How to Represent Goal in AI Systems

Zhuoying Li
Yuanpei College
Peking University
joy@stu.pku.edu.cn

Abstract

Comprehending and interpreting goals is a fundamental aspect of human cognition, essential for successful social interaction and cooperation. This paper explores the psychological foundations that enable humans to deduce goals and intentions from the actions they observe and broadens the conversation to encompass the incorporation of these notions into artificial intelligence systems. The primary focus is on inverse planning and parsing as two computational strategies used to represent goals within machines. These techniques provide insight into the intricacies of goal-directed analysis and come with unique advantages and drawbacks. In conclusion, the paper discusses the hurdles and prospects of integrating the human-like understanding of goals into artificial entities.

1 Introduction

As humans, we are naturally attuned to perceive and understand the goals and intentions of others; this cognitive ability is fundamental to social interactions and collaboration. The quest to endow machines with similar capabilities has sparked considerable interest within both the fields of psychology and artificial intelligence. The endeavor to decode and represent human goals computationally not only enhances our understanding of human cognition but also advances the development of more sophisticated, empathetic, and intuitive AI systems. This essay discusses the psychological perspectives on human goal detection and contrasts them with current computational methods designed to model and represent goals. Through an exploration of several computational methods, we analyze how these different approaches represent goal from different perspective, discussing their relative merits and shortcomings in the journey towards more intelligent and perceptive machines.

2 A psychological view about human goal detection

When you observe someone taking a detour to reach their destination by car, it's natural to assume that they have done so deliberately, perhaps because the direct route is congested, rather than thinking they are foolish. It is widely recognized that people tend to interpret the actions of others as purposeful and directed towards a goal. There is also substantial evidence suggesting that even young infants view certain behaviors as goal-oriented [6, 2]. So, what underlies this tendency to perceive intentional behavior?

One theory that attempts to explain this is the intuitive agency theory, which includes the so-called "rationality principle". This theory argues that humans regard themselves and others as intentional agents who: (i) commit their limited time and resources to actions that will bring about desired changes in the world; and (ii) achieve their intentions rationally by maximizing their utility while minimizing their costs, given their beliefs about the world [3]. Based on this assumption, approaches such as Bayesian Inference try to model intentions through inverse planning, which will be further discussed in Sec. 3.

In [7], the authors contrasted three distinct mechanisms: (i) action-effect association, (ii) simulation procedures, and (iii) teleological reasoning. However, they believe that these mechanisms do not

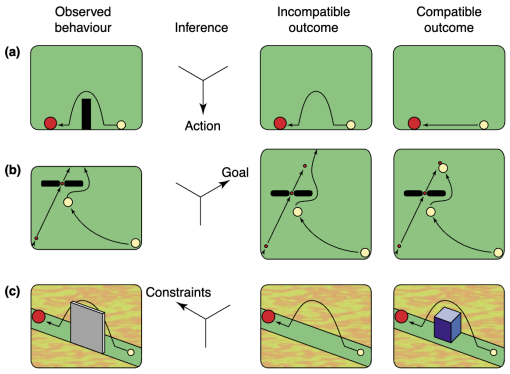

Figure 1: One-year-old infants were habituated to the event depicted in the first column (Observed behaviour). Their interpretation of this event was tested by presenting them with two different outcomes, one of them being incompatible (second column), the other one being compatible (third column) with a possible inference based on a teleological representation of the event. The experiment in [2] shows that infants looked longer at the incompatible outcome than the compatible outcome events, indicating that they view it as a goal-directed action.

compete against one another; instead, they complement one another: The fast effect prediction provided by action-effect associations can serve as a starting hypothesis for teleological reasoning or simulation procedure. In turn, the solutions provided by teleological reasoning in social learning can also be stored as action-effect associations for quick retrieval in the future.

3 Current computational method to represent goal

In this section, I will explore two distinct methods to represent goal, including their respective advantages and disadvantages.

3.1 Inverse planning

Planning is a process in which intent causes action. Conversely, inverse planning aims to deduce intent from executed actions within specific states, which implicitly assumes the "rationality principle". A commonly employed method for inverse planning is Bayesian Inference.

Baker et al. [1] proposed a framework that made use of the Markov Decision Process (MDP) model along with Bayesian rules. More precisely, the model calculates the conditional posterior probability of a designated Goal, based on the observed Actions and the Environment, by applying Bayes rules:

$$P(\text{Goal} \mid \text{Actions}, \text{Environment}) \propto P(\text{Actions} \mid \text{Goal}, \text{Environment}) P(\text{Goal} \mid \text{Environment})$$

In this equation, $P(\text{Goal} \mid \text{Environment})$ is the prior probability that sets up a hypothesis space of goals that are realizable in the environment,. On the other hand, $P(\text{Actions} \mid \text{Goal}, \text{Environment})$ provides bottom-up information from observed actions. The authors simplify Goal in three different ways: (i) M1: goal as a single state of the environment that an agent pursues until it is achieved; (ii) M2: Agents' goals can change over the course of an action sequence; (iii) M3: Agents can have subgoals along their way to final goals.

