Three Dogmas of Reinforcement Learning

David Abel
dmabel@google.com
Google DeepMind

Mark K. Ho
mkh260@nyu.edu
New York University

Anna Harutyunyan
harutyunyan@google.com
Google DeepMind

Abstract

Modern reinforcement learning has been conditioned by at least three dogmas. The first is the *environment spotlight*, which refers to our tendency to focus on modeling environments rather than agents. The second is our treatment of *learning as finding the solution to a task*, rather than adaptation. The third is the *reward hypothesis*, which states that all goals and purposes can be well thought of as maximization of a reward signal. These three dogmas shape much of what we think of as the science of reinforcement learning. While each of the dogmas have played an important role in developing the field, it is time we bring them to the surface and reflect on whether they belong as basic ingredients of our scientific paradigm. In order to realize the potential of reinforcement learning as a canonical frame for researching intelligent agents, we suggest that it is time we shed dogmas one and two entirely, and embrace a nuanced approach to the third.

1 On a Paradigm for Intelligent Agents

In *The Structure of Scientific Revolution*, Thomas Kuhn distinguishes between two phases of scientific activity (Kuhn, 1962). The first Kuhn calls "normal science" which he likens to puzzle-solving, and the second he calls the "revolutionary" phase, which consists of a re-imagining of the basic values, methods, and commitments of the science that Kuhn collectively calls a "paradigm".

The history of artificial intelligence (AI) arguably includes several swings between these two phases, and several paradigms. The first phase began with the 1956 Dartmouth workshop (McCarthy et al., 2006) and arguably continued up until sometime around the publication of the report by Lighthill et al. (1973) that is thought to have heavily contributed to the onset of the first AI winter (Haenlein & Kaplan, 2019). In the decades since, we have witnessed the rise of a variety of methods and research frames such as symbolic AI (Newell & Simon, 1961; 2007), knowledge-based systems (Buchanan et al., 1969) and statistical learning theory (Vapnik & Chervonenkis, 1971; Valiant, 1984; Cortes & Vapnik, 1995), culminating in the most recent emergence of deep learning (Krizhevsky et al., 2012; LeCun et al., 2015; Vaswani et al., 2017) and large language models (Brown et al., 2020; Bommasani et al., 2021; Achiam et al., 2023).

In the last few years, the proliferation of AI systems and applications has hopelessly outpaced our best scientific theories of learning and intelligence. Yet, it is our duty as scientists to provide the means to understand the current and future artifacts borne from the field, especially as these artifacts are set to transform society. It is our view that reflecting on the current paradigm and looking beyond it is a key requirement for unlocking this understanding.

In this position paper, we make two claims. First, reinforcement learning (RL) is a good candidate for a complete paradigm for the science of intelligent agents, precisely because "it explicitly considers the whole problem of a goal-directed agent interacting with an uncertain environment" (p. 3, Sutton & Barto, 2018). Second, in order for RL to play this role, we must reflect on the ingredients of our science and shift a few points of emphasis. These shifts are each subtle departures from three "dogmas", or implicit assumptions, summarized as follows:

1. The Environment-Spotlight (Section 2): Our emphasis on modeling environments rather than agents.
2. Learning as Finding a Solution (Section 3): Our search for agents that learn to solve tasks.
3. The Reward Hypothesis (Section 4): Assuming all goals are well thought of in terms of reward maximization.

When we relax these dogmas, we arrive at a view of RL as *the scientific study of agents*, a vision closely aligned with the stated goals of both RL and AI from their classic textbooks (Sutton & Barto, 2018; Russell & Norvig, 1995), as well as cybernetics (Wiener, 2019). As important special cases, these agents might interact with a Markov decision process (MDP; Bellman, 1957; Puterman, 2014), seek to identify solutions to specific problems, or learn in the presence of a reward signal with the goal of maximizing it, but these are not the only cases of interest.

2 Dogma One: The Environment Spotlight

The first dogma we call *the environment spotlight* (Figure 1), which refers to our collective focus on modeling environments and environment-centric concepts rather than agents. For example, the agent is essentially the means to deliver a solution to an MDP, rather than a grounded model in itself.

We do not fully reject this behaviourist view, but suggest balancing it; after all the classical RL diagram features two boxes, not just one. We believe that the science of AI is ultimately about *intelligent agents*, as argued by Russell & Norvig (1995); yet, much of our thinking, as well as our mathematical models, analysis, and central results tend to orbit around solving specific problems, and *not* around agents themselves. In other words, we lack a canonical formal model of an agent. This is the essence of the first dogma.

> Dogma 1: The Environment Spotlight
>
> Our collective focus on environments and environment-centric concepts, rather than agents.

