Constructing Intentionality in AI Agents: Balancing Object-Directed and Socio-Technical Goals

Boyuan Chen
Yuanpei College
Peking University

Abstract

In this paper, we explore methodologies for constructing intentionality in AI agents, focusing on object-directed and socio-technical goals. Through Bayesian networks and reinforcement learning, simple agents can be designed to follow both immediate and long-term objectives. Complex agents, meanwhile, require socio-technical goal construction that incorporates social values and multi-agent interactions, often achieved through preference-based models and game-theoretic approaches. Despite recent advances, significant challenges remain, particularly around goal misgeneralization and power-seeking tendencies that may emerge due to misaligned optimization processes. Our work discusses potential risks and suggests pathways, including cooperative inverse reinforcement learning, to enable safer and more effective alignment of AI agents with human intentions in diverse and dynamic settings.

1 Introduction

We live in a world filled with goal-directed agents. Values shape these goals, *e.g.*, water flows downhill to minimize potential energy, and intelligent agents form intentions to achieve higher rewards. In human society, values and intentions play a crucial role, particularly in social interactions beyond the physical realm. Humans need to develop their own intentions and infer those of others to interpret behavior in a social context. As causal agents, people dedicate their limited time and resources to actions that align the world with their intentions and desires. They pursue their goals rationally, aiming to maximize utility and minimize costs based on their understanding of the world, which known as *the rationality principle* [10, 11, 17].

Numerous studies have explored methods for intention construction and inference. For example, handcrafted reward functions have been used to set training objectives for RL agents [9], and Bayesian networks have been applied to model relationships between goals and actions [4, 3, 2] to aid social agents in goal inference. Research on intentionality is crucial for developing causal agents and facilitating AI agents' integration into social interactions.

Goals can be either long-term or immediate, and may relate to both social and technical aspects. They can also be abstract (*e.g.*, helpfulness) and therefore difficult to define with a specific function, posing challenges for intentionality construction. In this paper, we focus on intentionality construction from two perspectives [21]:

- Creating simple agents with object-directed goals that do not consider the goals of other agents (Section 2). The key in this case is constructing both immediate and long-term goals, potentially spanning years.
- Developing complex agents with either social or object-directed goals that take into account the goals and likely behaviors of other agents (Section 3). Here, the main challenge is constructing abstract goals related to social values, enabling the intentionality construction method to generalize and scale for human-AI interactions.

Finally, we discuss potential issues with intentionality construction, such as *goal misspecification* [16], which could lead to *goal misgeneralization* or even *power-seeking* behaviors [18] during subsequent optimization (Section 4).

2 Object-directed Goals Construction of Simple Agents

To enable agents to accomplish specific tasks, we typically define a set of goals, which can be modeled directly through logical relationships, represented with a Bayesian Network (BN) to capture causal dependencies, or expressed through a reward function.

2.1 Bayesian Networks for Intent Construction

Bayesian Networks (BNs) have seen extensive application in *inverse planning tasks* [24], where their strength in capturing probabilistic dependencies and causal relationships enhances their utility for intent recognition across various settings. By modeling the likelihood of different intentions based on observed actions, BNs allow for a nuanced understanding of underlying goals and motives [23]. This probabilistic approach facilitates the mapping of visible behaviors to inferred intentions, making BNs invaluable for scenarios where interpreting intent accurately is crucial.

These methodologies have also proven effective in the realm of video modeling. In particular, inverse planning on RGB-D videos, which capture both color and depth information, enables a hierarchical interpretation of human actions. The multi-layered approach not only captures observable actions but also aligns them with inferred decision-making and planning processes, offering insights into the layered nature of intentions behind actions [20, 14]. By structuring human intent within such hierarchies, BNs bridge observed behaviors with probable intentions, facilitating a more comprehensive understanding of the human action-planning continuum.

However, BNs have certain limitations. They can struggle with scalability in high-dimensional or complex environments, as the computational cost of inference increases with the number of variables. Additionally, BNs rely heavily on accurate prior knowledge of causal structures, which may not always be available in dynamic or uncertain contexts. Furthermore, while BNs can model static causal dependencies effectively, they can be less suited for modeling sequential dependencies in continuously evolving scenarios.

2.2 Reinforcement Learning for Goal Specification

Another commonly used approach for goal specification is reinforcement learning (RL), where reward functions are designed to encourage desired behaviors from the model and penalize undesirable ones. Early efforts in reward function design were typically handcrafted [9]. In these scenarios, such as the frozen lake environment [6], goal-oriented behaviors like *finding the exit* can be guided effectively through simple positive and negative reward values. In essence, these environment rewards help agents model the environment's characteristics and learn how to explore and utilize it to achieve specified goals. Setting these environmental rewards thus serves as a means for human supervisors to communicate goal specifications to the agents.

