Research on the Computational Framework of Goal Representation in the Context of Cognitive Artificial Intelligence

Yao Yuanhang[*]
Peking University
2000017813@stu.pku.edu.cn

Abstract

After reading the relevant literature listed in the course, I have gained insights into various methods of goal representation in the field of artificial intelligence. From the "intentional stance" framework to "teleological reasoning" and "inverse planning," these are proven effective methods. This essay will analyze the relationship between the former and the latter. Recognizing the significance of the latter two as mainstream computational frameworks, I will further analyze their respective advantages and disadvantages, as well as their practical effects and trade-offs. These studies and analyses will contribute to a more comprehensive understanding of the problem of goal representation in the field of artificial intelligence.

1 Introduction

In the era of cognitive artificial intelligence, accurately and faithfully simulating human thought processes has become a crucial research direction. Goals, along with beliefs, desires, and other cognitive processes, hold a pivotal position in human cognition. Therefore, the key to simulating human thought in cognitive artificial intelligence lies in the rational representation of goals. First and foremost is the "intentional stance" framework, a concept primarily derived from the field of psychology rather than focusing on computational frameworks. Studies on this framework, from Gyorgy Gergely's work in 1995 [3] to Baldwin's in 2001 [2], mostly originate from the psychological domain and explore how human individuals, including infants, infer the intentions of others. As a psychological framework, it comprises three main components: cognitive, emotional, and behavioral components (intention or intent).

The essay prompt requires us to discuss goal representation methods with computational frameworks. Although the "intentional stance" framework is rooted in psychological experiments, it can be translated into a computational framework. For instance, some studies utilize Bayesian models to infer attitudes in human social behavior, extracting social purposes such as helping or hindering from actions. "Intentional stance" is a concept in cognitive science, philosophy, and psychology, typically referring to psychological states like intentions and beliefs. To some extent, it can be seen as encompassing the psychological states of goals, intentions, etc., involved in the subsequent "teleological reasoning" and "inverse planning" that we will discuss. It serves as an overarching concept that coordinates the latter two methods. Therefore, our research will primarily delve into the computational frameworks, advantages, disadvantages, and application scenarios of the latter two methods.

2 Teleological Reasoning

Teleological reasoning is a type of inference that involves interpreting the actions of an agent as directed towards specific goals or purposes [4]. It is based on the explicit understanding of goals

[*] Use footnote for providing further information about author (webpage, co-first authors, *etc*.).

within the agent's environment, whether it be a robot, an intelligent system, or any other form of artificial intelligence entity. This reasoning process aims to comprehend the agent's behavior, viewing it as a series of intentional steps assumed to be taken to achieve a particular goal. In other words, it is unlikely to assume that intelligent agents would engage in purposeless actions not directed towards specific objectives.

Teleological reasoning is a theoretical framework with the core idea of regarding goals as states or outcomes that agents strive to achieve. These goals may be clearly defined at the beginning of a task or gradually inferred from the agent's task environment or context. In this reasoning process, the agent relies on its knowledge of the world to infer which specific actions would effectively move towards achieving the goal.

To achieve the goal, the agent continuously evaluates the current context and selects appropriate actions based on its understanding of the world. This involves considering various variables in the current environment, potential impacts, and possible obstacles. The agent utilizes information from its knowledge base to infer which actions are most advantageous for goal attainment. This reasoning process involves not only analyzing the impact of individual actions but may also include multi-step planning and sequences for more efficient goal achievement. During the inference process, the agent gradually takes actions to approach or achieve the goal. This may involve adjusting strategies, making decisions based on new information, and adapting to changes in the environment. Through continuous interaction with the environment and learning, the agent optimizes its behavior, enhancing efficiency and success rates in goal achievement.

In summary, teleological reasoning emphasizes the flexible application of knowledge and reasoning abilities by agents in the process of achieving specific goals. This theoretical framework provides valuable guiding principles for decision-making and behavior of artificial intelligence systems and other agent-like entities.

3 Inverse Planning

Inverse planning is an innovative planning method in the field of artificial intelligence, distinct from traditional forward planning methods. Forward planning typically starts from an initial state and generates a sequence of actions to reach the goal state. In contrast, inverse planning adopts a completely opposite strategy, initiating the process from the goal state and gradually deducing the sequence of actions required to achieve the goal.

