An overview of computational methods for goal modeling

Rui Yang
Yuanpei College
Peking University
ypyangrui@pku.edu.cn

Abstract

Goals play an crucial role in driving humans for their daily lives. They can be as abstract as a long-term blueprint or as specific as an immediate directive. For sophisticated AI agents, goals are essential as they not only define the objectives to be achieved but also mimic human decision-making processes, enabling agents to understand and interact in human-centric environments. In this essay, we reflect on the psychological understanding of goal perception and attribution with behavioural experiments and proposed mechanisms. Based on this point, we delve into the contemporary computational techniques for goal modeling. We specifically focus on Bayesian models and inverse reinforcement learning. In the end, we conclude with the pros and cons of existing methods and discuss potential future directions.

1 Introduction

Humans, as a special species with sophisticated social structures, can be “obsessed with goals". Due to the highly specialized division of labor in society, humans often need to achieve various goals in order to obtain the necessary material or social status for survival. Students are expected to complete their coursework with excellent grades in order to land in good jobs or further education opportunities. Assistant professors need to work hard to produce rich research results, apply for funding, and socialize with the academic community in order to get tenure.

In all these activities, goals play a significant role. They serve as fundamental drivers which provide direction and purpose, guiding individuals towards desired outcomes. By setting goals, people can organize their actions and resources more effectively. In addition, the ability to perceive others' goals is indispensable for interaction and collaboration between different individuals, which is crucial for the shaping of human community.

Therefore, the accurate and efficient modeling of goals is key to building AI agents with human-level ability for action planning and multi-agent collaboration. In the realm of computational systems, goal representation involves defining and structuring objectives that the algorithm aims to achieve, translating abstract aspirations into quantifiable and actionable directives. This stands as not just a technical necessity but also a philosophical inquiry about how machines interpret, prioritize, and pursue objectives. It is essential for mirroring the human process of goal setting and achieving in a digital context.

In this essay, we review the concept of intentionality and goal attribution in psychology. We then go on to list several classic computational methods for modeling goals, and compare their strengths and weaknesses. In the end, we discuss the challenges faced by goal modeling and point out potential future directions.

2 Psychological evidences for goal attribution

Goal attribution has long been paid close attention by the psychology community. Before delving into the computational methods, let us first look back on the studies by psychologists about human goal perception.

It is amazing that infants as young as six-month-old already demonstrate the ability to perceive and understand agents' intent by seeing human activities as goal-directed behavior [16]. By the age of eighteen-month-old, they can not only infer but also imitate the goal of an action even observed with continuous failure [4]. More notably, they can analyze the underlying hierarchy structure of concrete action goals, higher order plans, and collaborative intentions [17]. These altogether reflect an innate and refined mechanism for goal attribution.

Generally speaking, there are three mechanism proposed for the teleological interpretation of actions in humans [6]:

- Action-effect associations. Based on the ideomotor principle [12], this view lays stress on the synthesized representation of goals and motor actions in the cognitive system [11]. In essence, the perception of goals is achieved by bidirectional associations between actions and their effects. Numerous developmental psychology experiments provide evidence for this link [14, 16, 13].
- Simulation procedures. In this theory, individuals comprehend the thoughts of others by "putting themselves in their shoes", thereby creating a simulation of the mental states (such as beliefs, desires, and intentions) if they were in each others' situation [8, 9]. This is useful for understanding goal-directed actions, and especially important for "reducing the possible range of solutions through relying on the 'equivalence' assumption that the observed actor has the same motor constraints and preferences as the observer" [6].
- Teleological reasoning. While the former theoretical approaches are well suited for online action monitoring and prediction, they lack the inferential productivity for social learning of new means actions and artefact functions. Taking relevant constraints of the situation into consideration, the teleological interpretation system evaluates goal-directed action with rationality principle [5]. This is achieved by the computational representation of "functional stance" [19].

In summary, these three mechanisms complement each other and serve as an insightful view into the realization of human goal attribution.

3 Goal representation with computational methods

In this section, we compose the review on the computational methods for goal modeling.

From several prominent experiments which reveals the irresistible visual perception of goals and intention of even the simplest moving shapes [10, 3, 7], an important theory of "rationality principle" [5] is derived for the judgement of goal-redirected action. This theory believes that humans are rational creatures who attempt to maximize the utility while minimizing the costs.

Inverse planning governed by Bayesian inference come into shape and formally models the principle. The main idea of this method is to integrate the likelihood of observed behaviours with prior mental states via Bayesian inference. Hence the observer may infer the latent intention through rational planning model inversion.

