Reward is (almost) enough

Arjun Jagota
arjun.jagota22@imperial.ac.uk
Imperial College London

Sonali Parbhoo
sparbhoo@imperial.ac.uk
Imperial College London

Abstract

For centuries, humans have sought to understand intelligence and its associated mechanisms that drive how we think. While some have hypothesized that distinct signals or objectives are required for different types of abilities including learning, perception, social intelligence, generalization and imitation, others have suggested that learning through trial and error to maximise rewards can help develop behaviours that encompass all of these abilities. In this paper, we posit that while maximising rewards is central to developing a diverse range of abilities, the way in which we think about and formulate these rewards has to be re-framed as the conventional approach to using rewards in reinforcement learning can be prohibitive and is known to underperform in various settings, including sparse environments and noisy reward conditions. We suggest that these rewards need to be reformulated to incorporate different notions of i) uncertainty, ii) human preferences, and iii) nested or mixed compositions, iv) non-stationarity as well as account for v) situations where no reward is necessary. We suggest that doing so could enable more powerful reinforcement learning agents as a step towards artificial general intelligence.

1 Introduction

The richness and diversity of intelligence manifested in both animal and human behavior give rise to a multitude of associated abilities, ranging from social intelligence and language to perception, knowledge representation, planning, imagination, memory, and motor control. Exploring what motivates agents, whether natural or artificial, to display such varied forms of intelligence poses an intriguing question and is one that has fascinated humans for centuries.

Several researchers have posited that these abilities have emerged as a result of pursuing specific goals that are tailored to evoke them. For instance, language skills may have emerged from a combination of parsing, tagging various parts of speech, lexical analysis, and sentiment analysis goals. Similarly, perception may have developed from skills of recognition, segmentation and recall (Biederman, 1987). Others such as Silver et al. (2021) have instead suggested that a generic objective of maximising reward alone is enough to drive behaviour that exhibits most, if not all, abilities studied in both natural and artificial intelligence. A simple example of this might be how animals and humans require sophisticated abilities if they seek to survive in a complex world; consequently, achieving success would require a diverse range of intelligent capabilities and any behaviour that maximises the reward would thus exhibit those abilities.

Rewards in RL serve two fundamental purposes: first, different reward signals in different environments might produce different forms of intelligence (Silver et al., 2021), and second, the intelligence of an individual human or animal consists of a multitude of different skills or abilities. Together these abilities guide a human towards survival and should thus be consistent with any behaviours the human adopts.

In our work we posit that while the idea of rewards and reward maximisation might be sufficient to acquire a diverse range of intelligent abilities, this requires an adequate formulation of rewards. We

present some ideas on how this might be achieved in the context of several different examples. Our ideas are built on the premise that rewards should be *modular* in nature and might be comprised of different facets that enable accounting for uncertainty, different time horizons, non-stationarity and incorporating human preferences into RL where necessary. This formulation might also allow for the possibility of "switching off" all modules of a reward such that some behaviours might be captured without the need of a reward at all, similar to the notion of exploration in RL (Amin et al., 2021) and others that discuss reward-free RL e.g. Wagenmaker et al. (2022); Jin et al. (2020). We, like Silver et al. (2021), believe it is not the maximisation of rewards that limits the application of RL in practise; however the formulation of these rewards remains a key bottleneck for widespread application of RL in practice.

2 The Reinforcement Learning Problem Rethought

The reinforcement learning paradigm is commonly described in terms of an agent interacting with an environment to learn a task (Sutton & Barto, 1998). This is formalised with a Markov Decision Process (MDP) where an agent is in a state s_t at any point in time t, performs an action a_t and observes a reward from the environment r_t which it uses to update its state for the next time $t+1$, s_t, and perform subsequent actions such that the cumulative reward $\sum_{\forall t} \gamma^t r_t$ is maximised accordingly. Note that $\gamma \in (0, 1]$ is a discount factor that accounts for whether actions that occur immediately are prioritised over actions that occur in the future.

Importantly in this formulation, while the agent spends time computing its next action, the environment continues to process these actions and respond accordingly in terms of the reward produced. Subsequently, the agent might alter his course of behaviour based on the reward observed or act accordingly. However, this framing overlooks the fact that the agent too must maintain a mental model of what how the reward was produced. Though we may not be able to change the way in which an environment produces rewards, we can change the way we use this information and retain a mental model of the reward itself to guide subsequent behaviour. This mental model should encompass a variety of different facets or modules that might make learning easier. These include but are not limited to, accounting for uncertainty in the environment e.g. An et al. (2021) and incorporating human preferences similar to the way in which Reinforcement Learning with Human Feedback (RLHF) has done in language modelling e.g. Kirk et al. (2023). We might also consider whether the reward observed might be relevant for longer term of shorter time horizons so that we can adjust discounting accordingly e.g. Rathnam et al. (2023) since temporal abstraction plays a major role in planning and reasoning. We discuss these in the context of learning, perception, generalisation and imitation.

