UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title Al Agents Learn to Trust

Permalink https://escholarship.org/uc/item/1tg688k5

Journal Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors Nobandegani, Ardavan S. Rish, Irina Shultz, Thomas

Publication Date

2023

Peer reviewed

AI Agents Learn to Trust

Ardavan S. Nobandegani Mila - Quebec AI Institute, Montreal, Quebec, Canada

Irina Rish Mila - Quebec AI Institute, Montreal, Quebec, Canada

Thomas Shultz McGill University, Montreal, Quebec, Canada

Abstract

Widely considered a cornerstone of human morality, trust shapes many aspects of human social interactions. In this work, we show that artificial intelligence (AI) agents learn to trust — and rationally adapt the extent to which they trust — by interacting with their counterparts in the trust game (Berg, Dickhaut, & McCabe, 1995), the canonical task for studying trust in behavioral and brain sciences. Leveraging reinforcement learning (RL) to train our AI agents, we systematically investigate learning trust under various parameterizations of this task and show that trust can arise from pure self interest. Our work sheds new light on the computational underpinnings of trust, taking us a step closer to developing AI agents whose values are aligned with human values.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. ©2023 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).