One ground-breaking work is *Goal Inference as Inverse Planning* by Baker et al. [2]. In this work, the authors propose a framework to invert a probabilistic generative model of goal-dependent plans for goal inference. This framework combines the observations of actions with prior knowledge of goal space and converts to specific models under different circumstances. Three models targeting single underlying goal, complex goals, and changing goals are elaborated and demonstrated with experiments. Following works along this line advance from symbolic input to real videos, encompassing aerial events [15], outdoor path trajectories [18] and so on.

Methods based on Bayesian inference excel in their flexibility in modeling probabilistic dependencies and causal relationships. They are also an efficient computational modeling for the "action-effect

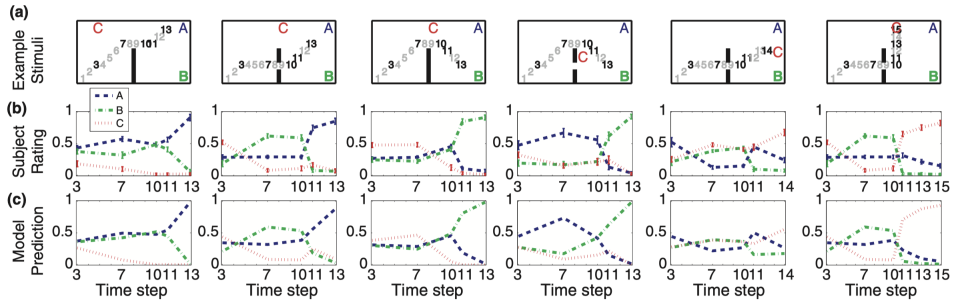

Figure 1: Experiment 1 from *Goal Inference as Inverse Planning* by Baker et al. [2].

associations" and “simulation procedures" mechanisms mentioned in the former section. In a word, inverse planning successfully combines top-down prior knowledge about goals with bottom-up observations of behaviors, and achieve powerful induction given sparse observation sequences.

The analogue of inverse planning in modern artificial intelligence is Inverse Reinforcement Learning (IRL), which is a method for inferring the reward function of an agent based on its policy or observed behavior. This is useful for understanding the decision-making process of other agents without manually specified reward function. The scheme of IRL can be explained through Figure 2 from Arora and Doshi [1].

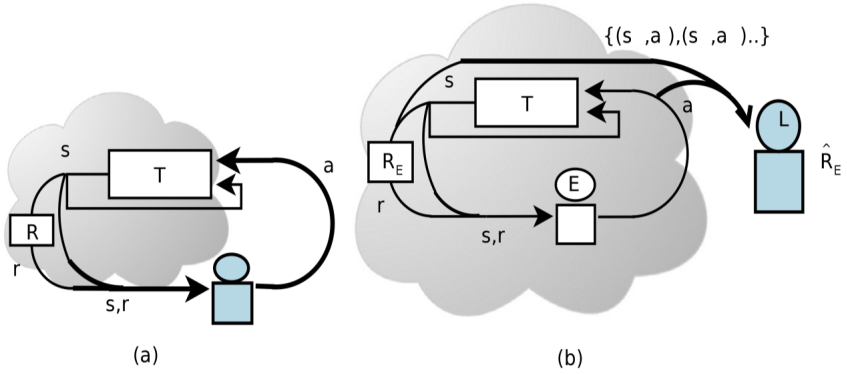

Figure 2: (a) The subject agent (shaded in blue) performing RL. The agent chooses an action at a known state and receives a reward generated by a reward function R. The state changes according to transition function T. (b) In inverse learning or IRL, the input and output for the learner L are reversed. L perceives the states and actions of expert E and learns a reward function R that best explains E’s behavior.

Apart from Bayesian inference, there are other categories of methods which solve the problem of inverse reinforcement learning. For example, marginal optimization, including margin of optimal from other actions or policies, margin of observed from learned feature expectations, and observed & learned policy distributions over action. Other categories involve entropy optimization, classification and regression, which we will not elaborate on due to the limitation of space.

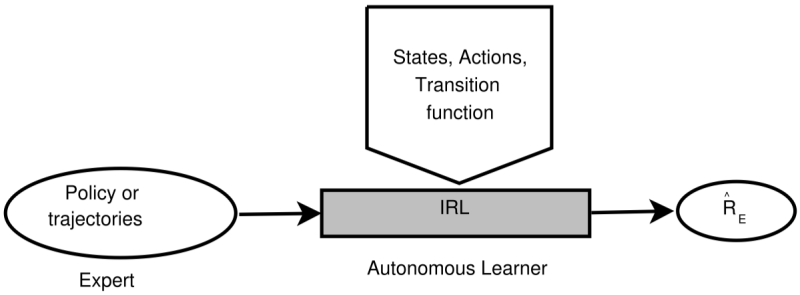

Figure 3: Generalized IRL pipeline with incomplete model of transition probabilities.