3 Reward is (almost) enough for knowledge and learning

Certain environments necessitate innate knowledge, crucial for immediate responses in novel situations. However, the capacity for innate knowledge is constrained by the agent's limitations and the challenge of constructing useful prior knowledge. Unlike other abilities, such as perception or action, innate knowledge cannot be operationalized and must precede experience.

Alternatively, environments may demand learned knowledge when future experiences are uncertain, leading to a vast array of potential knowledge requirements. In rich and complex environments, the space of potential knowledge surpasses the agent's capacity. Therefore, knowledge acquisition becomes a function of the agent's experience, necessitating learning processes.

Environments may require a combination of both innate and learned knowledge, with the balance shifting towards learned knowledge in richer and more extended environments. Reward maximising agents can integrate this innate and learned knowledge provided is adequately represented and adapts according to non-stationarity of changing environments. This may require rethinking how rewards are traditionally framed in terms of only state and action based information and adapt

these to account for other types of information including, but not limited to latent parameters and potential confounders.

4 Reward is (almost) enough for perception

In the human world, a range of perceptual abilities is crucial for obtaining rewards, with examples including image segmentation for safety, object recognition for food classification, face recognition for social interactions, scene parsing while driving, and speech recognition for immediate response (Silver et al., 2021). These abilities, spanning visual, auditory, olfactory, somatosensory, and proprioceptive senses, were traditionally approached as separate problems. However, recent advances in machine learning have aimed to unify them under supervised learning frameworks, effectively minimizing classification errors in test sets using labeled training data. An alternative view is that perception may be better understood as reward maximisation (Silver et al., 2021). This perspective aligns with observations in animals and could lead to more versatile perceptual behaviors. Active perception, context-dependent utility, information acquisition costs, contextual data distribution, and the absence of labeled data in many scenarios underscore the complexity and adaptive nature of perception. As a result, the way in which rewards are formulated for learning needs to be carefully considered. Given that as humans have a range of ways in which to perceive information based on our advanced sensory systems, it is unsurprising that for any reward model to be useful to guide behaviours, it too should capture a range of diverse information such as that used in perception. Some of this information might be more accessible than others, while in the absence of such information, we might need to draw on techniques such as reward shaping to assist in guiding behaviours in sparse reward situations. It should be noted however, that different situations may require varied degrees of shaping and adaptive forms of shaping should be considered, rather than traditional works for reward shaping that focus solely on defining reward shaping functions based on state and action information alone.

5 Reward is (almost) enough for generalisation

Generalization is commonly described as the capacity to apply a solution learned in one scenario to solve another (Pan & Yang, 2009; Taylor & Stone, 2009). For instance, in supervised learning, generalization may involve applying a solution acquired from one dataset, like photographs, to another dataset, such as paintings. In meta-learning, recent efforts have concentrated on transferring an agent from one environment to another.

Generalization may also be viewed in terms of maximising the reward based on a series of interactions with an agent and an environment. Based on the fact that the agent might encounter different aspects of the environment at different times, the agent will have to adapt accordingly and their behaviours will have to generalise in these situations. For this to work however, the agent must be able to adequately capture thee diverse aspects of the environment and retain this information in their mental model of the reward. Some ways of doing this might be to formalise rewards in terms of hierarchies where different elements of the hierarchy might be achieved as the agent acquires different skills to reach an overall goal (Shi et al., 2022). An alternative might be to consider a framing similar to options (Stolle & Precup, 2002) for rewards where, upon accounting some new information, the agent determines whether this information should be combined with a variety of elements that overlap and recur at different time-scales, or whether it should be ignored.

6 The Biological Perspective

The brain's complexity is evident in its myriad systems involved in cognition and behavior. Understanding these systems is crucial for building on learning processes and reward mechanisms.

Learning fundamentally relies on predicting the occurrence of rewards or penalties. Systems evolve to maximize reward probability. (Schultz, 2006) demonstrates that dopaminergic neurons reflect an

error signifying the difference between predicted and received rewards. These predictions manifest in various forms. The simplest is a passive prediction: an association between a stimulus and reward based on prior experience. Conversely, an active prediction is constructed upon learning associations between stimuli, responses, and rewards, encompassing the actions necessary to obtain the reward (O'Doherty, 2011).