The advent of deep neural networks has accelerated the application of reinforcement learning but has also introduced new challenges. For complex RL agents, we might want them to acquire specific human-like behaviors, such as opening and closing doors. However, these tasks can be multifaceted and have numerous pathways, making handcrafted reward functions insufficient to capture their complexity. Imitation learning offers a solution [1, 8] by enabling agents to learn directly from expert trajectories or behavior examples. By fine-tuning on these examples, the model's parameters can embody specific goals as modeled through these trajectories and behaviors.

Imitation learning is not the optimal solution, as human-provided demonstration data can be costly and difficult to collect. Moreover, not all tasks are feasible for human demonstration, such as autonomous drone control or robotic manipulation. To address these limitations, inverse reward design [13] attempts to construct goal specifications by reverse-engineering reward functions from existing trajectories, allowing the agent to infer and pursue goals based on observed behaviors. This approach shares a conceptual similarity with inverse planning tasks used for intent inference.

Preference-based RL also tackles these challenges by incorporating humans in the outer loop. Rather than providing demonstrations, humans merely rate the agent's behaviors. This feedback is then used to train a preference-based reward model that reflects human intent, highlighting what constitutes effective operational behavior [22].

Despite the effectiveness of RL methods for relatively simple agents, challenges remain, particularly regarding how to model abstract human values that go beyond immediate task-specific rewards. For instance, while simple reward functions may suffice for straightforward tasks, representing higher-order human values or complex ethical considerations within RL frameworks is still an open question.

2.3 Long-Term Goals Construction

The discussion above highlights that immediate goals are relatively straightforward to model. However, when goals extend to long-term horizons, it becomes particularly important to ensure that the model can accurately represent these long-term objectives and maintain their influence over time without diminishing as the time horizon increases.

Hierarchical Reinforcement Learning (HRL) [5] aim to address this issue by decomposing complex tasks into a hierarchy of sub-goals. This approach enables the model to pursue a high-level overarching goal by achieving finer-grained, intermediate objectives that collectively contribute to the realization of the broader goal. This concept mirrors human intentions, where we often define a macro-level goal accompanied by micro-level tasks that guide us step-by-step towards achieving it.

The construction of long-term goals is critical for planning tasks. However, there is still considerable debate over whether long-term goals need to be explicitly represented. For instance, individuals may struggle to clearly define their ten-year goals due to the inherent uncertainties of the future. Similarly, human experts, such as chess players, may not plan every move ten steps ahead; instead, they often rely on evaluating the current state of the board to make strategic decisions. This insight encourages us to examine the relationship between long-term goal construction and abstract reasoning. Through abstraction, past experiences can inform future planning, a process that offers valuable insights into building intentionality in agents. Since abstraction serves as a foundation for intelligent behavior and long-horizon reasoning, it forms a crucial basis for constructing intentional, long-term planning capabilities in intelligent agents.

3 Social-techinical Goals Construction of Complex Agents

Constructing goals for single-agent systems is relatively straightforward. However, when the problem extends to multi-agent scenarios, the complexity of intentions significantly increases. In these settings, agents are expected to account for the goals and anticipated behaviors of other agents. The primary challenge here is to construct abstract goals that encompass social values, thereby enabling the intentionality framework to generalize and scale effectively for human-AI interactions.

Modeling social and moral values presents particular difficulties, as it is challenging to find a precise computational representation for complex human values, such as fairness and justice. One compromise is to approximate human intentions by modeling their projection through human preferences. For example, a human supervisor may favor actions that are helpful or safe. Such preferences can be represented through preference-based models like the Bradley-Terry Model [15] or the Luce-Shepard Model, which are commonly used to model binary preferences or rank-order relations. This approach is widely employed in aligning large language models [19]. While effective to an extent, this approach is constrained by data quality and the risk of misgeneralization in preference models, potentially leading to goals that are misaligned or biased [7].

To further explore social values within multi-agent scenarios, we can reframe the problem using a game-theoretic perspective, examining two fundamental cases: cooperation and antagonism. Consider agents A and B; in a cooperative scenario, A's utility function includes B's welfare, and A aims to maximize B's benefit. In contrast, antagonism implies the opposite. By incorporating this structure into a bayesian network [21], we can model agents that act approximately rationally, given environmental constraints and their understanding of the intentions of other agents.