What do we mean when we say that we focus on environments? We suggest that it is easy to answer only one of the following two questions:

1. *What is at least one canonical mathematical model of an environment in reinforcement learning?*
2. *What is at least one canonical mathematical model of an agent in reinforcement learning?*

The first question has a straightforward answer: the MDP, or any of its nearby variants such as a k-armed bandit (Lattimore & Szepesvári, 2020), a contextual bandit (Langford & Zhang, 2007), or a partially observable MDP (POMDP; Cassandra et al., 1994). These each codify different versions

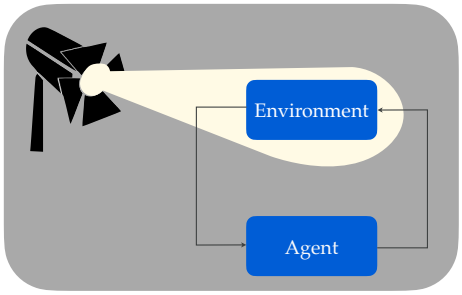

Figure 1: The first dogma, the Environment Spotlight.

of decision making problems, subject to different structural assumptions—in the case of an MDP, for instance, we make the Markov assumption by supposing there is a maintainable bundle of information we call the *state* that is a sufficient statistic of the next reward and next distribution over this same bundle of information. We assume these states are defined by the environment and are directly observable by the agent at each time step for use in learning and decision making. The POMDP relaxes this assumption and instead only reveals an observation to the agent, rather than the state. By embracing the MDP, we are allowed to import a variety of fundamental results and algorithms that define much of our primary research objectives and pathways. For example, we know every MDP has at least one deterministic, optimal, stationary policy, and that dynamic programming can be used to identify this policy (Bellman, 1957; Blackwell, 1962; Puterman, 2014). Moreover, our community has spent a great deal of effort in exploring variations of the MDP such as the Block MDP (Du et al., 2019) or Rich Observation MDP (Azizzadenesheli et al., 2016), the Object-Oriented MDP (Diuk et al., 2008), the Dec-POMDP (Oliehoek et al., 2016), Linear MDPs (Todorov, 2006), and Factored MDPs (Guestrin et al., 2003), to name a few. These models each forefront different kinds of problems or structural assumptions, and have inspired a great deal of illuminating research.

In contrast, this second question ("what is a canonical agent model?") has no clear answer (Harutyunyan, 2020). We might be tempted to respond in the form of a specific kind of a popular learning algorithm, such as Q-learning (Watkins & Dayan, 1992), but we suggest that this is a mistake. Q-learning is just one instance of the logic that *could* underlie an agent, but it is not a generic abstraction of what an agent actually is, not in the same way that a MDP is a model for a broad family of sequential decision making problems. As discussed by Harutyunyan (2020), we lack a canonical model of an agent, or even a basic conceptual picture. We believe that at this stage of the field, this is becoming a limitation, and is due in part to our focus on environments.

Indeed, the exclusive focus on environment-centric concepts (such as the dynamics model, environment state, optimal policy, and so on) can often obscure the vital role of the agent itself. But, here we wish to reignite interest in an agent-centric paradigm that can give us the conceptual clarity we need to be able to develop and discover general principles of agency. Without such ground currently, we struggle to even precisely define and differentiate between key agent families such as "model-based" and "model-free" agents (though some precise definitions have been given by Strehl et al. 2006 and Sun et al., 2019), or study more complex questions about the agent-environment boundary (Jiang, 2019; Harutyunyan, 2020), the extended-mind (Clark & Chalmers, 1998), embedded agency (Orseau & Ring, 2012), the effect of embodiment (Ziemke, 2013; Martin, 2022), or the impact of resource-constraints (Simon, 1955; Griffiths et al., 2015; Kumar et al., 2023; Aronowitz, 2023) on our agents in a general way. Most agent-centric concepts are typically beyond the scope of the basic mathematical language of our field, and are consequently not featured in our experimental work.

The Alternative: Shine the Spotlight on Agents, Too. Our suggestion is simple: it is important to define, model, and analyse agents in addition to environments. We should build toward a canonical mathematical model of an agent that can open us to the possibility of discovering general laws governing agents (if they exist), building on the work of Russell & Subramanian (1994), Wooldridge & Jennings (1995), Kenton et al. (2023), and echoing the call of Sutton (2022). We should engage in foundational work to establish axioms that characterize important agent properties and families, as in work by Sunehag & Hutter (2011; 2015) and Richens & Everitt (2024). We should do this in a way that is confluent with our latest empirical data about agents, drawing from the variety of disciplines that study agents, from psychology,[1] cognitive science, and philosophy, to biology, AI, and game theory. Doing so can expand the purview of our scientific efforts to understand and design intelligent agents.

[1]Tomasello makes a similar case that the field of psychology should center around the concept of agency: "Every scientific discipline begins with a proper domain, a first principle. In biology, that proper domain or first principle is life: physical substances organized in particular ways to perform particular organismic functions. In psychology, depending on one's theoretical predilections, that proper domain or first principle might be either behavior or mentality. But my preferred candidate would be agency, precisely because agency is the organizational framework within which both behavioral and mental processes operate." (p. 134, Tomasello, 2022).