6.1 Reward Prediction Error

Reward prediction error (RPE) manifests differently in the brain for various reward types. Stimulus-bound reward prediction error activates in the central and lateral orbitofrontal cortex, amygdala, and ventral striatum ((Shidara et al., 1998), (Gottfried et al., 2002), (Cromwell & Schultz, 2003), (Day et al., 2006)). Instrumental-action outcome associations are present in the vmPFC, medial OFC, medial prefrontal cortex, and anterior medial striatum ((Carmichael & Price, 1995), (Daw et al., 2006), (Hampton et al., 2006), (Kim et al., 2006), (Valentin et al., 2007), (Tanaka et al., 2008), (Gläscher et al., 2009)).

Instrumental-action outcome associations can be further categorized into goal-directed and habit-directed reward predictions. Goal-directed learning is observed in the vmPFC and medial OFC, while habit-directed reward predictions appear more prominent in the dorsolateral striatum, though this latter area warrants further investigation (Tanaka et al., 2008).

While our understanding of brain components related to RPEs remains nascent, this framework offers potential inspiration for artificial agent design. Agents could compute RPEs in diverse manners corresponding to different reward types within a task, potentially offering greater insight into agent cognition and learning processes. It is worth noting that this categorization of agent RPEs need not be limited to stimulus-bound and action-outcome associations; rather, it serves as an initial framework. Furthermore, this concept could extend to task hierarchies necessitating similar or distinct types of RPEs. We can extend this sentiment to hierarchical rewards.

6.2 Hierarchical Rewards

A critical aspect of human reward and penalty perception lies in the hierarchies comprising targeted actions or goals. Tasks can be made of different hierarchies of subgoals that help one achieve a larger goal, or subroutines that should be avoided to ensure completion.

The hierarchical nature of the human experience is mirrored in neural reward processing. (Ribas-Fernandes et al., 2011) conducted an experiment involving a delivery task game with a subgoal of envelope retrieval. Their study revealed RPE signals in the medial prefrontal cortex for both subgoals and superordinate goals. (Diuk et al., 2013) demonstrated that RPEs arise at different task levels in the basal ganglia, with the ventral striatum (VS) responding to both subgoal and goal-related prediction errors. (Ribas Fernandes et al., 2018) later expanded on their initial work, reaffirming that mPFC activity correlated with prediction errors specific to changes in goal and subgoal distances. Their experiments suggested that mPFC's subgoal or goal-related PE is task-specific and appears to be modulated by attentional factors. The study contrasted PE signaling between mPFC and VS, with VS exhibiting sensitivity only to money-driven PEs when a monetary incentive for efficient game completion was introduced. This finding suggests that mPFC is more engaged in tasks involving extended sequences and intentional behavior, aligning with earlier discussions on stimulus-bound predictions versus instrumental action-outcome associations.

Similarly to the types of learning, different hierarchical forms necessitate distinct types of PE signaling, and these hierarchies manifest in various learning contexts. Research efforts have aimed to isolate the roles of different brain regions with respect to these PEs and the hierarchies they influence. Future research should explore tasks of greater complexity to understand how different brain regions interact and the types of PE signals that emerge in such contexts.Adapting to computational systems would allow agents to further structure their RPEs in different types of hierarchies.

6.3 Putting things together

While these two concepts act as frameworks that help dictate the form of learning, there are characteristics within the system that can be integrated as well. Distributional rewards could assist with learning to give better context to agents. Biological systems have been seen to produce multiple RPEs, varying in magnitude or optimism (Dabney et al., 2020). Humans from a young age can learn causal structures. (McCormack et al., 2016) showed children's ability to perform interventions in simple models to demonstrate learning causal structures behaviourally. Further studies have aimed at mapping causality with neuroimaging and other electrophysiology techniques (Siddiqi et al., 2022), though much is yet to be understood. Causality extends far past just rewards, encompassing a great deal of state and action information as well as types of memory. There are methods to integrate these structures into these reward environments as well. One can consider Bayesian approaches to introducing priors an agent may have within the framework of various types of RPEs at different hierarchies. While models and neuroscience also maintain clear differences for good reason, inspiration can also flow in both directions. Much of modern reinforcement learning has been adapted from neuroscience and psychology, though further experimentation on models could help inspire theories and explanations as to how the brain functions so well. Given the immense complexity of systems like these, modularity is incredibly important. The ability to easily reshape as well as interpret these neurologically inspired structures is key to understanding how and why both models and brains function as they do.

7 Final Thoughts

Our exploration into the nature of intelligence and its underlying mechanisms underscores the ongoing quest to comprehend the complexities of human cognition. While historical perspectives have debated the necessity of distinct signals or objectives for various cognitive abilities, we propose an alternative viewpoint: that the pursuit of reward maximization can serve as a unifying framework for developing a spectrum of intelligent behaviors. However, our analysis suggests that the traditional approach to formulating rewards in reinforcement learning may hinder progress, particularly in challenging environments. We advocate for a paradigm shift in reward formulation, emphasizing considerations such as uncertainty, human preferences, dynamic environments, and scenarios where rewards may not be applicable. By reimagining rewards in this manner, we envision the potential for more robust reinforcement learning agents, representing a significant stride towards the realization of artificial general intelligence.

