm returns.

We enhanced the empathic agents concept to allow for the use of an argumentation approach (*i.e.*, Dung's abstract argumentation with maximal ideal semantics [4]) to synchronize inconsistent beliefs about acceptability rules [8]. As a proof-of-concept, we implemented the argumentation-enabled empathic agents using the agent development framework Jason in a recommender system scenario. To enable the agents to argue, we created a Jason extension that allows the agents to resolve argumentation frameworks[4].

4 RELATED WORK

While our empathic agent is grounded in existing research on game theory and agreement technologies[5], its placement at the intersection of mixed-motive and fully cooperative games limits the body of existing similar research. Still, conceptual comparisons to some existing works are possible. For example, the persuasive agents approach proposed by Black and Atkinson [2] is similar to our concept in that their agent makes use of formal argumentation techniques and attempts to model the other agent's preferences. However, their agent maintains a model of the preferences of others purely for *persuasive* purposes and does not attempt to compromise between its own preferences and the preferences of others. Consequently, a rigid theoretical comparison to alternative approaches in regards to properties like computational complexity, as well as an evaluation based on established benchmarks do not seem to be possible. The lack of comprehensive comparability implies that a thorough practice-oriented assessment of the proposed concept is necessary and hence motivates our plan to evaluate empathic agents in a human-computer interaction study.

5 FUTURE WORK

To extend the theoretical foundations of the developed empathic agent approach and to further advance its applicability, we plan to conduct the following research:

- **Provide a Markovian perspective on our developed agent that adds a temporal perspective to the problem in focus. (Q1)** It can be expected that a formalization of empathic agents as (multi-agent) Markov decision processes will highlight challenges in regards to the complexity of the problem (large state space that grows exponentially with the number of agents involved in the scenario).
- **Improve the agent's ability to handle incomplete and subjective information and enable it to learn. (Q2)** In particular, the capabilities of empathic agents to reach consensus in case of inconsistent/subjective beliefs as introduced in our previous work [8] is planned to be enhanced by employing value-based [1] and possibilistic [11] argumentation approaches. Moreover, an inverse reinforcement learning perspective (*c.f.* [10]) on the Markovian approach can be developed.
- **Develop frameworks and libraries that enable the application in real-world scenarios. (Q3, Q4)** Considering complexity constraints, it makes sense to first focus on providing engineering abstractions for the recommender system domain, which allows for relatively simple models of environment and users/agents.
- **Evaluate empathic agents empirically in a recommender system scenario. (Q5)** A first evaluation can utilize multi-agent simulations. Later, a human-computer interaction study can assess the usefulness of empathic agents for improving an application's user experience.

ACKNOWLEDGMENTS

The author thanks Helena Lindgren and Juan Carlos Nieves for their supervision and support. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

[3] Different aggregation functions to determine the maximal shared utility are available: while the *sum* of all agents' utility seems most intuitive, using the *product* can facilitate a "fair" distribution of utility among the agents.

[4] The code of the running examples (including the extension and documentation) is available at https://github.com/TimKam/empathic-jason (archived at https://zenodo.org/record/2585805).

[5] For an overview of multi-agent negotiation, see for example Fatima and Rahwan [5].

REFERENCES

[1] Trevor JM Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13, 3 (2003), 429–448.

[2] Elizabeth Black and Katie Atkinson. 2011. Choosing persuasive arguments for action. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 905–912.

[3] Amy Coplan. 2011. Will the real empathy please stand up? A case for a narrow conceptualization. *The Southern Journal of Philosophy* 49, s1 (2011), 40–65.

[4] Phan Minh Dung, Paolo Mancarella, and Francesca Toni. 2007. Computing ideal sceptical argumentation. *Artificial Intelligence* 171, 10-15 (2007), 642–674.

[5] Shaheen Fatima and Iyad Rahwan. 2013. Negotiation and bargaining. In *Multiagent Systems, Second Edition*, Gerhard Weiss (Ed.). MIT Press, Cambridge, Chapter 4, 143–176.

[6] Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. 2018. Coercion and deception in persuasive technologies. In *20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018*. CEUR-WS, 38–49.

[7] Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. 2019. Empathic Autonomous Agents. In *EMAS 2018: Engineering Multi-Agent Systems*, Viviana Mascardi, Alessandro Ricci, and Danny Weyns (Eds.). Springer International Publishing, Cham.

[8] Timotheus Kampik, Juan Carlos Nieves, and Helena Lindgren. 2019. Implementing Argumentation-Enabled Empathic Agents. In *EUMAS 2018: Multi-Agent Systems*, Marija Slavkovik (Ed.). Springer International Publishing, Cham, 140–155.

[9] Alan M Leslie. 1987. Pretense and representation: The origins of "theory of mind.". *Psychological review* 94, 4 (1987), 412.

[10] Andrew Y Ng, Stuart J Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *ICML*. 663–670.

[11] Juan Carlos Nieves and Roberto Confalonieri. 2011. A Possibilistic Argumentation Decision Making Framework with Default Reasoning. *Fundamenta Informaticae* 113, 1 (2011), 41–61.

[12] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine Theory of Mind. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 4218–4227.

[13] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. Recommender systems: introduction and challenges. In *Recommender systems handbook*. Springer, 1–34.

[14] Cass R Sunstein. 2018. *# Republic: Divided democracy in the age of social media*. Princeton University Press.