AI Alignment Dialogues

An Interactive Approach to AI Alignment in Support Agents

Chen, Pei Yu

DOI
10.1145/3514094.3539531

Publication date
2022

Document Version
Final published version

Published in
AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society

AI Alignment Dialogues: An Interactive Approach to AI Alignment in Support Agents

Pei-Yu Chen
Delft University of Technology
Delft, The Netherlands
p.y.chen@tudelft.nl

ABSTRACT

This project proposes a different way of looking at AI alignment, namely by introducing AI Alignment Dialogues. We argue that alignment dialogues have a number of advantages in comparison to data-driven approaches, especially for behaviour support agents, which aim to support users in achieving their desired future behaviours rather than their current behaviours. The advantages of alignment dialogues include allowing the users to directly convey higher-level concepts to the agent and making the agent more transparent and trusted.

KEYWORDS

AI alignment, Human values, Responsible AI, Behaviour support agent, Conversational agent

ACM Reference Format:
Pei-Yu Chen. 2022. AI Alignment Dialogues: An Interactive Approach to AI Alignment in Support Agents. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3514094.3539531

As behaviour support systems are increasingly deployed in the real world over a long period of time [6], it is inevitable that there will be situations where the provided support does not fully match the user's ideal choices or needs. We refer to this as 'misalignment'. In this PhD project, we focus on addressing these misalignment situations with the ultimate goal to achieve alignment.

The central question of the project is related to the notion of AI alignment. Russell [5] proposes to achieve AI alignment by building AI systems that learn and promote human values so that they do not pose a threat to us. One such approach usually use inverse reinforcement learning (see, e.g., [3, 4]). Inverse reinforcement learning, among other data-driven approaches, largely depends on behavioural data and have their limitations. First, the existing data only reflects past or current user behaviour. It does not account for desired future behaviour that the user is not yet doing. Moreover, data-driven approaches can lack transparency because of the complex relation between input data and a model's output [2]. This makes it difficult for users to understand how the system works and also to adapt the system to their preferences.

AIES '22, August 1–3, 2022, Oxford, United Kingdom.

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3539531

To overcome these problems, we propose to approach AI alignment in the context of support agents with a direct conversation between user and agent. This is referred to as an 'AI Alignment Dialogue': a dialogue where the user and the support agent talk about the situation and how to solve it together to achieve and maintain AI alignment. Having an AI alignment dialogue allows the users to directly convey higher-level concepts to the agent, such as their values, goals, or wishes. These high-level concepts can be difficult to learn from the data [1]. In addition, when the user converses with the agent, the agent could give insight into how it works or why it gives the advice using the same high-level concepts. These high-level concepts may be easier to understand for the user than the lower-level features on which behavioural data-driven approaches work. These aspects together could make the agent more transparent and trusted. Moreover, the agent may not always have full knowledge of the user's situation and wishes. Via a dialogue we can have the user in the loop to create the required understanding. By doing this, we ensure the AI is responsible: humans and machines work together in a way that is beneficial for the user and allows them to shape their lives in accordance to what they find important.

To the best of our knowledge, this is the first proposal for approaching the AI alignment problem via a dialogue. We take a different but important point of view about alignment and venture into how alignment dialogues might go beyond the learning of assumed 'preferences' to reach goals, values, aspirations, and the like. We hope this project provides another perspective within the community and offers new possibilities for a major theme in AI ethics: alignment.

ACKNOWLEDGMENTS

This work was supported by the Hybrid Intelligence Gravitation Programme (project number 024.004.022).

REFERENCES

[1] Stuart Armstrong and Sören Mindermann. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. *Advances in Neural Information Processing Systems* 31 (2018).
[2] Virginia Dignum. 2017. Responsible autonomy. *arXiv preprint arXiv:1706.02513* (2017).
[3] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016), 3909–3917.
[4] Michael Herman, Tobias Gindele, Jörg Wagner, Felix Schmitt, and Wolfram Burgard. 2016. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Artificial Intelligence and Statistics.* PMLR, 102–110.
[5] Stuart Russell. 2019. *Human compatible: Artificial intelligence and the problem of control.* Penguin.
[6] Sanne van Waveren, Elizabeth J Carter, and Iolanda Leite. 2019. Take one for the team: The effects of error severity in collaborative tasks with social robots. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents.* 151–158.