CoRe Essay 6
Intentionality

Yuntian Gu
Yuanpei College
Peking University
guyuntian@stu.pku.edu.cn

Abstract

Understanding and representing goals within artificial intelligence systems is a critical challenge that bridges the gap between simple task execution and complex decision-making. This essay reviews the current paradigms in cognitive AI, with a specific focus on the hierarchical organization of goals and their non-observable nature. We discuss various computational models for representing goals, examining their advantages and disadvantages in details.

1 Introduction

The concept of intentionality is pivotal in understanding human cognition and behavior. It serves as the bedrock for interpreting a wide array of complex and dynamic actions that unfold in the real world. This multifaceted nature of intentionality not only influences interpersonal relationships and social dynamics but also has profound implications for the development of artificial intelligence (AI). In recent years [4], there has been an increasing emphasis on creating AI with humanlike common sense—a shift from a purely reactive system to one capable of understanding, predicting, and acting upon a set of structured goals. This paper delves into the intricacies of goal representation in computational systems, anchoring the discussion in contemporary literature that explores cognitive AI and the discernment of human intentions.

In everyday life, goals manifest in various forms: as immediate tasks to be accomplished, as stepping stones toward a larger objective, or as abstract, long-term visions that guide behavior over extended periods. The translation of this multifarious goal landscape into a computational framework presents both opportunities and challenges. It requires a system that can not only execute predefined tasks but also exhibit the flexibility and foresight characteristic of human planning and decision-making.

2 Hierarchical Goal Representation

Hierarchical goal representation in computational frameworks is predicated on the idea that goals can be decomposed into a series of subgoals, which are organized in a manner that reflects their interdependencies and relative importance. This framework is inspired by cognitive models of human decision-making [2], where long-term plans are operationalized through a series of smaller, more immediate tasks. In cognitive AI, this principle is captured in systems designed to mimic human common sense and reasoning, which necessitates an understanding of how goals are structured and how they drive behavior.

2.1 Advantages

By dividing complex goals into manageable units, AI systems can tackle them with greater efficiency and flexibility. Hierarchies allow for the easy addition or removal of goal components, which makes the system scalable to various complexities and sizes of problems. Also, subgoals at lower levels of the hierarchy can be context-specific, allowing the system to adapt to different situations by activating

relevant sub-trees. Updates to the system's knowledge base or goal specifications can be localized to specific subgoals without necessitating a complete overhaul of the goal structure.

2.2 Disadvantages

Crafting a comprehensive hierarchy that accurately reflects the nuances of goal relationships can be daunting and resource-intensive. While hierarchies provide structure, they can also introduce rigidity, making it difficult for the system to adapt to unforeseen circumstances that do not fit neatly within the predefined goal structure. The maintenance and traversal of a goal hierarchy can introduce computational overhead, especially in dynamically changing environments where goal priorities shift frequently. Also, managing dependencies between goals can become complex, especially when changes to one goal can have cascading effects on the hierarchy.

3 Dark Goal Representation

Dark goal representation is a term derived from the cognitive AI literature to describe goals as entities that exist beyond the observable, pixel-represented data, akin to the unobservable mental states in humans that drive behavior [1, 3]. This concept stems from the realization that intentions can be inferred from the aggregation of actions, environmental contexts, and their outcomes over time, rather than from any single, immediate perceptual cue. In AI, this translates to the construction of internal models that predict and simulate the potential trajectories of actions to achieve a set goal, without explicit and immediate feedback.

3.1 Advantages

The advantage of dark representations lies in their ability to capture the subtlety and complexity of real-world scenarios where the intention behind actions is not immediately clear. Systems with dark goal capabilities can operate in environments where the state space is poorly defined or where sensory data do not directly indicate the purpose of actions. For example, in social robotics, a robot might use dark goal representation to infer the intentions behind human gestures or expressions in order to respond appropriately, even if it has never encountered those specific gestures or expressions before.

Moreover, dark goal representation aligns with the predictive processing model of the human brain, which posits that the brain is constantly predicting sensory input and updating its models based on prediction errors. AI systems using this framework can continuously refine their goal-directed behavior as new data is processed, leading to a form of artificial common sense.

3.2 Disadvantages

The primary challenge with dark goal representation is the complexity involved in accurately inferring intentions from sparse or noisy data. It requires extensive training data and robust error correction mechanisms to ensure that the system does not misinterpret the data and form incorrect goal-oriented predictions. This could lead to AI behaviors that are unpredictable or difficult to explain, which poses challenges for transparency and trustworthiness.

Additionally, there is a risk of overfitting, where the AI system becomes too tailored to the training data and unable to generalize its goal inference to new, unseen contexts. This could limit the system's flexibility and adaptability, which are crucial for operating in dynamic real-world environments. Finally, the computational cost of training and running models that support dark goal representation can be significant, demanding substantial resources and potentially limiting the scalability of such systems.

4 Conclusion

The task of representing goals in artificial intelligence is not merely a technical challenge but also a conceptual one, reflecting deep questions about the nature of agency and intelligence. The hierarchical model of goal representation brings an organized structure to AI systems that mirrors human planning and execution processes, allowing for a clear delineation of steps towards achieving complex objectives. However, as with human endeavors, the rigidity of predefined hierarchies often

fails to accommodate the fluidity of changing circumstances, reflecting a key limitation in the current design of adaptive AI systems. Dark goal representation, on the other hand, pushes the boundaries of AI's capability to infer and act upon unobservable indicators. This model's strength lies in its potential for profound contextual understanding, predicting behavior, and proactive interaction within an environment. However, it also brings to the forefront the perennial issue of interpretability and the difficulty of ascertaining the "correctness" of inferred intentions. Ensuring the transparency and accountability of AI systems with dark goals is crucial, particularly in applications that require trust and explainability, such as healthcare, law, and autonomous vehicles.

References

[1] Dare A Baldwin and Jodie A Baird. Discerning intentions in dynamic human action. *Trends in cognitive sciences*, 5(4):171–178, 2001. 2

[2] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017. 1

[3] Dan Xie, Sinisa Todorovic, and Song-Chun Zhu. Inferring" dark matter" and" dark energy" from videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2224–2231, 2013. 2

[4] 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