3 Dogma Two: Learning as Finding a Solution

The second dogma is embedded in the way we treat the concept of learning. We tend to view learning as a finite process involving the search for—and eventual discovery of—a solution to a given task. For example, consider the classical problem of an RL agent learning to play a board game, such as Backgammon (Tesauro et al., 1995) or Go (Silver et al., 2016). In each of these cases, we tend to assume a good agent is one that will play a vast number of games to learn how to play the game effectively. Then, eventually, after enough games, the agent will reach optimal play and can stop learning as the desired knowledge has been acquired.

In other words, we tend to implicitly assume that the learning agents we design will eventually find a solution to the task at hand, at which point learning can cease. This is present in many of our classical benchmarks, too, such as mountain car (Taylor et al., 2008) or Atari (Bellemare et al., 2013), in which agents learn until they reach a goal. On one view, such agents can be understood as searching through a space of representable functions that captures the possible action-selection strategies available to an agent (Abel et al., 2023b), similar to the Problem Space Hypothesis (Newell, 1994). And, critically, this space contains at least one function—such as the optimal policy of an MDP—that is of sufficient quality to consider the task of interested solved. Often, we are then interested in designing learning agents that are guaranteed to *converge* to such an endpoint, at which point the agent can stop its search (and thus, stop its learning). This process is pictured in Figure 2, and is summarized in the second dogma.

> Dogma 2: Learning as Finding a Solution
>
> Our implicit focus on designing agents that find a solution, then stop learning.

This view is embedded into many of our objectives, and follows quite naturally from the use of the MDP as a model of the decision making problem. It is well established that every MDP has at least one optimal deterministic policy, and that such a policy can be learned or computed through dynamic programming or approximations thereof. The same tends to be true of many of the alternative learning settings we consider.

The Alternative: Learning as Adaptation. Our suggestion is to embrace the view that learning can also be treated as adaptation (Barron et al., 2015). As a consequence, our focus will drift away from optimality and toward a version of the RL problem in which agents continually improve, rather than focus on agents that are trying to solve a specific problem. Of course, versions of this problem have already been explored through the lens of lifelong (Brunskill & Li, 2014; Schaul et al., 2018), multi-task (Brunskill & Li, 2013), and continual RL (Ring, 1994; 1997; 2005; Khetarpal et al., 2022; Anand & Precup, 2023; Abel et al., 2023b; Kumar et al., 2023). Indeed, this perspective is highlighted in the introduction of the textbook by Sutton & Barto (2018):

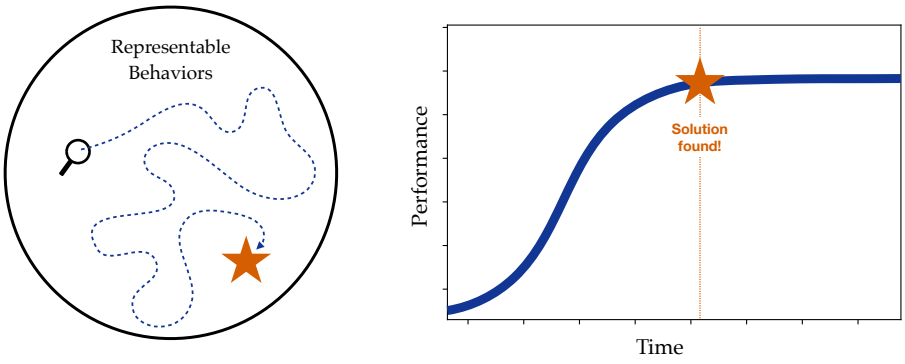

Figure 2: Dogma 2: Learning as Finding a Solution.

> When we say that a reinforcement learning agent's goal is to maximize a numerical reward signal, we of course are not insisting that the agent has to actually achieve the goal of maximum reward. Trying to maximize a quantity does not mean that that quantity is ever maximized. The point is that a reinforcement learning agent is always trying to increase the amount of reward it receives. (p. 10, Sutton & Barto, 2018).

This is a matter of a shift of emphasis: when we move away from optimality, how do we think about evaluation? How, precisely, can we define this form of learning, and differentiate it from others? What are the basic algorithmic building blocks that carry out this form of learning, and how are they different from the algorithms we use today? Do our standard analysis tools such as regret and sample complexity still apply? These questions are important, and require reorienting around this alternate view of learning. We suggest that we as a community shed the second dogma and study these questions directly.

4 Dogma Three: The Reward Hypothesis

The third dogma is the *reward hypothesis* (Sutton, 2004; Littman, 2015; Christian, 2021; Abel et al., 2021; Bowling et al., 2023), which states "All of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)."

