RL in context - towards a framing that enables cybernetics-style questions

Vincent Létourneau
vincentmillions@gmail.com
Université de Montréal

Maia Fraser
mfrase8@uottawa.ca
University of Ottawa

Abstract

Early work in cybernetics and artificial intelligence (AI) was aimed as much at developing computational models that enable mathematical understanding of key principles in living organisms as it was at developing computing machines based on such models. Reinforcement learning of course traces its roots back to these early days as well, with significant development since then, both in terms of mathematical formalization and applied success. As we gather in this workshop to examine the conceptual foundations of RL and how these impact framing problems, we aim in this short paper especially in the direction of building computational models to capture and shed light on aspects of biological life. Given the generality of RL, we argue it is an especially fruitful paradigm to bring into a novel formulation of continual learning - one that may not immediately fit the bitter lesson's emphasis on scaling-above-all, but that we hope can be used to formulate cybernetics-style questions about the rich hierarchy of learning processes found in complex living systems. Strengthening this foundational thrust also holds promise to help make sense of safety issues in the ongoing deployment of AI.

1 Background

The fertile two-way exchange that originally existed in the field of AI between research aimed at understanding biological intelligence and research aimed at developing intelligent machines has shifted heavily over the years towards an emphasis on the latter, in part driven by dramatic engineering successes. Although machine learning theory has continued to develop, this has often been directed towards explaining and designing machines rather than elucidating biological learning. The theoretical underpinnings of RL—its concepts, definitions, axioms, and assumptions—have very much followed this trend.

In this workshop, that provides the opportunity to re-examine some of the underlying framework in RL and how it enables or impedes the framing of problems, our contribution looks specifically at how the framing of RL could be shaped to promote *learning-theoretic analysis* of questions concerning complex living systems. In particular, we propose to capitalize on the generality of RL as a learning paradigm (i.e. the possibility to view supervised learning, active learning, semi- and un-supervised learning as special cases of RL) and then to place such general learning units in the context of a hierarchical formulation of continual learning. We are thus exploring a version of continual RL that involves a possibly endless arising and vanishing of RL-units, most of which are themselves not immortal. The immortality emerges only potentially at the top of the hierarchy, and possibly there is no highest level. This proposed hierarchical formulation is inspired by a systems-theoretic viewpoint as expressed in early work of mathematician Norbert Wiener (Wiener, 1948; 1950) and we consider cybernetics-style questions that are natural to pose in this setting.

1.1 Cybernetics

The term *cybernetics* was coined by Wiener in his book of the same name (Wiener, 1948) in order to designate the science of communication and control in both living systems and machines. It derives from the Greek verb "kubernan", which means to steer and which is also, for example, at the root of the English word "govern". Wiener was interested in mathematically studying the phenomenon of steering. In particular, he argued that feedback mechanisms were central to all intelligent behaviour. In this and subsequent books of his aimed at general readers, Wiener considered a wide variety of questions in society, technology and biology, highlighting the role of feedback in each, as an organism or organization. There was significant overlap with systems theory, though he did not specifically use the terminology *complex (living) system*. The defining property of such a system is that it be composed of many components, together with a self-regulation mechanism to maintain a certain global pattern among these components, each of which may in turn be a complex system of its own. This sets up hierarchy as a key property of life. We return to this notion again in Section 2.

1.2 Reinforcement learning as a general paradigm

Reinforcement learning (RL) is most commonly formulated in terms of a Markov decision process (MDP), namely, a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, \mathcal{R} \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are spaces of states and actions respectively, P is the transition probability kernel, and $\mathcal{R}$ the reward function; see (Sutton & Barto, 1998; Puterman, 1994, 2005). The agent knows $\mathcal{S}$ and $\mathcal{A}$. It sees a current state—revealed by the environment—and, upon taking an action a from $\mathcal{A}$, sees the effect of P and $\mathcal{R}$ as they generate respectively the next state $s' \sim P_{a,s}$ and the associated reward $\mathcal{R}_a(s, s')$. The agent's goal is to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ (or to distributions on $\mathcal{A}$) that maximizes the expectation of some form of cumulative reward, which could be the average reward over an infinite horizon, discounted total reward with fixed discount factor $\gamma \in (0, 1)$, or un-discounted total over a finite horizon. Since $P, \mathcal{R}$ are not known[1] to the agent, this challenge is a learning problem, referred to as RL. We focus especially on the interaction this sets up between the agent and environment, defining an RL agent as in Figure 1a. Supervised learning can be seen as a special case where the only actions are requesting new samples and predicting labels, neither of which changes the state of the environment. An active-learning variant of this could offer more tailored sample-requesting actions, and so on. We may thus view the basic RL agent in Figure 1a as a general learning unit.