The key shift in inverse planning lies in the reverse problem-solving approach [1]. The agent system or artificial intelligence algorithm sets the goal state as the starting point and, through reverse path planning, progressively retraces steps back to the current state. The advantage of this method is its direct focus on determining how to move from the current state to the goal state without the need to pre-generate a complete sequence of actions. Through the reverse deduction process, the system identifies a series of actions to achieve the goal, offering a more flexible and efficient planning approach.

Inverse planning demonstrates significant advantages in specific applications within the field of artificial intelligence, particularly in scenarios with a large problem space and complex action sequences, especially in the process of seeking goals. For example, in intelligent games, especially maze problems, inverse planning can assist intelligent agents (such as in-game characters) in effectively finding a path from the goal position back to the starting position. Consider a scenario where an intelligent character is in an unknown maze with the goal of returning to the maze's starting point. Traditional forward planning might require generating a complete path from the starting point to the goal and following that path. In contrast, inverse planning starts from the goal position, gradually deducing the sequence of actions needed to return to the starting point. This approach allows the character to focus more directly on how to initiate actions from the current position, leading directly back to the starting point. This inverse planning method proves practical in maze problems. By thinking in reverse, the character can deduce a path to the starting point based solely on the goal position, eliminating the need for a complete exploration of the entire maze structure. Such intelligent decision-making enhances the adaptability of the agent system, enabling it to find solutions more rapidly and efficiently.

In conclusion, inverse planning, as a unique planning strategy, provides an effective problem-solving approach for artificial intelligence systems. Emphasizing reverse thinking by deducing steps starting from the goal state allows the system to flexibly address various complex problems and tasks.

4 Compare the advantages and disadvantages of the two methods

The two methods highlighted above possess distinct advantages and limitations when addressing specific problems. Through an in-depth comparative analysis, our aim is to provide a basis for choosing the appropriate method in specific scenarios. Key considerations include the clarity of goals, the level of the agent's understanding of world knowledge, and the availability of computational resources.

4.1 the clarity of goals

Teleological reasoning performs exceptionally well when the agent has a clear understanding of its goals. For explicitly defined objectives, such as "move the robot to a specified location," teleological reasoning can directly represent this goal as a state the agent is attempting to achieve, making the problem intuitive and easy to handle. Conversely, inverse planning is more suitable for situations where the agent's goals are not explicitly defined. By observing the agent's actions, the system can infer the goals the agent might be attempting to achieve, adding flexibility and applicability to scenarios with ambiguous objectives.

4.2 the level of the agent's understanding of world knowledge

Teleological reasoning performs better when the agent has a clear understanding of the world's knowledge. When the agent accurately comprehends the relationship between goals and the environment, teleological reasoning can effectively deduce which actions will contribute to goal achievement. Meanwhile, inverse planning is relatively more suitable for situations where the agent's knowledge of the world is incomplete. By observing actions, the system can deduce goals in reverse, adapting to the incompleteness of agent knowledge, but it may perform less effectively when the agent's understanding of the environment is limited.

4.3 the availability of computational resources

Teleological reasoning may be computationally expensive, especially when dealing with complex goals or incomplete agent knowledge. The need for intricate logical reasoning can result in higher computational costs, particularly when the goals are complex or unclear.

In contrast, inverse planning is generally less computationally expensive. It can adapt to the incompleteness of agent knowledge and performs relatively well in handling fuzzy goals. However, it may require a substantial amount of data to effectively deduce the agent's goals.

5 Select a method based on the application scenario

After a detailed analysis of the strengths and weaknesses of teleological reasoning and inverse planning, it becomes natural to choose the more suitable method based on the specific application scenario. For situations with explicit goals and clear knowledge, teleological reasoning might be the preferred choice. It provides an intuitive problem representation and clear logical reasoning, suitable for cases where the agent has a clear understanding of goals and the environment.

In scenarios where goals are not explicitly defined, and environmental knowledge is relatively incomplete, inverse planning may be more appropriate. It adapts by reverse inference of goals through observing agent actions, making it suitable for handling fuzzy goals or incomplete agent knowledge. In cases of limited computational resources, inverse planning could be considered, as it is generally less computationally expensive, especially suitable for dealing with environmental uncertainty or limited data.

In summary, whether to choose teleological reasoning or inverse planning depends on the specific problem scenario and system requirements.

References

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

[2] Simon Baron-Cohen. Discerning intentions in dynamic human action–it's all in the timing. *TiCS*, 5(4):171–178, 2001. 1

[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. 1

[4] . . *()*, 13(1):1–5, 2011. 1