Holtzen et al. [4] takes a step further. They represent intent by a hierarchical probabilistic And-Or graph structure which illustrates a relationship between actions and plans. By reverse-engineering the decision-making and action-planning processes of a human, they deduce human intent within a Bayesian probabilistic programming framework. Besides, the framework takes RGBD video as inputs instead of symbolic inputs.

Advantages (i) Given prior observations of behavior in similar situations, the model has a good generalization ability; (ii) It's flexible and performs well in small datasets [1].

Disadvantages (i) Human goals are often complex and layered in their hierarchy of importance, which means that creating models using Bayesian inverse planning can require significant human labor; (ii) Acquiring the prior probabilities $P(\text{Goal} \mid \text{Environment})$ is challenging, and it is difficult to ensure they encompass all possible scenarios.

3.2 Parsing

We can view agents' behaviours from another way: actions are like words and activities are like languages. Based on this, taking sequence as inputs, parsing method use grammars parser to parse and predict human activities. A typical work of this kind of method is [5]. As shown in Fig. 2, the generalized Earley parser parses sequence data and finds the optimal segmentation and labels in the language defined by the input grammar. Then the model makes top-down predictions based on the parsing graph.

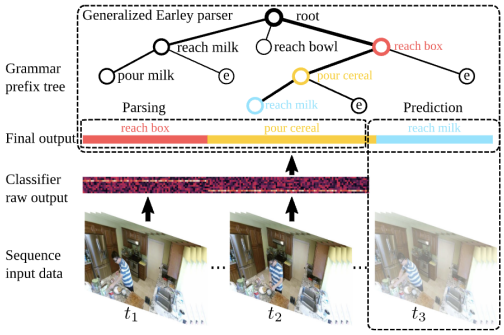

Figure 2: The generalized Earley parser segments and labels the sequence data into a label sentence in the language of a given grammar [5].

Advantages (i) Method based on parsing generates a grammar parse tree for input data sequence, which is highly explainable; (ii) It effectively integrates a high-level grammar with the low-level classifier. The grammatical structure aids in the segmentation and annotation of the sequence data, offering valuable direction for these processes.

Disadvantages (i) In my opinion, given that the method treats the input data as sentences to be parsed, the data must be sequential. This requirement may impose certain limitations; (ii) It is designed for human activity predictions, but there exists a noticeable gap between activity predictions and goal detection. Therefore, it is imperative to devise a strategy for deducing the agent's goals from the parse tree.

4 Conclusion

In my view, significant work remains to be done in representing goals within AI systems. For instance, a critical challenge is how to constrain the parameter space of a goal, given its often indistinct boundaries and potential for considerable complexity.

References

[1] 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

[2] György Gergely and Gergely Csibra. Teleological reasoning in infancy: the naıve theory of rational action. *Trends in Cognitive Sciences*, 7(7):287–292, 2003. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(03)00128-1. URL `https://www.sciencedirect.com/science/article/pii/S1364661303001281`. 1, 2

[3] György Gergely, Zoltán Nádasdy, Gergely Csibra, and Szilvia Bíró. Taking the intentional stance at 12 months of age. *Cognition*, 56(2):165–193, 1995. ISSN 0010-0277. doi: https://doi.org/10.1016/0010-0277(95)00661-H. URL `https://www.sciencedirect.com/science/article/pii/001002779500661H`. 1

[4] Steven Holtzen, Yibiao Zhao, Tao Gao, Joshua B. Tenenbaum, and Song-Chun Zhu. Inferring human intent from video by sampling hierarchical plans. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 1489–1496. IEEE Press, 2016. doi: 10.

1109/IROS.2016.7759242. URL https://doi.org/10.1109/IROS.2016.7759242. 2

[5] Siyuan Qi, Baoxiong Jia, Siyuan Huang, Ping Wei, and Song-Chun Zhu. A generalized earley parser for human activity parsing and prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2538–2554, 2021. doi: 10.1109/TPAMI.2020.2976971. 3

[6] Amanda L. Woodward and Jessica A. Sommerville. Twelve-month-old infants interpret action in context. *Psychological Science*, 11(1):73–77, 2000. ISSN 09567976, 14679280. URL http://www.jstor.org/stable/40063499. 1

[7] 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. 1