Inferring Goal Representation from Actions

Xizhi Xiao
Yuanpei College
Peking University
xxz3470608107@stu.pku.edu.cn

Abstract

Goals are invisible mental states representing the desired outcomes of actions. They are essential for understanding human behaviors and predicting their future actions. However, goals are not directly observable and need to be inferred from actions. In this essay, we review the recent progress in inferring goal representation from actions. We first introduce psychological study on goal inference in infants and children, and then review a well established Bayesian model of action understanding. We conclude with a discussion on the future direction of goal inference.

1 Introduction

Representing goals has been a challenging problems in AI research, because goals are purely invisible by pixels of images and need to be inferred from physically visible properties. In court, the severity of punishment is influenced by intentionality, and often the criminal suspect will explain their crime as unintentional or passive in order to mitigate their sentence. However, the judges can infer his or her inner intentions from the behavior. In the field of robotics, the robot needs to understand the goal of the human partner in order to cooperate with him or her. For example, when a human is approaching the door with the hand raising, the robot need to understand that the person is trying to open the door in order to assist him. All these examples show that goal inference is a fundamental ability for human and artificial intelligence.

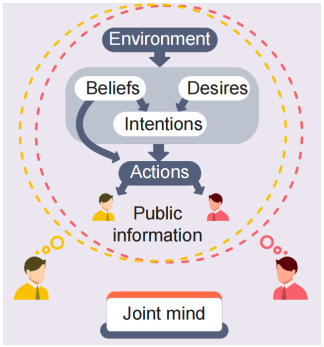

Figure 1: The Theory of Mind (ToM) framework (adopted from [4]).

Despite the significance and ubiquity of goal inference, it remains to be explored how humans infer goals from actions. In the Theory of Mind (ToM) framework (see Fig. 1), intentions are generated by beliefs and desires, and derive specific actions. Due to the rational principle of human actions, we can infer goals from observed actions, which is call the inverse planning [1]. In the remainder of the essay, we examine the advancements in inferring goals from actions both psychologically and computationally.

2 Psychological study

Humans have the tendency to infer others intentions from their actions. Even young children can seletively focus on the aspects of an actor's behavior that are relevant to his or her underlying intentions. 9-month-old infants looked longer when the actor grasped a new toy than when she moved through a new path, and 5-month-old infants showed similar, though weaker, patterns [5]. These results suggest that infants care little about the surface behaviors an actor exhibits, but instead focus more on the underlying intentions that drive those behaviors. Other research has discovered that infants' sensitivity to the actor's goal was correlated with their engagement in object-directed contact with the toys, which indicates that infants can rapidly form goal-based action representations and suggests a developmental link between infants' goal directed actions and their ability to detect goals in the actions of others [3]. Infants are able to distinguish unwilling intentions from unable intentions. When the experimenter was unwilling to give the toy to the infant, the infant reacted with more impatience (e.g., reaching, looking away) than when the experimenter was simply unable to give it [2]. All these research indicate that humans can infer others' intentions from their actions from a very young age.

3 A computational model

In this section, we introduce several well established computational framework for inferring goals or intentions from actions.

Baker et al. [1] formulizes action understanding as Bayesian inverse planning: the Bayesian inversion of models of probabilistic planning in Markov decision problems (MDPs). This framework uses MDPs to capture observers' mental models of intentional agents' goal- and environment-based planning. Given an MDP model of goal-directed planning, Bayesian inverse planning computes the posterior probability of a **Goal**, conditioned on observed **Actions** and the **Environment**, using Bayes' rule:

$$P(Goal|Actions, Environment) \propto P(Actions|Goal, Environment)P(Goal|Environment)$$

Inverse planning enables goal-based prediction of future actions in novel situations, given prior observations of behavior in similar situations. This framework has been solidly validated with psychophysical experiments and successfully applied to infer goals from human actions.

4 Conclusion and outlook

In conclusion, we have reviewed the recent progress in inferring goal representation from actions in the field of psychology and artificial intelligence. Further research could benefit from:

- **Goal inference with nonverbal signals.** In real life, people often use nonverbal signals to express their intentions, such as facial expressions, gestures, and body movements. Therefore, it is necessary to study how to infer goals from nonverbal signals.
- **Goal inference in multi-agent interaction.** In real life, people often interact with multiple agents at the same time. Current studies focus more on the interaction between two agents, and it is necessary to study how to infer goals in multi-agent interaction.

References

[1] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009. 1, 2

[2] Tanya Behne, Malinda Carpenter, Josep Call, and Michael Tomasello. Unwilling versus unable: infants' understanding of intentional action. *Developmental psychology*, 41(2):328, 2005. 2

[3] Jessica A Sommerville, Amanda L Woodward, and Amy Needham. Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, 96(1):B1–B11, 2005. 2

[4] Stephanie Stacy, Chenfei Li, Minglu Zhao, Yiling Yun, Qingyi Zhao, Max Kleiman-Weiner, and Tao Gao. Modeling communication to coordinate perspectives in cooperation. *arXiv preprint arXiv:2106.02164*, 2021. 1

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