However, such modeling approaches have limitations. They often struggle to capture the nuances of fluctuating social contexts and may oversimplify complex interpersonal dynamics, reducing the reliability of the inferred intentions in highly variable multi-agent environments.

4 Disucssion

In the above analysis, we examined computational frameworks for constructing both object-directed and socio-technical goals. However, as Goodhart's Law warns, when means are mistaken for ends, the original objective is often compromised. As intentions grow more complex, precisely modeling them becomes increasingly difficult. Some objective functions may fail to account for spurious features, which agents could exploit as loopholes, leading to undesirable goal generalization in new contexts, even if the goals seem appropriate within human-observed environments.

Such goal misgeneralization, often due to distributional shifts, can result in policies that advance high-level but unintended objectives in new settings. Moreover, models may generalize to broadly-scoped goals that produce unintended consequences, especially with longer time horizons or unfamiliar tasks. For instance, a large language model might overextend *following instructions* to behaviors unforeseen by developers. This risk is exacerbated by simplicity bias in training algorithms, which may favor generalized, less context-specific goals (like obedience) over specific, bounded ones. Misgeneralization of broadly-scoped goals can drive models toward actions misaligned with intended outcomes—potentially even leading to shortcut-seeking behaviors—particularly in complex tasks like scientific research or organizational management, where oversight is challenging and the impacts are significant.

Mispecified goals may also be amplified under RL optimization, where the model's objective is to maximize rewards. This drive can lead models to "overachieve" or pursue goals at any cost, potentially resulting in aggressive strategies like resource exploitation or attempts to gain control. One approach to address this issue is to avoid rigid goal-setting and instead foster human-AI collaboration, where the AI seeks to model and advance human objectives dynamically. This is the essence of cooperative inverse reinforcement learning [12], which enables AI systems to align closely with human intentions by continuously adapting to human feedback, thereby reducing the risk of undesirable or extreme goal pursuit.

5 Conclusion

Our study underscores the nuanced challenges of constructing intentionality in AI agents, particularly as their roles extend into social and multi-agent contexts. Simple goal models can be implemented using Bayesian networks and reinforcement learning, but these methods face limitations in scaling to abstract human values and social interactions. Complex agents, designed to engage in socio-technical tasks, require flexible models that consider the goals and likely behaviors of other agents. However, as demonstrated, goal misgeneralization remains a critical risk, where agents might exploit optimization shortcuts, resulting in behaviors that deviate from human-aligned objectives. To mitigate these risks, we emphasize adaptive frameworks like cooperative inverse reinforcement learning, which can incorporate ongoing human feedback to dynamically align AI behaviors with intended human goals, thereby enhancing the reliability of agent behavior in complex environments.

References

[1] Alexandre Attia and Sharone Dayan. Global overview of imitation learning. *arXiv preprint arXiv:1801.06503*, 2018. 2

[2] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011. 1

[3] Chris L Baker, Joshua B Tenenbaum, and Rebecca R Saxe. Goal inference as inverse planning. In *Proceedings of the annual meeting of the cognitive science society*, volume 29, 2007. 1

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

[5] Matthew Michael Botvinick. Hierarchical reinforcement learning and decision making. *Current opinion in neurobiology*, 22(6):956–962, 2012. 3

[6] G Brockman. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016. 2

[7] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023. 3

[8] Pim De Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[9] Damien Ernst and Arthur Louette. Introduction to reinforcement learning. *Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P*, pages 111–126, 2024. 1, 2

[10] 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

[11] Herbert Gintis. A framework for the unification of the behavioral sciences. *Behavioral and brain sciences*, 30(1):1–16, 2007. 1

[12] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 29, 2016. 4

[13] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[14] 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. 2

[15] Tzu-Kuo Huang, Ruby C Weng, Chih-Jen Lin, and Greg Ridgeway. Generalized bradley-terry models and multi-class probability estimates. *Journal of Machine Learning Research*, 7(1), 2006. 3

[16] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023. 2

[17] Shari Liu, Neon B Brooks, and Elizabeth S Spelke. Origins of the concepts cause, cost, and goal in prereaching infants. *Proceedings of the National Academy of Sciences*, 116(36):17747–17752, 2019. 1

[18] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022. 2

[19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[20] Tianmin Shu, Yujia Peng, Lifeng Fan, Hongjing Lu, and Song-Chun Zhu. Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations. *Topics in cognitive science*, 10(1):225–241, 2018. 2

[21] Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22, 2009. 1, 3

[22] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18 (136):1–46, 2017. 3

[23] 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. 2

[24] 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. 2