Empathetic Reinforcement Learning Agents

Doctoral Consortium

Manisha Senadeera
Applied Artificial Intelligence Institute, Deakin University
Geelong, Australia
manisha.senadeera@deakin.edu.au

ABSTRACT

With the increased interaction between artificial agents and humans, the need to have agents who can respond to their human counterparts appropriately will be crucial for the deployment of trustworthy systems. A key behaviour to permit this, one which humans and other living beings exhibit naturally, is empathy. In my research I explore the potential for agents to behave in ways that may be considered empathetic. Empathy is a two stage process involving the identification of the feelings or goals of the other, and having that same feeling be evoked in oneself. I began my work towards this objective by initially designing an agent who exhibits sympathy - the ability to identify the goals of another. Empathy is slightly more complex as it involves a process of projecting the state of the other back onto oneself and observing one's own response. In my research I hope to draw inspiration from this and evoke empathy through a process of mapping the other's goals back to oneself. By drawing upon empathetic responses, the hope is that this will lead to a faster and deeper understanding of the other.

KEYWORDS

Reinforcement Learning; Empathy; Sympathy; Inverse Reinforcement Learning

ACM Reference Format:
Manisha Senadeera. 2022. Empathetic Reinforcement Learning Agents: Doctoral Consortium. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

Machine learning and artificial intelligence has played a prominent role in the acceleration of the current information and data driven world of today. As the prevalence of these technologies grows in their impact and influence, more awareness is coming to the forefront about their limitations and risks. Work is currently underway in areas of safety, governance, and policy in AI [1]. There are significant concerns around the potential harm that such technologies could bring. As our use of these technologies extend from one directional interactions (like a user with their smart device or car) to more bi-directional interactions (a virtual secretary, or robotic pet), the need to create agents who are able to reason, act morally, and evaluate the social and ethical implications of their actions becomes increasingly important. The advancements made in algorithms in the fields of computer vision, learning, reasoning and planning have been incorporated into robotic technologies, much of which is now commercially available to the average consumer. This interaction with humans has seen itself extend to critical applications, such as AI doctors [5] and agents that interact with the elderly [15]. As such these technologies need to be designed to exhibit more nuanced care in their actions, laying particular emphasis on the safety and comfort of the other agents (humans).

In human-to-human interactions, there are many factors that lead to harmonious relationships. Humans are able to easily sympathise and empathise with others. This ability to care extends beyond being directed at just other humans, and includes animals, plants and to some extent, non-living objects [6]. It is this ability to preempt how others will behave and feel, learnt from an understanding of how oneself will behave and feel under a similar situation, that allows us to exist cohesively. As such, in order for AI technologies to better integrate with humans and be accepted, important human-like characteristics will need to be exhibited. There are many key components for this integration, but the focus of this research will be limited to a fundamental part of human interaction - empathy. As defined by Hoffman [7], empathy is "*any process where the attended perception of the object's state generates a state in the subject that is more applicable to the object's state or situation than to the subject's own prior state or situation*". This can be broken down into two processes - the first identifies that the other is experiencing a particular emotion, and the second generates a state within the observer that more closely resembles the other's state.

Using reinforcement learning (RL) [14], my thesis aims to develop agents who can more readily identify with others through empathy. Although this endeavour involves modelling other agents, my focus differs from that of prior RL work in Theory of Mind (ToM) [8] and goal recognition [12]. Specifically, goal recognition observes sequences of actions to predict the goal of another agent, while ToM attempts to interpret the others' mental states, intentions and beliefs [3]. An empathy-based approach will differ, as additional information based on oneself is incorporated into this modelling process, thereby allowing for a richer understanding of the other and the ability to better identify the best actions to take in the interests of the other. Previous works have attempted to model empathetic agents but many have fallen short of the definition above [4, 10]. I base my approach of inferring behaviours of other agents using inverse reinforcement learning (IRL), [9] which we will boost with empathy information. Once this understanding has been developed, it will be used to evoke considerate behaviours. In particular, we consider learning a sympathy function to determine the degree to which an agent should act selfishly or selflessly based on the situation at hand. This function is designed to operate dynamically in an online fashion whilst ensuring versatility and generalisation. As my intended application is in human-robot interaction scenarios, we hope to eventually adapt my empathy framework to operate

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

in more complex and realistic situations, and be adaptable to the stochastic nature of human behaviours.

