Unraveling Human Goals: Computational Explorations

Du Jiang
School of Intelligence Science and Technology
Peking University
dujiang@pku.edu.cn

Abstract

In the realm of artificial intelligence, the representation of goals is essential for systems that exhibit cognitive abilities. This essay explores three crucial aspects—teleological reasoning, inverse planning and Bayesian goal inference, and preferenced-based reinforcement learning—that collectively form a comprehensive framework for the computational representation of goals.

1 Introduction

Computational explorations in goal representation play an essential role in shaping the capabilities of artificial intelligence systems and are indispensable for achieving a reasonable understanding of human intentionality. The significance of these explorations lies in their potential to bridge the gap between raw data and human-like cognitive processes, empowering computational models to interpret, predict, and adapt to the diverse and dynamic goals inherent in human behavior.

This essay summarizes some computational explorations into the representation of human goals, emphasizing three key aspects—teleological reasoning, inverse planning and Bayesian goal inference, and preferenced-based reinforcement learning. Teleological reasoning grounds the understanding of actions in their goal-oriented context. Inverse planning and Bayesian goal inference unravel implicit intentions by inferring goals from observed actions. Preferenced-based reinforcement learning ensures flexibility and scalability in goal representation. The synthesis of these aspects promises an AI framework capable of discerning explicit and implicit human intentions and navigating the diverse landscape of goal specifications.

2 Computational Approaches for Goal Representation

2.1 Teleological Reasoning

Teleological reasoning focuses on understanding actions in the context of their goals[7, 4]. In this paradigm, actions are not merely recognized as isolated events; rather, they are seen as purposeful endeavors aimed at achieving specific outcomes. In the computational realm, this involves establishing a mapping between actions and their intended results.

By incorporating teleological reasoning, a computational model gains the ability to discern the motivations and objectives that underlie observed behaviors. Different from traditional action recognition, this approach offers a contextually rich understanding of human-like behavior and enables the model to handle the complexities of diverse goals, ranging from abstract long-term visions to detailed immediate objectives.

Teleological reasoning, while foundational, may have limitations in comprehensively understanding the contextual nuances of human actions[1]. It might struggle to capture the intricate interplay of situational factors that influence goal-oriented behavior, leading to potential misinterpretations. In scenarios where the goals are ambiguous or multifaceted, teleological reasoning may face challenges in accurately discerning the intended outcomes. The inherent complexity of human intentions, especially in social or emotional contexts, poses difficulties for purely teleological approaches.

2.2 Inverse Planning and Bayesian Goal Inference

While teleological reasoning provides a foundation for understanding explicit goals, human behavior often involves implicit intentions that are not explicitly stated. Inverse planning and Bayesian goal inference tackle the challenge of inferring goals from observed actions[2, 3]. In the related computational frameworks, the model treats action understanding as the reverse process of planning and attempts to uncover the goals that could have led to a particular set of actions.

The Bayesian perspective introduces probabilistic models, allowing the model to make inferences about goals while accounting for the inherent uncertainty in human decision-making. This kind of computational framework is designed to be capable of robustly attributing goals, even when they exist in the realm of implicit and unspoken motivations.

However, inverse planning and Bayesian goal inference approaches often demand substantial amounts of high-quality training data to accurately model and infer goals. Insufficient or biased datasets can lead to suboptimal performance and limit the generalizability of the models to diverse contexts. The probabilistic nature of Bayesian models can result in increased computational overhead. The complexity of inference algorithms may hinder real-time applications, making them less suitable for scenarios requiring rapid decision-making.

2.3 Preferenced-Based Reinforcement Learning

Preferenced-based reinforcement learning offers a solution by representing goals as preferences over states or trajectories[5, 6]. The learning process is guided by comparisons or rankings between different trajectories or actions. Instead of receiving traditional scalar rewards, the agent receives feedback in the form of user preferences, indicating whether one trajectory or action is preferred over another.

This approach provides a flexible and scalable framework, allowing for the specification of a wide range of goals. By ranking goals in terms of preference, the model accommodates both abstract, long-term visions and detailed, immediate goal conditions. This adaptability is crucial in real-world applications where goals can vary greatly among users or in dynamic environments. Preferenced-based reinforcement learning thus equips computational models with the capability to understand and handle the diverse goals of human users.

One of the key challenges in preferenced-based reinforcement learning is the proper specification of reward functions. Designing reward functions that accurately reflect human preferences without introducing biases or unintended consequences can be a daunting task, requiring careful tuning and validation. The model's reliance on preferences assumes that user preferences are well-defined and consistent, which might not always be the case. Users' preferences can evolve, be context-dependent, or even subject to irrationality, leading to potential misalignment between the model's assumptions and the user's true preferences.

3 Discussion

By designing computational frameworks with the ability to understand actions in the context of their goals, infer implicit intentions, and flexibly accommodate diverse goal specifications, previous works pave the way for intelligent systems that navigate the intricacies of human intentionality.

It's important to note that the mentioned disadvantages are not insurmountable, and ongoing research seeks to address these challenges. Hybrid approaches that combine the strengths of these methods or incorporate additional contextual information are emerging to enhance the robustness and applicability of goal representation in diverse real-world scenarios. Moreover, the choice of which approach to use often depends on the specific application and the nature of the goals being represented.

References

[1] Ian A. Apperly and Stephen A. Butterfill. Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116 4:953–70, 2009. URL https://api.semanticscholar.org/CorpusID:4997651. 1

[2] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2009.07.005. URL https://www.sciencedirect.com/science/article/pii/S0010027709001607. Reinforcement learning and higher cognition. 2

[3] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1, 2017. URL https://api.semanticscholar.org/CorpusID:3338320. 2

[4] Gergely Csibra and György Gergely. 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, 124(1):60–78, 2007. ISSN 0001-6918. doi: https://doi.org/10.1016/j.actpsy.2006.09.007. URL https://www.sciencedirect.com/science/article/pii/S0001691806001235. Becoming an Intentional Agent: Early Development of Action Interpretation and Action Control. 1

[5] W. Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the Fifth International Conference on Knowledge Capture*, K-CAP '09, page 9–16, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605586588. doi: 10.1145/1597735.1597738. URL https://doi.org/10.1145/1597735.1597738. 2

[6] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8. 2

[7] Amanda L Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34, 1998. ISSN 0010-0277. doi: https://doi.org/10.1016/S0010-0277(98)00058-4. URL https://www.sciencedirect.com/science/article/pii/S0010027798000584. 1