From Mechanisms of Goal Attribution to a Review of Grammar Parsing and Inverse Inference

Xinyi Yang
Department of Automation
Tsinghua Universtiy
xy-yang21@mails.tsinghua.edu.cn

Abstract

Humans have a strong inclination to and are also good at, interpreting others' actions as a series of goals driven by intentions. In this essay, we analyze three mechanisms that account for this in humans. One of them may be better depending on the task demands and the information available, but they support each other under most circumstances. Thus, as for computational frameworks, we pay more attention to how they adapt to these three mechanisms. From this perspective, two classical methods, grammar parsing and inverse inference, are discussed about their advantages and limitations.

1 Introduction

When humans observe others in motion, they usually care little about the surface behaviors displayed, but they have a strong inclination to interpret events as a series of goals driven by intentions.[4] Cognition experiments [7] also show that even 12-month-old infants can take the "intentional stance" in interpreting the goal-directed spatial behavior of a rational agent. This goal-recognition capability is essential for social interaction, which enables humans to better understand others' mental states (beliefs, desires, and intentions) and predict others' behaviors when engaging in cooperative activities or a simple social occasion. For example, when a man walks towards a closed door with hands full of books and asks for your help, it will be really awkward if you pick a book from him and start to read. The whole issue is, that the goal of him is going through that door but not showing you these books.

Humans are so adept at inferring the mental states underlying other agent's actions, that the above embarrassing situation rarely happens. However, how to represent goals in artificial intelligence? Goals can not only be immediate effects of short-term actions, but also high-level results of long-term sequences of actions. A kid reaches for a toy as his goal is exactly to grasp it; Tom takes a bottle of milk from the refrigerator, walks towards the cabinet and takes out a cup, the combination of these actions implies that he may want to drink milk with the cup. Moreover, an action could be done to achieve different goals, and a goal could be achieved in different ways and each sub-goal towards the goal could be reached by countless routes. Take Tom's example again, the action of taking out a cup may be for drinking milk, water, or something else which depends on other actions and the environment. Both drinking straight from the bottle and pouring milk into a cup will achieve the goal of drinking milk. Since these problems exist, it is hard to build a computational framework to model goal inference in humans.

There are two mainstream kinds of methods proposed: grammar parsing [16] (receives a sequential input and applies a grammar parser to obtain a representation of the combination of actions) and inverse inference [2] (models the intuitive causal relation between beliefs, goals and actions as planning, and invert this relation to infer beliefs and goals from actions). In order to analyze their advantages and disadvantages, in this essay we pay attention to how well they adapted to goal attribution mechanisms in humans. Therefore, in Sec. 2, three distinct mechanisms proposed based on observations of human cognition will be described in turn. Then grammar parsing and inverse inference methods will be reviewed with these three mechanisms in Sec. 3. At last, we will give the conclusion in Sec. 4.

2 Mechanisms of Goal Attribution in Humans

2.1 Action-Effect Associations

In action planning, the ideomotor principle [11] emphasizes the role of goal representation in the generation of motor actions. Based on this view, the representation of goals in the actor's cognitive system focuses on action-effect representations and their bidirectional associations. Thus, the goal is defined as the desired effect of the corresponding action. These links are established by simple associations upon observing the effects that one's actions have produced, and these associations start to build up from early on in infancy.[4] For example, when 6-month-old infants see a hand repeatedly grasping one of two objects, they anticipate that the same object will be grasped again even when the spatial location of the objects is rearranged.[20] Infants indeed tend to attend to the effects of actions they observe and expect the actor the produce the same effect again.

Many of the routine goal-directed actions are performed by humans the same way every time towards the same goal. When goals are attributed on the basis of observed action-effect associations, it will be easy and fast for the observer to infer the goal state relies on the assumption that an action is directed towards the same goal state that has been produced earlier, and the same goal state will be achieved by an actor in a similar way. However, this mechanism is severely limited to the observer's own knowledge about action-effect associations. If the observed action is novel, or if the current environment does not afford the actor to reach its goal in a similar way, this mechanism does not offer a solution. And it can not deal with the situation "an action *vs*. multiple effects" since there is no further selection to identify the goal according to the particular environment.