First, it is important to acknowledge that this hypothesis is not deserving of the title "dogma" at all. As originally stated, the reward hypothesis was intended to organize our thinking around goals and purposes, much like the expected utility hypothesis before it (Machina, 1990). And, the reward hypothesis seeded the research program of RL in a way that has led to the development of many of our most celebrated results, applications, and algorithms.

> Dogma 3: The Reward Hypothesis
>
> All goals can be well thought of in terms of reward maximization.

However, as we continue our quest for the design of intelligent agents (Sutton, 2022), it is important to recognize the nuance in the hypothesis.

In particular, recent analysis by Bowling et al. (2023), building on the work of Pitis (2019); Abel et al. (2021) and Shakerinava & Ravanbakhsh (2022), fully characterizes the implicit conditions required for the hypothesis to be true. These conditions come in two forms. First, Bowling et al. provide a pair of interpretative assumptions that clarify what it would mean for the reward hypothesis

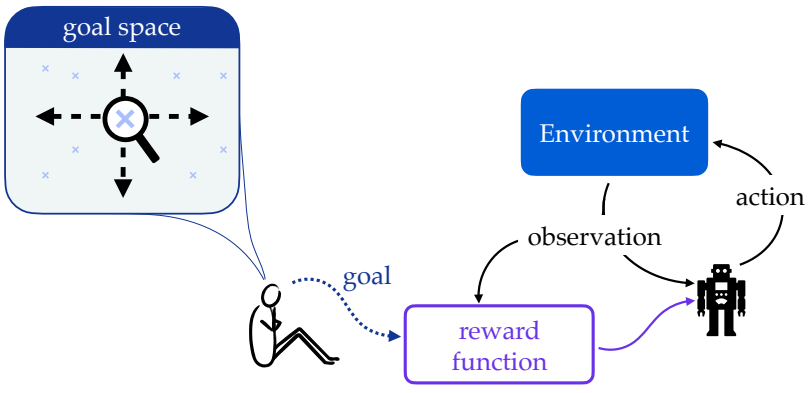

Figure 3: The third dogma, the Reward Hypothesis. Any goal that a designer might conceive of can be well thought of in terms of the maximization of a reward signal by a learning agent.

to be true or false—roughly, these amount to saying two things. First, that "goals and purposes" can be understood in terms of a preference relation on possible outcomes. Second, that a reward function captures these preferences if the ordering over agents induced by value functions matches that of the ordering induced by preference on agent outcomes. Then, under this interpretation, a Markov reward function exists to capture a preference relation if and only if the preference relation satisfies the four von Neumann-Morgenstern axioms (von Neumann & Morgenstern, 1953), and a fifth Bowling et al. call γ-Temporal Indifference.

This is significant, as it suggests that when we write down a Markov reward function to capture a desired goal or purpose, we are *forcing* our goal or purpose to adhere to the five axioms, and we must ask ourselves if it is always appropriate. As an example, consider the classical challenge on the incomparability (or incommeasurability) of values in ethics, as discussed by Chang (2015). That is, certain abstract virtues such as happiness and justice might be thought to be incomparable to one another. Or, similarly, two concrete experiences might be incommeasurable, such as a walk on the beach and eating breakfast—how might we assign measure to each of these experiences in the same "currency"? Chang notes that two items might not be comparable without further reference to a particular use, or context: "A stick can't be greater than a billiard ball...it must be greater in some respect, such as mass or length." However, the first axiom, completeness, strictly requires that the implicit preference relation assigns a genuine preference between all pairs of experiences. As such, if we take the reward hypothesis to be true, we can only encode goals or purposes in a reward function that reject both incomparability and incommeasurability. It is worth noting that completeness in particular has been criticized by Aumann (1962) due to the demands it places on the individual holding the preference relation. Finally, the completeness axiom is not the only one restricting the space of viable goals and purposes; axiom three, independence of irrelevant alternatives, famously rejects risk-sensitive objectives as well due to the Allais paradox (Allais, 1953; Machina, 1982).

The Alternative: Recognize and Embrace Nuance. Our suggestion is to simply call attention to the limitations of scalar rewards, and to be open to other languages for describing an agent's goals. It is important that we are aware of the implicit restrictions we are placing on the viable goals and purposes under consideration when we represent a goal or purpose through a reward signal. We should become familiar with the requirements imposed by the five axioms, and be aware of what specifically we might be giving up when we choose to write down a reward function. On this latter point there is a profound opportunity for future work. It is also worth highlighting the fact that preferences are themselves just another language for characterizing goals—there are likely to be others, and it is important to cast a wide net in our approach to thinking about goal-seeking.

5 Discussion

We have here argued that the long-term vision of RL should be to provide a holistic paradigm for the science of intelligent agents. To realise this vision, we suggest that it is time to reconcile our relationship with three implicit dogmas that have shaped aspects of RL so far. These three dogmas amount to over-emphasis (1) on environments, (2) on finding solutions, and (3) on rewards as a language for describing goals. Further, we have initial suggestions on how to pursue research that makes subtle departures from these dogmas. First, we should treat agents as one of our central objects of study. Second, we must move beyond studying agents that find solutions for specific tasks, and also study agents that learn to endlessly improve from experience. Third, we should recognize the limits of embracing reward as our language for goals, and consider alternatives.