2 CONTRIBUTIONS

The objective of my research is to develop a framework that trains agents to behave empathetically as described by Hoffman [7]. This can be described as agents who can concurrently:

(1) understand the goals of other agents they interact with,
(2) understand when and how their actions impact the other agent,
(3) map similar goals from their own value system to that of the other agent (similar actions/rewards).
(4) adjust their actions to behave considerately of the other.

3 SYMPATHETIC AGENTS

My first piece of work [13] focused on the development of a sympathetic agent. Sympathy differs from empathy as it is primarily focused just on the identification of motivations or feelings in the other.

In this work I presented a sympathetic framework which takes into consideration the goals of both agents. Our environmental setting is motivated by situations where a human may interact with a robot. The human (referred to as the other agent, or the independent agent) is assumed to already possess their own beliefs and behave according to a fixed policy. The robot (learning agent) however, must learn to concurrently complete its task, whilst also behaving considerately of the other agent. To do so, my framework infers the reward function (goals) of the other agent through IRL. Using a convex weighted sum of the inferred rewards and the learning agent's own rewards returned from the environment, a sympathetic reward is constructed, based on which the learning agent is trained. The weighting of the two rewards is determined by a dynamically adjusted sympathy function, which considers the long term outcomes for both agents to ensure the learning of considerate behaviours, while simultaneously giving due importance to its own rewards. Previous works [2, 4, 10] have also proposed the use of such a weighting, however in these works this weighting was considered a fixed hyperparameter that the user was required to specify beforehand. This was a key limitation, as the degree of weighting could not adapt to the situation the agent is faced with within the game.

My proposal was applied to both adversarial and assistive games, and was shown to successfully induce considerate behaviours whilst also allowing the learning agent to complete its task.

4 EMPATHETIC AGENTS

Following my work on sympathetic behaviour, I now focus on extending the previous work to mimic and evoke an empathetic response. Empathy differs by the added process whereby the emotional state of the other is also evoked within oneself. Though the differences between an empathetic agent and a benevolent agent may not be overly apparent, I believe the empathy based process of modelling the other will induce slightly different behaviours. As there will be a strong emphasis on mapping states back to oneself, a deeper understanding of another's experience will manifest itself in more considerate actions as the learning agent will behave in ways that it would have wanted another to behave had it been in a similar situation.

In a virtual game or real-world robot, what this may look like is the ability to develop a model of the other faster (as you can use yourself as a source of information). Additionally, empathy allows a more nuanced understanding of the other, more so than sympathy. For example if the other agent is hurt, the learning agent can reflect upon themselves to garner how they would want to be treated by the other if they themselves were hurt. This provides a clearer understanding on the best action the learning agent should take in response to the other agent's situation.

Though still in the preliminary stages, one avenue of thought is to identify when the rewards of the two agents are similar, and use this to determine whether the situation would evoke empathy. Compared to the previous sympathy project where IRL was used to infer the rewards of the independent agent, with empathy, I could instead map the rewards from the independent agent to the learning agent (and vice-versa). When such a link is made, agent rewards can be divided into intrinsic rewards (rewards that are commonly shared between the two agents e.g. negative rewards from being harmed), and environmental rewards (rewards that are unique to the objective of each agent). The identification of intrinsic rewards could lead to a better understanding of the potential true reward of the other agent, thus enabling the learning agent to leverage its own value function to infer the value function of the other agent.

