CoRe Essay 6
Rethinking the Expressivity of Markov Reward: Reward is all you need

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

Abstract

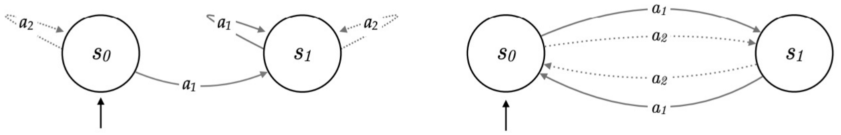
This paper engages in the ongoing debate on the nature and role of reward systems in human and artificial intelligence. Central to our discussion is the "Reward is enough" hypothesis, which posits that complex behaviors and cognitive functions in both humans and artificial intelligence can be reduced to and driven by reward maximization. We explore the theoretical underpinnings of this hypothesis, its practical implications, and the critiques and alternatives proposed in the literature. In particular, we focus on the expressivity of Markov Chain models and the potential for augmenting state spaces to encapsulate complex tasks that were previously thought to defy representation by traditional Markov reward functions. Our analysis provides insights into the sufficiency of reward maximization as a unifying principle for intelligence and the potential of expanded state spaces in addressing the limitations of current models.

1 Introduction

The concept of utility, a cornerstone in understanding human decision-making, presents an intricate challenge when translated into computational frameworks. The essence of human utility, characterized by its subjective and individualistic nature, defies straightforward quantification and generalization. In the quest to bridge this gap, the field of artificial intelligence and machine learning has pivoted towards exploring the dynamics of reward systems. This exploration is rooted in the hypothesis that intelligence, whether natural or artificial, fundamentally operates on the principle of maximizing [3]. This principle suggests that a comprehensive understanding of reward mechanisms could illuminate the pathways to replicating aspects of human intelligence in computational models.

The debate around this hypothesis is vibrant and diverse, with various research efforts contributing to an evolving narrative. On one end of the spectrum, some scholars advocate for a paradigm where reward maximization is deemed sufficient to drive behaviors encapsulating a wide array of cognitive abilities. This perspective, which aligns with the "Reward is enough" hypothesis, posits that the complexity and richness of intelligent behavior can be distilled into the pursuit of reward maximization. On the other end, alternative viewpoints highlight the limitations and potential oversimplifications of this approach, advocating for more nuanced or multifaceted models. For example, some work [2] introduces a paradigm shift in RL by proposing a method where tasks are specified not through laboriously crafted reward functions but through examples of successful outcomes. This approach aligns closely with how humans often learn – not through explicit quantification of utility but through observation and imitation of desired outcomes. This method's direct learning of value functions from outcomes without an intermediate reward function representation offers a simplified and potentially more robust framework for understanding human utility.

In this essay, we delve into this debate, primarily focusing on the "Reward is enough" hypothesis, while also acknowledging and examining the critiques and alternative models proposed in the



(a) Steady State Case

(b) Entailment Case

Figure 1: Two CMPs in which there is a SOAP that is not expressible under any Markov reward function. On the left, $\Pi_G = \{\pi_{21}\}$ is not realizable, as π_{21} cannot be made better than π_{22} because s_1 is never reached. On the right, the XOR-like-SOAP, $\Pi_G = \{\pi_{12}, \pi_{21}\}$ is not realizable: To make these two policies optimal, it is entailed that π_{22} and π_{11} must be optimal, too.

literature. By dissecting these perspectives, we aim to shed light on the potential of reward-based frameworks in capturing the essence of human utility and decision-making. The discussion will not only consider the theoretical underpinnings but also the practical implications, especially in terms of data collection, generalization, and computational efficiency. This exploration is pivotal for advancing our understanding of artificial intelligence and its capacity to mimic, complement, or enhance human cognitive processes.