Open Questions. Each of these suggestions can be translated into important research questions we encourage the community to explore further. First, *what is our canonical model of an agent?* Several recent proposals have emerged, and agree on many aspects. What are the consequences of adopting one view, rather than another? Which ingredients of an agent are necessary, rather than extraneous? We suggest that it is important to think carefully about these questions, and adopt conventions for the standard model of an agent. Such a model can be used to clarify old questions, and open new

lines of study around agent-centric concepts such as the agent-environment boundary (Todd & Gigerenzer, 2007; Orseau & Ring, 2012; Harutyunyan, 2020), embodiment (Ziemke, 2013; Martin, 2022), resource-constraints (Simon, 1955; Ortega, 2011; Braun & Ortega, 2014; Ortega et al., 2015; Griffiths et al., 2015; Kumar et al., 2023; Aronowitz, 2023), and embedded agency (Orseau & Ring, 2012). Second, what is the goal of learning when we give up the concept of a task's solution? In other words: how do we think about learning when no optimal solution can be found? How do we begin to evaluate such agents, and measure their learning progress? Third, we suggest embracing a wide variety of views about plausible accounts of the objectives of an agent. This includes continuing to embrace classical accounts of reward maximization, but also considering varied objectives like average reward (Mahadevan, 1996), risk (Howard & Matheson, 1972; Mihatsch & Neuneier, 2002), constraints (Altman, 2021), logical goals (Littman et al., 2017), or even open-ended goals (Samvelyan et al., 2023).

On the term "Dogma". The title of this paper and use of the term "dogma" are an homage to "Two Dogmas of Empiricism" by Quine (1951). The term "dogma" casts a more negative light on each of the principles than we intend (though, as Kuhn (1963) notes, there is a role for dogma in the sciences). Indeed, as discussed, the reward hypothesis was originally conceived of as a *hypothesis* as its name suggests. Still, it is a principle that is often taken as a presupposition that frames the rest of the field of RL similar to the way that the Church-Turing Thesis frames computation—they are both standard pre-scientific commitments that are part of most research programmes (Lakatos, 2014). The other two dogmas are both *implicit* rather than conventions we regularly state openly and embrace; it is rare to see work in RL actively argue against the importance of thinking about agents or agency, for instance. Instead, it is a convention to begin most RL research by framing our research questions around dynamic programming and MDPs. In this sense, the community has been drawn to specific well-tread research paths that involve modeling environments first, rather than *agents* directly. The same implicit character is true of the second dogma: due to our focus on MDPs and related models, it also tends to be the case that instances of the RL problem we study have a well structured *solution* that is known to be discoverable through means such as dynamic programming or temporal difference learning. We then often use language involving an algorithm *solving* a task by converging to an optimal policy, reflecting the influence of the second dogma. It is in this sense that we take the term "dogma" to be fitting of the first two: we tend not to question these aspects of our research programme, yet they influence much of our methods and goals.

It is worth noting that it is understandable why the sentiments underlying the three dogmas were adopted: by building our study from Markov models, we can make use of the suite of well-understood, efficient algorithms based on dynamic programming, thanks to the seminal work by Bellman (1957), Sutton (1988), Watkins (1989), and others. This is further supported by the way that fundamental results from stochastic approximation (Robbins & Monro, 1951) have influenced many classical results, such as the convergence of Q-learning by Watkins & Dayan (1992) or TD-learning with function approximation by Tsitsiklis & Van Roy (1996).

Inspiration. We are not the first to suggest moving beyond some of these conventions. The work on *general* reinforcement learning by Hutter (2000; 2002; 2004) and colleagues (Lattimore & Hutter, 2011; Leike, 2016; Cohen et al., 2019) has long studied RL in the most general possible setting. Indeed, the stated goal of the original work on AIXI by Hutter (2000) was "...to introduce the universal AI model" (p. 3). Similarly, a variety of work has explicitly focused on agents. For instance, the classical AI textbook by Russell & Norvig (1995) defines AI "as the study of agents that receive percepts from the environment and perform actions" (p. viii), and frames the book around "the concept of the intelligent agent" (p. vii). Russell & Subramanian (1994) also feature a general take on goal-directed agents that has shaped much of the agent-centric literature that follows—the agent functions there introduced have been more recently adopted as one model of an agent (Abel et al., 2023a;b). Sutton (2022) proposes the "quest for a common model of the intelligent decision maker", and provides initial suggestions for how to frame this quest. Work by Dong et al. (2022) and Lu et al. (2021) have built on the traditions of agent-centric modeling, providing detailed accounts of the possible

constituents of an agent's internal mechanism, similar to Sutton. Further work by Kenton et al. (2023) and Richens & Everitt (2024) explore a causal perspective on agents, giving both concrete definitions and insightful results. Outside of AI, the subject of *agency* is an important subject of discourse in its own right—we refer the reader to the work by Barandiaran et al. (2009) and Dretske (1999) or the books by Tomasello (2022) and Dennett (1989) for further insights from nearby communities.