2.2 Simulation Procedures

According to Goldman [8], humans understand other agents' mental states by imaging themselves in others' position, and simulatively generating the mental states that they would possess were they in the same situation. Based on this view, goal-directed actions understanding is achieved through a predictive simulation procedure in which the goal ascribed to the actor is taken as an input and output actions that the observer self would perform to reach the goal. After applying the simulation procedure in the opposite way, the likely goal could be recovered from the observed actions. Theoretical evidence of this mechanism lies in the discovery of mirror neurons in motion keys.[17] It is also found that the sudden emergence of certain social cognitive skills around 9 months of age occur simultaneously with the emergence of infants' own unfolding capacity for means-end actions which could be used to interpret others' actions through simulation.[19]

The advantage of relying on such simulation procedures rather than empirical laws, is that the observer can exploit her/his own existing mental mechanisms to link the goals and actions. Based on the assumption that the observed actor has the same motor constraints and preferences as the observer [12, 13], an effective simulation may lead to valid goal inferences by reducing the possible range of solutions on many occasions. However, this mechanism does not work when the above assumption isn't meet, *e.g.*, a non-human actor, or action that could not be performed due to individual motor deficits.

2.3 Teleological Reasoning

The principle of rational action [3], which emphasizes the relative efficiency of the action performed to achieve the goal within the current environment constraints given, demonstrates that interpreting an action as goal-directed should incorporate normative evaluation. Whether or not an outcome may be seen as the goal depends on whether the outcome is judged to justify the action in the given situation.[4] Thus, based on this principle, teleological reasoning follows the assertion that a mentalistic action explanation is well-formed (and therefore acceptable) if, and only if, the action (represented by the agent's *intention*) realizes the goal state (represented by the agent's *desire*) in a rational manner within the situational constraints (represented by the agent's *beliefs*).[6] (Fig. 1) In a violation-of-expectation study[7], twelve-month-olds looked longer at the action that seemed an inefficient means to the goal, but showed no dishabituation to the most efficient means to the goal. It is indicated that by 12 months infants can (1) interpret other's actions as goal-directed, (2) evaluate which one of the alternative actions available within the constraints of the situation is the most efficient means to the goal, and (3) expect the agent to perform the most efficient means available.

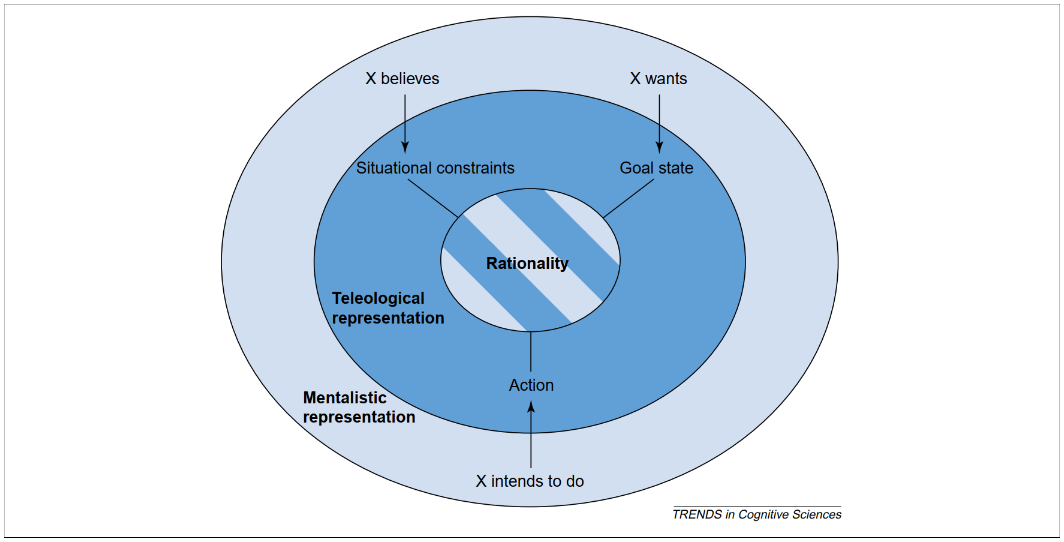

Figure 1: Teleological and mentalistic representations of actions.

Since biological systems tend to conserve energy, this mechanism is likely to produce valid predictions and goal attributions, and has been shown to be a computationally viable way (*e.g*., Baker et al. [1] and Baker et al. [2]). Furthermore, it can also solve some of the "an action *vs*. multiple effects" problems encountered with Action-Effect Associations, and the similar problem "multiple actions *vs*. an effect". However, its accuracy depends heavily on the agent's beliefs about the current environment, and such beliefs are always not equivalent to the real environment. Insufficient or inaccurate knowledge about the constraints of the actor or the situation may produce wrong predictions or goal attribution by teleological reasoning.