The primary challenges faced by IRL include the difficulty of performing accurate inference, the generalizability of the algorithms, sensitivity to prior knowledge, the complexity of solutions (which

tends to grow disproportionately with the size of the problem), and so on. Some challenges are mitigated or conquered through methods mentioned above, but more remain as open problems for future research. Efficient methods with little time complexity and good scalability are called for, as well as benchmark suites for evaluating the performance of such algorithms.

In contrast to goal inference and attribution, the forward modeling of goal is more straightforward and allows for all kinds of methods, depending on different needs. A phenomenal showcase is AlphaGo's triumph in defeating world champion Lee Se-dol, exemplifying the power of abstract goals in AI. The precision of industrial robots, on the other hand, illustrates a different emphasis on the efficacy of detailed goals. More complex problems such as the hierarchical representation of goal structure and complex goal organization with different priority may be taken into consideration as well.

Computational modeling of long-term goals involving value judgement may be more vague at current stage. There have been some attempts with value functions and U-V systems. Surely this would be the key to more human-like general artificial intelligence in the future exploration.

4 Conclusion

In this essay, we delved into the significance of goals in both human psychology and artificial intelligence. We observed that goal understanding begins early in human life, highlighting its fundamental role in shaping our behavior. Psychological studies discussed shed light on the potential underlying mechanism for goal attribution.

In the realm of AI, techniques like Bayesian inference and inverse reinforcement learning have shown promise in emulating human-like goal perception and inference. These methods effectively merge observations with prior knowledge to infer intentions, though they still face challenges in efficiency and scalability.

We also notice the comparison between AI agents which deal with immediate goals, like in robotics or games, and those modeling long-term goals and values. The latter is a more complex area left for future research.

In summary, the integration of psychological insights with AI methodologies is key to creating AI systems with goal perception abilities. As we continue to refine these approaches, hopefully AI will become increasingly adept at understanding human goals and cooperation.

References

[1] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021. 3

[2] Chris L Baker, Joshua B Tenenbaum, and Rebecca R Saxe. Goal inference as inverse planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007. 2, 3

[3] John N Bassili. Temporal and spatial contingencies in the perception of social events. *Journal of personality and social psychology*, 33(6):680, 1976. 2

[4] Szilvia Biro and Bernhard Hommel. Becoming an intentional agent: introduction to the special issue. 2007. 2

[5] Gergely Csibra and György Gergely. The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental science*, 1(2):255–259, 1998. 2

[6] 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. 2

[7] Winand H Dittrich and Stephen EG Lea. Visual perception of intentional motion. *Perception*, 23(3):253–268, 1994. 2

[8] Alvin I Goldman. *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press, 2006. 2

[9] Robert M Gordon. Folk psychology as simulation. *Mind & language*, 1(2):158–171, 1986. 2

[10] Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944. 2

[11] Bernhard Hommel, Jochen Müsseler, Gisa Aschersleben, and Wolfgang Prinz. The theory of event coding (tec): A framework for perception and action planning. *Behavioral and brain sciences*, 24(5):849–878, 2001. 2

[12] William James. The principles of psychology, volume i. *New York: Holt*, 1390, 1890. 2

[13] Ildikó Király, Bianca Jovanovic, Wolfgang Prinz, Gisa Aschersleben, and György Gergely. The early origins of goal attribution in infancy. *Consciousness and cognition*, 12(4):752–769, 2003. 2

[14] Alan M Leslie. Infant perception of a manual pick-up event. *British Journal of Developmental Psychology*, 2(1):19–32, 1984. 2

[15] Tianmin Shu, Yujia Peng, Lifeng Fan, Hongjing Lu, and Song-Chun Zhu. Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations. *Topics in cognitive science*, 10(1):225–241, 2018. 2

[16] Amanda L Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34, 1998. ISSN 0010-0277. 2

[17] Amanda L Woodward, Jessica A Sommerville, Sarah Gerson, Annette ME Henderson, and Jennifer Buresh. The emergence of intention attribution in infancy. *Psychology of learning and motivation*, 51:187–222, 2009. 2

[18] Dan Xie, Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Learning and inferring "dark matter" and predicting human intents and trajectories in videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1639–1652, 2017. 2

[19] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 2