Similarly, a variety of work has explored alternative ways to think about goals. For instance, Little & Sommer (2013) study an agent that learns a predictive model of its environment, and ground this study using the tools of information theory. This is similar in spirit to the Free-Energy Principle advocated for by Friston (2010), with recent work by Hafner et al. (2020) exploring connections to RL. Preferences have also been used as an alternative to rewards, as in preference-based RL (Wirth et al., 2017), with a more recent line of work on RL from human feedback (Christiano et al., 2017; MacGlashan et al., 2016; 2017) now playing a significant role in the current wave of language model research (Achiam et al., 2023). Others have proposed the use of various logical languages for grounding goals, such as linear temporal logic (Littman et al., 2017; Li et al., 2017; Hammond et al., 2021) and nearby structures such as reward machines (Icarte et al., 2022). Another perspective presented by Shah et al. (2021) explicitly contrasts the framing of assistance games (Hadfield-Menell et al., 2016) with reward maximization, and suggests that the former provides a more compelling path to designing assistive agents. Lastly, a variety of work has considered forms of goal-seeking beyond expected cumulative reward, as in ordinal dynamic programming (Koopmans, 1960; Sobel, 1975), convex RL (Zahavy et al., 2021; Mutti et al., 2022; 2023), other departures from the expectation (Bellemare et al., 2017; 2023), or by incorporating other objectives such as constraints (Le et al., 2019; Altman, 2021) or risk (Mihatsch & Neuneier, 2002; Shen et al., 2014; Wang et al., 2023).

Other Dogmas. There are many other assumptions inherent to the basic philosophy of reinforcement learning that we did not discuss. For instance, it has been common to focus on agents that learn from a *tabula rasa* state, rather than consider other stages of learning. We also tend to adopt the cumulative discounted reward with a geometric discounting schedule as the objective, rather than using a hyperbolic schedule (Fedus et al., 2019), or consider the existence of environment-state rather than a partially observable setting (Cassandra et al., 1994; Dong et al., 2022). We take it that reflecting on these and other perspectives is also important, but that they have already received significant attention by the community.

Conclusion. We hope this paper can reinvigorate the RL community to explore beyond our current frames. We believe this begins by embracing the vision that RL is a good candidate for a holistic paradigm of intelligent agents, and continues with a careful reflection of the values, methods, and ingredients of our scientific practice that will enable this paradigm to flourish.

Acknowledgements

The authors are grateful André Barreto and Dilip Arumugam for detailed comments on a draft of the paper, and to the anonymous RLC reviewers for their helpful feedback. We would further like to thank the many people involved in discussions that helped shape the authors' thoughts on each of the three dogmas: Sara Aronowitz, André Barreto, Mike Bowling, Brian Christian, Will Dabney, Michael Dennis, Steven Hansen, Khimya Khetarpal, Michael Littman, John Martin, Doina Precup, Mark Ring, Mark Rowland, Tom Schaul, Satinder Singh, Rich Sutton, Hado van Hasselt, Ben Van Roy, and all of the members of the Agency team.

References

David Abel, Will Dabney, Anna Harutyunyan, Mark K. Ho, Michael L Littman, Doina Precup, and Satinder Singh. On the expressivity of Markov reward. In *Advances in Neural Information Processing Systems*, 2021.

David Abel, André Barreto, Hado van Hasselt, Benjamin Van Roy, Doina Precup, and Satinder Singh. On the convergence of bounded agents. *arXiv preprint arXiv:2307.11044*, 2023a.

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023b.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Maurice Allais. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: journal of the Econometric Society*, pp. 503–546, 1953.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Nishanth Anand and Doina Precup. Prediction and control in continual reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Sara Aronowitz. The parts of an imperfect agent. *Oxford Studies in Philosophy of Mind Volume 3*, pp. 1, 2023.

Robert J Aumann. Utility theory without the completeness axiom. *Econometrica: Journal of the Econometric Society*, pp. 445–462, 1962.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning in rich-observation MDPs using spectral methods. *arXiv preprint arXiv:1611.03907*, 2016.

Xabier E Barandiaran, Ezequiel Di Paolo, and Marieke Rohde. Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5):367–386, 2009.

Andrew B Barron, Eileen A Hebets, Thomas A Cleland, Courtney L Fitzpatrick, Mark E Hauber, and Jeffrey R Stevens. Embracing multiple definitions of learning. *Trends in neurosciences*, 38(7): 405–407, 2015.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, pp. 679–684, 1957.

David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, pp. 719–726, 1962.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *Proceedings of the International Conference on Machine Learning*, 2023.