References

Susan Amin, Maziar Gomrokchi, Harsh Satija, Herke van Hoof, and Doina Precup. A survey of exploration methods in reinforcement learning. *arXiv preprint arXiv:2109.00157*, 2021.

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34:7436–7447, 2021.

Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

ST Carmichael and Joseph L Price. Sensory and premotor connections of the orbital and medial prefrontal cortex of macaque monkeys. *Journal of Comparative Neurology*, 363(4):642–664, 1995.

Howard C Cromwell and Wolfram Schultz. Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *Journal of neurophysiology*, 89(5):2823–2838, 2003.

Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675, 2020.

Nathaniel D Daw, John P O'doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, 2006.

Jeremy J Day, Robert A Wheeler, Mitchell F Roitman, and Regina M Carelli. Nucleus accumbens neurons encode pavlovian approach behaviors: evidence from an autoshaping paradigm. *European Journal of Neuroscience*, 23(5):1341–1351, 2006.

Carlos Diuk, Karin Tsai, Jonathan Wallis, Matthew Botvinick, and Yael Niv. Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *Journal of Neuroscience*, 33(13):5797–5805, 2013.

Jan Gläscher, Alan N Hampton, and John P O'Doherty. Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral cortex*, 19(2):483–495, 2009.

Jay A Gottfried, John O'Doherty, and Raymond J Dolan. Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. *Journal of Neuroscience*, 22(24):10829–10837, 2002.

Alan N Hampton, Peter Bossaerts, and John P O'doherty. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, 26(32):8360–8367, 2006.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020.

Hackjin Kim, Shinsuke Shimojo, and John P O'Doherty. Is avoiding an aversive outcome rewarding? neural substrates of avoidance learning in the human brain. *PLoS biology*, 4(8):e233, 2006.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

Teresa McCormack, Neil Bramley, Caren Frosch, Fiona Patrick, and David Lagnado. Children's use of interventions to learn causal structure. *Journal of experimental child psychology*, 141:1–22, 2016.

John P O'Doherty. 14 reward predictions and computations. *Neurobiology of Sensation and Reward*, pp. 311, 2011.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Sarah Rathnam, Sonali Parbhoo, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. The unintended consequences of discount regularization: Improving regularization in certainty equivalence reinforcement learning. In *International Conference on Machine Learning*, pp. 28746–28767. PMLR, 2023.

Jose JF Ribas-Fernandes, Alec Solway, Carlos Diuk, Joseph T McGuire, Andrew G Barto, Yael Niv, and Matthew M Botvinick. A neural signature of hierarchical reinforcement learning. *Neuron*, 71 (2):370–379, 2011.

José JF Ribas Fernandes, Danesh Shahnazian, Clay B Holroyd, and Matthew M Botvinick. Subgoal- and goal-related prediction errors in medial prefrontal cortex. *bioRxiv*, pp. 245829, 2018.

Wolfram Schultz. Behavioral theories and the neurophysiology of reward. *Annu. Rev. Psychol.*, 57 (1):87–115, 2006.

Lucy Xiaoyang Shi, Joseph J Lim, and Youngwoon Lee. Skill-based model-based reinforcement learning. *arXiv preprint arXiv:2207.07560*, 2022.

Munetaka Shidara, Thomas G Aigner, and Barry J Richmond. Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. *Journal of neuroscience*, 18(7):2613–2625, 1998.

Shan H Siddiqi, Konrad P Kording, Josef Parvizi, and Michael D Fox. Causal mapping of human brain function. *Nature reviews neuroscience*, 23(6):361–375, 2022.

David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103535. URL https://www.sciencedirect.com/science/article/pii/S0004370221000862.

Martin Stolle and Doina Precup. Learning options in reinforcement learning. In *Abstraction, Reformulation, and Approximation: 5th International Symposium, SARA 2002 Kananaskis, Alberta, Canada August 2–4, 2002 Proceedings 5*, pp. 212–223. Springer, 2002.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

Saori C Tanaka, Bernard W Balleine, and John P O'Doherty. Calculating consequences: brain systems that encode the causal effects of actions. *Journal of Neuroscience*, 28(26):6750–6755, 2008.

Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Vivian V Valentin, Anthony Dickinson, and John P O'Doherty. Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, 27(15):4019–4026, 2007.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pp. 22430–22456. PMLR, 2022.