1.3 Continual learning

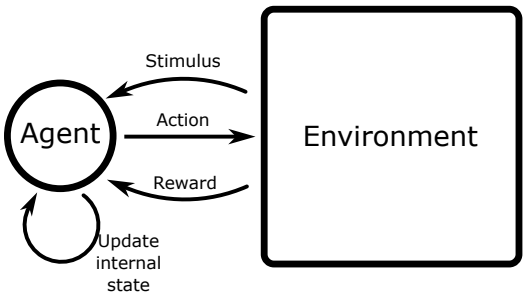
Continual learning (also called incremental or lifelong learning) has been formulated in various ways (see for example surveys in Wang et al. (2024); van de Ven & Tolias (2019) and in the RL setting Abel et al. (2023)). The basic idea is that training samples (or observations), generated by a distribution (environment) that may shift in time, will be observed in sequence and the learning agent should retain relevant information learned from the past to carry out a sequence of learning tasks. Concretely, for a supervised learning example, the t'th task might be to predict the mean of Y given X under the distribution $\mathcal{D}_t$ on $X \times Y$, where t is a possibly infinite-horizon time variable ($t \in \{0, ..., T\}$ or $t \in \mathbb{N}$) such that the training set $\mathcal{S}_t$ is sampled from the distribution D_t. A key difficulty of continual learning is maintaining *plasticity*, i.e. the ability to adapt quickly to new tasks, while avoiding the phenomenon called *catastrophic forgetting* where adaptation to new tasks has reduced the performance on previous tasks. In continual RL, first explored in the thesis of Ring (1994), recent work (Abel et al., 2023) addresses the prospect of considering the continual RL problem as one of *endless adaptation* rather than *finding a solution*.

2 Framing continual learning in terms of RL - a first sketch

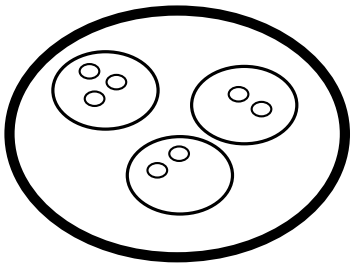
In this section we sketch an RL-based and multi-scale form of continual learning. This is a preliminary sketch only, intended to stimulate discussion about how one might capture some of these

[1] If they were, one could use dynamic programming.

aspects in the simplest way possible. What we are aiming to achieve with such a framework is to be able to address questions such as those in Section 3 in a learning-theoretic manner, and thereby to strengthen (mathematical) intuition about the corresponding properties of complex learning systems. There is inherently a wide gap between the messy, infinitely complex phenomena of living systems in the real world and the simplicity needed in a mathematical model in which one can carry out sound analysis. Our proposal is an attempted middle ground. We look forward to stimulating discussion in the workshop to improve the formulation.



(a) RL agent



(b) Hierarchy of complex agents composed of simpler ones

Figure 1: Ingredients of the proposed framework: RL agents of varying scales, with more complex agents being composed of simpler ones.

The basic ingredients of the proposed framework are as follows. We suppose an *environment* that produces stimuli at varying time- and space-scales, where the generating distribution is specified by the environment's multi-scale "state", and moreover, we assume this state can be altered by actions of agents (at respective scales). *Each agent (at a specific scale) is an RL agent* in the sense that it can not only sense the stimuli being emitted by the environment (at respective scale) but can also act upon the environment, thereby possibly altering the environment's state (at respective scale), and there is some form of *reward signal* communicated to the agent for each state-action-state triple it carries out. This signal is produced by an oracle that is either part of the (inductive bias of the) agent, or part of the environment. So far, there is nothing novel about this proposal except for the multiscale aspect. The setup has, moreover, been kept intentionally vague to allow for various more precise instantiations. We illustrate this simplistic form of RL agent in Figure 1a, where the updates that an agent may make to its internal state, e.g. under Q-learning, are also shown explicitly. Not shown are the other RL agents that might be interacting with the environment at other scales. Note: interaction between agents in this framework proceeds only via the environment; this may turn out to be inconvenient but has been postulated for now so each RL agent is close to the original single-agent, single-environment setup.

We now add several additional twists. First, we assume the environment evolves for all time but *agents are generally mortal*: every agent begins at its specific birth time $t_{\texttt{BIRTH}}$, and has a terminal internal state $s_{\texttt{DEAD}}$ after which it ceases to exist as an agent. (Its sub-agents may continue to exist.)

Second, as an optional richness for this framework (in the list below, this is only relevant for questions 1, 2, 4), it may be of interest to suppose an agent can arise either as a copy of an existing agent, or as a hybrid of two (or more?) existing agents. These modalities correspond somewhat to asexual and sexual reproduction respectively. Many details would need to be specified in an actual instantiation of such a framework, in particular, the cost of an agent producing offspring.