Daniel A. Braun and Pedro A. Ortega. Information-theoretic bounded rationality and ϵ-optimality. *Entropy*, 16(8):4662–4676, 2014.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. *arXiv preprint arXiv:1309.6821*, 2013.

Emma Brunskill and Lihong Li. PAC-inspired option discovery in lifelong reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2014.

Bruce Buchanan, Georgia Sutherland, and Edward A Feigenbaum. Heuristic DENDRAL: A program for generating explanatory hypotheses. *Organic Chemistry*, pp. 30, 1969.

Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. Acting optimally in partially observable stochastic domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1994.

Ruth Chang. Value incomparability and incommensurability. *The Oxford handbook of value theory*, pp. 205–224, 2015.

Brian Christian. *The Alignment Problem: Machine Learning and Human Values*, pp. 130–131. Atlantic Books, 2021.

Paul F Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Andy Clark and David Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998.

Michael K Cohen, Elliot Catt, and Marcus Hutter. A strongly asymptotically optimal agent in general environments. *arXiv preprint arXiv:1903.01021*, 2019.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Daniel C Dennett. *The intentional stance*. MIT press, 1989.

Carlos Diuk, Andre Cohen, and Michael L Littman. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the International conference on Machine learning*, 2008.

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent states. *Journal of Machine Learning Research*, 23(255):1–54, 2022.

Fred I Dretske. Machines, plants and animals: the origins of agency. *Erkenntnis (1975-)*, 51(1):19–31, 1999.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *Proceedings of the International Conference on Machine Learning*, 2019.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

Karl J Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2): 127–138, 2010.

Thomas L Griffiths, Falk Lieder, and Noah D Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229, 2015.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, 2016.

Michael Haenlein and Andreas Kaplan. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, 61(4):5–14, 2019.

Danijar Hafner, Pedro A Ortega, Jimmy Ba, Thomas Parr, Karl J Friston, and Nicolas Heess. Action and perception as divergence minimization. *arXiv preprint arXiv:2009.01791*, 2020.

Lewis Hammond, Alessandro Abate, Julian Gutierrez, and Michael Wooldridge. Multi-agent reinforcement learning with temporal logic specifications. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2021.

Anna Harutyunyan. What is an agent? http://anna.harutyunyan.net/wp-content/uploads/2020/09/What_is_an_agent.pdf, 2020.

Ronald A Howard and James E Matheson. Risk-sensitive Markov decision processes. *Management science*, 18(7):356–369, 1972.

Marcus Hutter. A theory of universal artificial intelligence based on algorithmic complexity. *arXiv preprint cs/0004001*, 2000.

Marcus Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Proceedings of the International Conference on Computational Learning Theory*, 2002.

Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

Nan Jiang. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents. *Artificial Intelligence*, pp. 103963, 2023.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

Tjalling C Koopmans. Stationary ordinal utility and impatience. *Econometrica: Journal of the Econometric Society*, pp. 287–309, 1960.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1962.

Thomas S Kuhn. The function of dogma in scientific research. *Scientific Change*, 1963.

Saurabh Kumar, Henrik Marklund, Ashish Rao, Yifan Zhu, Hong Jun Jeon, Yueyang Liu, and Benjamin Van Roy. Continual learning as computationally constrained reinforcement learning. *arXiv preprint arXiv:2307.04345*, 2023.

Imre Lakatos. Falsification and the methodology of scientific research programmes. In *Philosophy, Science, and History*, pp. 89–94. Routledge, 2014.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems*, 2007.

Tor Lattimore and Marcus Hutter. Asymptotically optimal agents. In *Proceedings of the International Conference on Algorithmic Learning Theory*, 2011.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *Proceedings of the International Conference on Machine Learning*, 2019.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Jan Leike. *Nonparametric general reinforcement learning*. PhD thesis, The Australian National University, 2016.

Xiao Li, Cristian-Ioan Vasile, and Calin Belta. Reinforcement learning with temporal logic rewards. In *Proceedings of the International Conference on Intelligent Robots and Systems*, 2017.

James Lighthill, Stuart Sutherland, Roger Needham, and Christopher Longuet-Higgins. Artificial intelligence: A general survey. *Science Research Council*, 1973.

Daniel Y Little and Friedrich T Sommer. Learning and exploration in action-perception loops. *Frontiers in neural circuits*, 7:37, 2013.

Michael L Littman. Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 521(7553):445–451, 2015.

Michael L Littman, Ufuk Topcu, Jie Fu, Charles Isbell, Min Wen, and James MacGlashan. Environment-independent task specifications via GLTL. *arXiv preprint arXiv:1704.04341*, 2017.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. Reinforcement learning, bit by bit. *arXiv preprint arXiv:2103.04047*, 2021.