It's worth noting that these three mechanisms do not compete but complement each other.[4] First, depending on the task demands and the information available, one or the other mechanism would provide faster or more valid answers and none of them is better in all situations. Second, these mechanisms usually support each other during their implementation.

3 Grammar Parsing and Inverse Inference Methods Review

3.1 Grammar Parsing

Grammar Parsing here refers to inferring goals through recognition and segmentation of long-term and complicated activities from vision with grammar models. Take the activity "making cereal" as an example, Fig. 2 [16] shows a temporal grammar representation. With such compositional/hierarchical models on actions, we can both infer low-level direct goals, *e.g*., "getting milk", and high-level abstract goals, *e.g*., "saving an empty stomach". There have been several grammar-based methods proposed. Pei et al. [14] detected atomic actions and used a stochastic context sensitive grammar for video parsing and intent prediction. Pirsiavash and Ramanan [15] described simple grammars that capture hierarchical temporal structure while admitting inference with a finite-state-machine, which makes parsing linear time, constant storage, and naturally online. Qi et al. [16] extended the work by generalizing the Earley parser to parse sequence data which is neither segmented nor labeled.

When utilizing such a temporal grammar representation of actions to infer the goal, an intuitive method would be taking the predicted final environment state as the agent's goal. During this process, a series of action-effect associations form a sequential whole that contributes to the outcome together. Given the past observations, this hierarchical structure could model non-Markovian events which are very common in humans. And because the final state is explicitly reflected in the structure, the inference is also explicable. However, parsing inference also inherits the shortcomings of action-

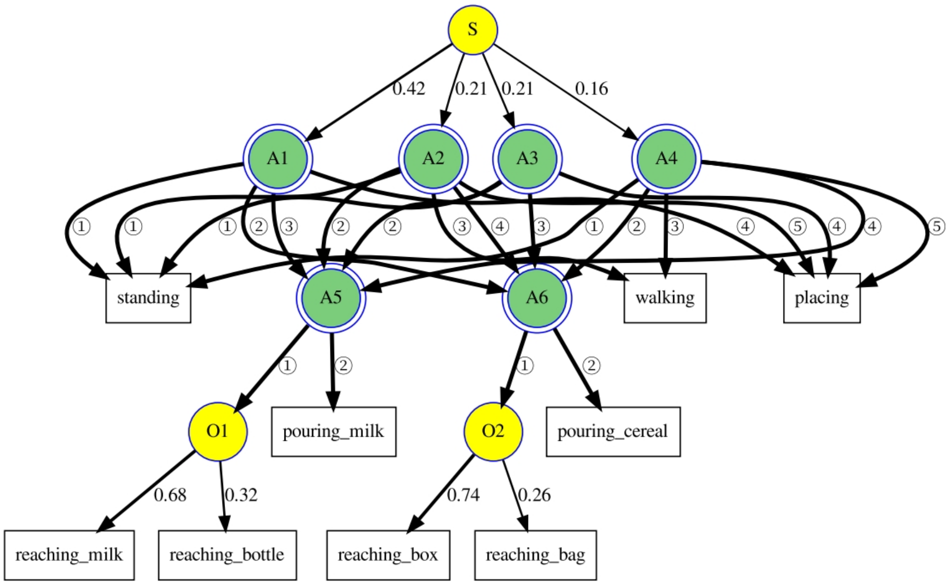

Figure 2: An example of a temporal grammar representing the activity "making cereal". The circled numbers on edges indicate the temporal order of expansion. The green nodes are And nodes while the yellow nodes are Or-nodes. The number on the branching edges of Or-nodes represents the branching probability.

effect associations. The same final environment state may reflect different goals, for example, the state that a computer and coffee are neatly laid out on the desk implies the agent may want to set the desk for work or playing. As Huang et al. [10] demonstrated, large language models (llms) can be prompted to generate plausible goal-driven action plans for goal-driven tasks in interactive, embodied environments. It is considerable to prompt llms with grammar representations and consequent changing states of the environment to generate inferred goals from the inverse process of planning. Just like simulation in humans, there exist simulation procedures according to the world knowledge incorporated in llms. And as being trained on large corpora of human-produced language, these models are thought to contain a lot of information about the world.[18] So if prompting together with personal characteristics of the agents, the assumption of the mechanism of simulation procedures will meet under most circumstances.