2 Reward is enough

In the paper "Reward is Enough," the authors propose a unifying theory that reward maximization serves as the fundamental driving force behind various cognitive abilities in both human and artificial intelligence. This concept suggests that an array of complex behaviors and skills, ranging from basic perception to sophisticated social interactions and general intelligence, can be understood and developed within the framework of seeking and maximizing rewards.

The most astonishing part is, even the ability to understand and interact within social contexts is framed as a strategy for optimizing rewards in interactions with other agents. Also, the mastery of language is seen as a key tool for achieving rewards in complex social and cultural settings. Therefore, the overarching concept of general intelligence is presented as the ultimate form of adaptability and problem-solving geared towards reward maximization in a wide array of contexts.

3 Rethinking the Expressivity of Markov Chain

In his discourse, David [1] presents the argument that while many tasks within certain categories can be encapsulated by reward functions, there exist specific instances within each of these categories that defy representation by any Markov reward function. The central thesis of this paper is to counter this viewpoint by demonstrating that any instance can be encompassed by a Markov reward function, provided that the state space is appropriately augmented.

David's construction of a Controlled Markov Process (CMP), incorporating a Set Of Acceptable Policies (SOAP) that elude expression under a traditional Markov reward function, forms the cornerstone of his argument. This approach, however, can be critiqued for its excessive compression of the state space, leading to the loss of critical information. This scenario is analogous to expecting a visually impaired individual to safely operate a vehicle, thereby showcasing the inherent limitations in utilizing a Markov reward function under such constrained conditions.

To illustrate his point, David employs a rudimentary example involving a "consistent directional movement" task in a grid world, where the state is defined merely by an (x, y) coordinate pair. The SOAP $\Pi G = \{\pi_{\leftarrow}, \pi_{\uparrow}, \pi_{\rightarrow}, \pi_{\downarrow}\}$ encapsulates this task but fails to be adequately represented by any Markov reward function, in terms of elevating these policies above others in value. However, upon deeper examination of the underlying rationale for this "consistent directional movement" and the conditions enabling such a SOAP, one may deduce that the agent possesses an inherent "velocity" attribute, and any deviation from this velocity is unacceptable. Thus, by expanding the state space to encompass velocity, now defined as $(x, y) \times \{\leftarrow, \uparrow, \rightarrow, \downarrow\}$, the representation of these policies through a Markov reward function becomes feasible.

Additionally, David elucidates his perspective through two case studies, as referenced in Figure 1. In the 'Steady State Case', a critical analysis reveals that vital information pertaining to state s_1 is omitted. If state s_1 is indeed achievable, the exclusivity of policy π_{w1} as the only acceptable policy is unfounded. Thus, it is conceivable that state s_1 could be reached, possibly due to operational errors or other factors, suggesting a nonzero probability of occurrence. In the 'Entailment Case', it is apparent that by augmenting the space to $\{s_0, s_1\} \times \{a_1, a_2\}$, where a denotes the preceding action, the existence of a Markov reward function that renders the policy optimal is evident.

4 Conclusion

In this paper, we explore the "Reward is enough" hypothesis. We highlight the ongoing debate on the sufficiency of reward maximization as a principle for understanding and developing intelligence. Our analysis reveals that, while reward maximization provides a compelling framework for a wide range of cognitive abilities, there are inherent limitations and complexities in its application, particularly in scenarios with constrained or compressed state spaces. The discussion on the expressivity of Markov Chains, exemplified by David's critique and our counterarguments, underscores the importance of appropriately augmenting state spaces to capture the full breadth of tasks and policies.

References

[1] David Abel, Will Dabney, Anna Harutyunyan, Mark K Ho, Michael Littman, Doina Precup, and Satinder Singh. On the expressivity of markov reward. *Advances in Neural Information Processing Systems*, 34:7799–7812, 2021. 2

[2] Ben Eysenbach, Sergey Levine, and Russ R Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification. *Advances in Neural Information Processing Systems*, 34:11541–11552, 2021. 1

[3] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. 1