James MacGlashan, Michael L Littman, David L Roberts, Robert Loftin, Bei Peng, and Matthew E Taylor. Convergent actor critic by humans. In *Proceedings of the International Conference on Intelligent Robots and Systems*, 2016.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive learning from policy-dependent human feedback. In *Proceedings of the International Conference on Machine Learning*, 2017.

Mark J Machina. "Expected Utility" analysis without the independence axiom. *Econometrica: Journal of the Econometric Society*, pp. 277–323, 1982.

Mark J Machina. Expected utility hypothesis. In *Utility and probability*, pp. 79–95. Springer, 1990.

Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1):159–195, 1996.

John D Martin. Time to take embodiment seriously. 2022.

John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the Dartmouth summer research project on Artificial Intelligence, August 31, 1955. *AI magazine*, 27(4): 12–12, 2006.

Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49: 267–290, 2002.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250):1–42, 2023.

Allen Newell. *Unified theories of cognition*. Harvard University Press, 1994.

Allen Newell and Herbert A Simon. Computer science as empirical inquiry: Symbols and search. In *ACM Turing award lectures*, pp. 1975. 2007.

Allen Newell and Herbert Alexander Simon. GPS, a program that simulates human thought. 1961.

Frans A Oliehoek, Christopher Amato, et al. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2016.

Laurent Orseau and Mark Ring. Space-time embedded intelligence. In *Proceedings of the International Conference on Artificial General Intelligence*, 2012.

Pedro A Ortega. *A unified framework for resource-bounded autonomous agents interacting with unknown environments*. PhD thesis, University of Cambridge, 2011.

Pedro A Ortega, Daniel A Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby. Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*, 2015.

Silviu Pitis. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Willard Van Orman Quine. Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43, 1951.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *Proceedings of the International Conference on Learning Representations*, 2024.

Mark B Ring. *Continual learning in reinforcement environments*. PhD thesis, The University of Texas at Austin, 1994.

Mark B Ring. Child: A first step towards continual learning. *Machine Learning*, 28(1):77–104, 1997.

Mark B Ring. Toward a formal framework for continual learning. In *NeurIPS Workshop on Inductive Transfer*, 2005.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Stuart J Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 1995. ISBN 0-13-103805-2.

Stuart J Russell and Devika Subramanian. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609, 1994.

Mikayel Samvelyan, Akbir Khan, Michael D Dennis, Minqi Jiang, Jack Parker-Holder, Jakob Nicolaus Foerster, Roberta Raileanu, and Tim Rocktäschel. MAESTRO: Open-ended environment design for multi-agent reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2023.

Tom Schaul, Hado van Hasselt, Joseph Modayil, Martha White, Adam White, Pierre-Luc Bacon, Jean Harb, Shibl Mourad, Marc Bellemare, and Doina Precup. The Barbados 2018 list of open issues in continual learning. *arXiv preprint arXiv:1811.07004*, 2018.

Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael D Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. Benefits of assistance over reward learning, 2021. URL https://openreview.net/forum?id=DFIoGDZejIB.

Mehran Shakerinava and Siamak Ravanbakhsh. Utility theory for sequential decision making. In *Proceedings of the International Conference on Machine Learning*, 2022.

Yun Shen, Michael J Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328, 2014.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1): 99–118, 1955.

Matthew J Sobel. Ordinal dynamic programming. *Management science*, 21(9):967–975, 1975.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. PAC model-free reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2006.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Proceedings of the Conference on Learning Theory*, 2019.

Peter Sunehag and Marcus Hutter. Axioms for rational reinforcement learning. In *Proceedings of the International Conference on Algorithmic Learning Theory*, 2011.

Peter Sunehag and Marcus Hutter. Rationality, optimism and guarantees in general reinforcement learning. *The Journal of Machine Learning Research*, 16(1):1345–1390, 2015.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3: 9–44, 1988.

Richard S Sutton. The reward hypothesis, 2004. URL http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html.

Richard S Sutton. The quest for a common model of the intelligent decision maker. *arXiv preprint arXiv:2202.13252*, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Matthew E Taylor, Gregory Kuhlmann, and Peter Stone. Autonomous transfer for reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2008.

Gerald Tesauro et al. Temporal difference learning and TD-gammon. *Communications of the ACM*, 38 (3):58–68, 1995.

Peter M Todd and Gerd Gigerenzer. Mechanisms of ecological rationality: heuristics and environments that make. *Oxford handbook of evolutionary psychology*, pp. 197, 2007.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Michael Tomasello. *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press, 2022.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*, 1996.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Vladimir N Vapnik and Aleksei Y Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1953.

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with CVaR. In *Proceedings of the International Conference on Machine Learning*, 2023.

Christopher J.C.H. Watkins. *Learning from delayed rewards*. PhD thesis, Cambridge University, 1989.

Christopher J.C.H. Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.

Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990, 2017.

Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex MDPs. In *Advances in Neural Information Processing Systems*, 2021.

Tom Ziemke. What's that thing called embodiment? In *Proceedings of the Annual Cognitive Science Society*. 2013.