Theory and Simulation of Goal Attribution

Jinfan He
Peking University
2200012979@stu.pku.edu.cn

Abstract

Humans have a tendency to interpret the actions of others as goal-directed, a tendency that manifests itself as early as infancy. Infants demonstrate the ability to differentiate between behavior influenced by external factors and behavior guided by an explicit goal. The goal attribution aids in predicting the actions of others and engaging in social learning. This paper introduces two possible cognitive hypotheses: teleological reasoning and action-goal association theory. It explains how both hypotheses fit into different aspects of human goal attribution ability. Then, a computational model is presented that emulates the human goal attribution ability by utilizing Bayesian inverse planning to represent potential goals based on observed action sequences and the environment.Finally, we discuss how to achieve goal attribution for abstract goals and highlight the potential issues associated with solely linking them to subgoals.

1 Introduction

Humans exhibit a strong tendency to interpret the observed behavior of others as goal-directed, a tendency that emerges as early as infancy. Researchers have found that infants as young as 15 months of age can understand that actions are guided by underlying goals, and they distinguish between the concept of action outcome and goal[3]. They are capable of recognizing which unexpected action outcomes are due to external factors and which stem from different goals of others. For example, when the experimenter attempted to hand them something and failed, the infants were able to discern whether this was the result of the experimenter's unwillingness or inability to do the task, to understand situations where the task could not be completed, and to maintain longer attention to instances of unwillingness.

This ability to interpret other people's actions as goal-directed aids in predicting their subsequent actions and final outcomes, while accounting for causal relationships between different actions. Simultaneously, the capacity for goal interpretation helps humans in social learning. When individuals associate an action with a goal, they perceive a causal relationship between the action and the goal. If they have not performed such an action before, they tend to view it as a new means of achieving a goal, showing a tendency to explore the causal relationship between the action and the goal. If causal relationships cannot be extracted from actions, individuals resort to fully imitating action sequences to achieve the desired outcome.

2 Some related hypotheses

So, how do people perceive goals from action sequences? [2]The teleological perspective posits that individuals understand all actions based on optimal solutions to their beliefs and goals. If an individual's behavior differs from the predicted outcome, one assumes the existence of unknown beliefs or different goals. Action-goal correlation theory suggests that when people see the result of an action, they take it as the goal of the action and immediately associate the two. Clearly, the former theory can better explain certain behaviors in life and infer underlying unknown beliefs, while the latter can help us analyze situations more quickly and make predictions. Both theories explain different aspects of the process of attributing goals to actions.

However, applying these psychological theories to artificial intelligence requires further consideration. For example, understanding how people adjust beliefs and goals to account for new situations when predictions and actions disagree, or how people judge which outcomes in a sequence of outcomes generated by an action sequence are desired goals.

3 A computational model for goal representation

An intuitive idea is to generate action sequences step by step based on the goal and the environment. We can use Bayesian inverse planning[1] to represent the possible goals as posterior distributions based on the observed action sequence and environment:

$$P(Goal \mid Actions, Environment)$$

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

If we have prior knowledge about possible goals in the environment and can generate planning policies for different goals and environments, we can infer goals from the environment and action sequences. However, this requires the agent to have human-like planning capabilities, and the goal may change with the action sequence, so we need to determine promptly which action sequences are generated by the same goal.

However, the space of possible goals is often vast, so using action-goal correlation theory can help us quickly narrow down the possibilities. We can consider objectives based on the probabilistic order of possible objectives for an action sequence.

4 Other problem: abstract goal representation

Another question is how to represent abstract goals. Currently, most studies focus on a sequence of behaviors that reflect possible concrete and immediate goals, but fail to represent and understand long-term, abstract goals.

This representation is undoubtedly challenging due to the lack of detailed criteria and the large volume of data generated over an extended period of time, such as maintaining a healthy body. From my understanding, people also lack an intuitive grasp of abstract goals and can only achieve them by associating them with specific subgoals.

Since abstract goals are difficult to evaluate quickly for optimal solutions, individuals tend to associate proximate subgoals that approximate the optimal achievement of the abstract goal..This approach reduces the amount of data to record action sequences associated with abstract goals..As a result, the attribution of abstract goals is often based on the attribution of subgoals, which is further facilitated by the correlation between subgoals and abstract goals.

However, this approach overlooks the collective effect of multiple subgoals and the influence of environmental factors, making it prone to errors in qualitative outcomes..For example, offering milk to someone is generally considered a gesture of kindness, but if given to someone who is lactose intolerant, it can undoubtedly become a harmful act.

Therefore, when analyzing abstract goals, attributing goals should not rely solely on the relevance of a particular subgoal to the abstract goal..Instead, it should consider other factors, such as beliefs, environment, and past behavior, in a comprehensive way.

5 Conclusion

This paper presents the psychological mechanism by which humans interpret the behavior of others as goal-directed, noting that this mechanism is formed early in infancy. This mechanism aids social learning and predicting the actions of others, and it accounts for two related hypotheses: teleological and action-goal correlation. A computational model is proposed to demonstrate the goal, which uses Bayesian inverse programming to infer the posterior distribution of possible goals based on the observed action sequences and the environment.Finally, we discuss how to achieve goal attribution for abstract goals and highlight the potential issues associated with solely linking them to subgoals.

References

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

[2] Gyo¨rgy Gergely Gergely Csibra. 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, 124:60–78, 2007. 1

[3] Josep Call Tanya Behne, Malinda Carpenter and Michael Tomasello Max Planck Institute for Evolutionary Anthropology. Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology*, 124(2):328–337, 2005. 1