3.2 Inverse Inference

Inverse inference here refers to inverse planning based on Bayesian, which is formularized as:

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

For modeling human goal understanding. Formalisms for the inverse process are often divided into model-based and model-free approaches. Most model-based approaches model the intuitive causal relation between beliefs, goals and actions as rational probabilistic planning in Markov decision problems (MDPs), and invert this relation to infer beliefs and goals from actions. Baker et al. [2] proposed a framework for goal inference, in which the bottom-up information of behavior observations and the top-down prior knowledge of goal space are integrated to infer underlying intent. Holtzen et al. [9] presented a method to infer hierarchical intentions by reverse-engineering decision-making and action-planning processes in human minds from partially observed RGB-D videos, moving from the symbolic input to real vision input. Xie et al. [21] extended to outdoor scenarios, which inferred object functionality and human intent by reasoning about human activities. For model-free approaches, Daw et al. [5] considered dual-action choice systems from a normative perspective, using the computational theory of reinforcement learning.

When inferring goals and intents through inverse planning, the adoptive intuitive theory of rational planning is strongly based on the mechanism of teleological reasoning, *i.e.*, agents will plan approximately rationally to achieve their goals, given their beliefs about the world. It is also a simulation process. Consequently, inverse inference can effectively model human mental states and environmental constraints. However, the same as teleological reasoning, the accuracy of beliefs has a heavy influence. In some methods, the researchers assumed that the agent's action depends directly on the environment and the goal, without a separate representation of the agent's beliefs.[2] But with vision input, the perspective is often too limited to acquire whole observations of the environment. In addition, the solution to the inverse problem requires strong prior knowledge of the structure and content of agents' mental states, which may need complex design. And evaluating a potentially very large space of possible mental state interpretations is also very costly.

4 Conclusion

From a philosophical perspective, three mechanisms of goal attributions, *action-effect associations*, *simulation procedures* and *teleogical reasoning*, are all found evidence early and common in humans. They have their respective strengths and weaknesses, and complement each other. Two classical methods, *grammar parsing* and *inverse inference*, both have made progress in computational frameworks for goal inference/detection. After reviewing main researches, it is found that grammar parsing well adapts to action-effect associations, and inverse inference well adapts to simulation procedures and teleological reasoning. We argue that these two methods may inherit the strengths and weaknesses of the corresponding mechanisms. However, mechanism integration may be a considerable choice to improve.

References

[1] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian models of human action understanding. *Advances in neural information processing systems*, 18, 2005. 3

[2] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009. 1, 3, 4, 5

[3] Gergely Csibra and György Gergely. The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental science*, 1(2):255–259, 1998. 2

[4] Gergely Csibra and György Gergely. obsessed with goals: Functions and mechanisms of teleological interpretation of actions in humans. *Acta psychologica*, 124(1):60–78, 2007. 1, 2, 3

[5] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12): 1704–1711, 2005. 4

[6] György Gergely and Gergely Csibra. Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, 7(7):287–292, 2003. 2

[7] 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, 2

[8] Alvin I Goldman. *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press, 2006. 2

[9] 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)*, pages 1489–1496. IEEE, 2016. 4

[10] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022. 4

[11] William James. *The principles of psychology*, volume 1. Cosimo, Inc., 2007. 2

[12] Andrew N Meltzoff. Imitation and other minds: The like me hypothesis. *Perspectives on imitation: From neuroscience to social science*, 2:55–77, 2005. 2

[13] Andrew N Meltzoff. The like meframework for recognizing and becoming an intentional agent. *Acta psychologica*, 124(1):26–43, 2007. 2

[14] Mingtao Pei, Yunde Jia, and Song-Chun Zhu. Parsing video events with goal inference and intent prediction. In *2011 International Conference on Computer Vision*, pages 487–494. IEEE, 2011. 3

[15] Hamed Pirsiavash and Deva Ramanan. Parsing videos of actions with segmental grammars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 612–619, 2014. 3

[16] 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, 2020. 1, 3

[17] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27: 169–192, 2004. 2

[18] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020. 4

[19] Michael Tomasello. *The cultural origins of human cognition*. Harvard university press, 2009. 2

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

[21] Dan Xie, Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Learning and inferring dark matter and predicting human intents and trajectories in videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1639–1652, 2017. 4