Previous works have examined empathy along a similar vein. Work by Bussmann et al. [4] imposes the complete value function of the learning agent over the other agent to imagine how it would feel if it was in the other agent's position. The limitation of this work, which I hope to alleviate is the assumption that the other agent also values completely what the learning agent values. Work by Raileanu et al. [11] removes this limiting assumption and uses the learning agent's model parameters and the observed actions of the other agent to infer what its goal may be. Although this work is most closely aligned with my objective, it is limited by constraints associated with the goals of the game and the training procedure. This is a constraint I aim to address in my own work.

5 FUTURE WORK

In the future, I aim to adapt my framework to handle complex and stochastic independent agent policies, a setting that more accurately represents real human behaviours. Additionally, it is currently assumed that reward features are shared between the agent and the human. However, humans are likely to have a vaster space of features and preferences. As such, exploration into adapting our framework to more generalised reward functions will be required. Both of these challenges could involve substantial improvements to the IRL component. The exact nature on how to implement this is still unclear, but is a very interesting area for future work.

We believe that equipping artificial agents with empathetic behaviours is an important area for further research, particularly with the proliferation of human-agent systems. Future work in this area could tackle a number of issues such as building trust between humans and artificial agents, and helping to robustly align agent behaviours with human values.

REFERENCES

[1] 2020. *Report on the safety and liability implications of artificial intelligence, the internet of things and robotics.* Technical Report. European Commission. https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en

[2] Parand Alizadeh Alamdari, Toryn Q Klassen, Rodrigo Toro Icarte, and Sheila A McIlraith. 2021. Be Considerate: Objectives, Side Effects, and Deciding How to Act. *arXiv preprint arXiv:2106.02617* (2021).

[3] James Blair, Carol Sellars, Ian Strickland, Fiona Clark, Akintude Williams, Margaret Smith, and Lawrence Jones. 1996. Theory of mind in the psychopath. *Journal of Forensic Psychiatry* 7, 1 (1996), 15–25.

[4] Bart Bussmann, Jacqueline Heinerman, and Joel Lehman. 2019. Towards empathic deep q-learning. In *Artificial Intelligence Safety 2019 (CEUR Workshop Proceedings)*. Artificial Intelligence Safety 2019, CEUR-WS.org, 1–7.

[5] Sumit Das, S Biswas, Aditi Paul, and Aritra Dey. 2018. AI doctor: an intelligent approach for medical diagnosis. In *Industry interactive innovations in science, engineering and technology*. Springer, 173–183.

[6] Alain Goudey and Gael Bonnin. 2016. Must smart objects look human? Study of the impact of anthropomorphism on the acceptance of companion robots. *Recherche et Applications en Marketing (English Edition)* 31, 2 (2016), 2–20.

[7] Martin L Hoffman. 1996. Empathy and moral development. *The Annual Report of Educational Psychology in Japan* 35 (1996), 157–162.

[8] Julian Jara-Ettinger. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences* 29 (2019), 105–110.

[9] Andrew Y. Ng and Stuart Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *in Proc. 17th International Conf. on Machine Learning*. in Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, 663–670.

[10] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R. Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. 2019. Teaching AI agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development* 63 (2019), 2–1.

[11] Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. 2018. Modeling Others using Oneself in Multi-Agent Reinforcement Learning. (Feb. 2018).

[12] Leonardo Rosa Amado, Reuth Mirsky, and Felipe Meneguzzi. 2022. Goal Recognition as Reinforcement Learning. *arXiv e-prints* (2022), arXiv–2202.

[13] Manisha Senadeera, Thommen George Karimpanal, Sunil Gupta, and Santu Rana. 2022. Sympathy based Reinforcement Learning Agents. In *AAMAS, 2022*. (to appear).

[14] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[15] Kazuyoshi Wada and Takanori Shibata. 2007. Living with seal robots–its sociopsychological and physiological influences on the elderly at a care house. *IEEE transactions on robotics* 23, 5 (2007), 972–980.