Finally, we add a hierarchical aspect. The goal of this is to capture the observation that complex living systems are composed in hierarchical fashion of simpler components. Components may come and go but a form of homeostasis maintains at least some global pattern among components (Wiener, 1948); this "pattern" is in a sense the identity of the large complex system. Our simple version of this is the following. We assume agents may contain sub-agents, always of smaller scale; see Figure 1b. To avoid specifying the global pattern that may define a "complex" agent of this kind, we simply assume

that there is an additional "integrative" signal passed from the large agent to its sub-agents whenever they take an action or update their internal parameters, i.e. whenever they change in any way. This signal - to keep things simple - will take values in $[-1, 1]$ with larger positive value indicating the extent to which the change reinforces the global pattern, and smaller negative values indicating the extent to which the change disrupts the pattern. The smaller agents should include such a signal (possibly heavily attenuated) as part of the total reward they maximize. The larger system is thus rewarding its components for acting in ways that maintain the integrity of the larger system. Admittedly, this formulation flies in the face of the (revolutionary but now generally accepted) view proposed by Kauffman (1993) that the larger system's pattern arises emergently by a process of self-organization of components and their interactions; it is in this sense not directed by the larger system but rather a spontaneously arising form of order. In our case, much as evolutionary pressures are simplistically represented as in the preceding paragraph, we have adopted in the present paragraph a simplistic version of homeostasis. These first choices may need to be revised if the framework does not allow realistic analysis of the questions we now pose.

3 Some questions to approach within this framing

The following are examples of questions one might address within an RL-based hierarchical continual learning framework as sketched above. In each case, by working on this type of question one may come up with reasons to alter the sketched framework. Our aim in this paper is for the sketch in the previous section to provoke thought and conversation, and to use questions like the ones listed below to further refine the framework and/or to define variants of the framework so that ultimately, progress can be made answering such questions.

1. Can we prove learning-theoretic benefits for acquiring inductive bias from a *single ancestor vs multiple ancestors*? (This question can be seen as studying asexual vs sexual reproduction through a learning-theoretic lens.) Presumably the relative advantages will depend on a tradeoff among different specifications such as computational capacity, memory, cost of fatality etc.. Is the proposed set-up suited to this type of study or should it be modified to better answer the question?

2. *Phylogenetic learning* (Wiener, 1948) is the negative feedback mechanism whereby traits that reduce the expected number of offspring (during lifetime) are less likely to appear in the inductive bias of individuals in subsequent generations. Can we already study this within the above framework or is there a need to incorporate more elaborate genetic-algorithm and/or multi-agent ideas explicitly?

3. Wiener refers to the feedback mechanism whereby behaviour is modified during a lifetime in response to experience as *ontogenetic learning* (Wiener, 1948). This can be basic supervised learning or RL. It can be formulated in an active-inference rephrasing (Friston et al., 2009) as reducing free energy between the agent's internal model and the environment; it has the effect of aligning the predictive capabilities of the agent with the dynamics of the environment. In this sense it involves a negative feedback loop between agent and environment, but where only the agent carries out planning/control, while the environment is sensed by the agent and also possibly modified by the agent. (Here we are viewing the sensor, controller, and effector of the feedback loop as belonging to the agent.)

    - Can we extend PAC-style learning theory to at least supervised learning in the above framework? This would mean a special case of the framework where the only actions an agent can take are requesting new samples and predicting labels, neither of which change the state of the environment.
    - If we adopt an extension of PAC-style learning-theoretic tools to RL, for example as formulated in (Fraser & Létourneau, 2022), can we further extend this to the above continual learning framework?

- Can we additionally give the environment learning capabilities? In fact we could suppose that the *only* distribution shift in time is due to the environment's adaptation to the agent's actions. Would this special case yield any conclusions that might not be available for general distribution shift, e.g. assuming a certain fixed learning "algorithm" and learning rate in the environment? Note: in the special case being considered, not only would the environment not shift independently of the agent, but the agent can potentially alter the distribution shift in the environment, something that is not possible in the traditional continual learning formulation.

4. If we view a continual learning super-agent as above being made up of a hierarchical "composition" of sub-agents whose acquired abilities they at least partially transfer forward to offspring, can we establish by learning-theoretic arguments a potential tradeoff so that a sweet spot between catastrophic forgetting and non-plasticity requires such a hierarchy, perhaps a deep one? This may be related to fascinating connections that have been discovered between such a sweet spot in continual learning on the one hand, and multi-scale aspects, resp. meta-learning, on the other (Kaplanis et al., 2018; Javed & White, 2019).

5. In similar spirit, is there some fundamental reason that hierarchical RL does not benefit from depth as much as neural networks do, or is this phenomenon observed in practice (where rarely more than two layers are effective) more due to engineering/algorithmic limitations?

6. Is there a notion of compositional sparsity similar to that for supervised learning (see Poggio & Fraser (2024) for a definition and discussion) that applies to hierarchical RL agents? or that can be formulated in some other way for the framework above?

Conclusion

We hope to gather further questions in conversation with workshop participants, as well as in our ongoing work on this framework. We have sketched here only a first draft of a possible framing of RL in the context of continual learning, inspired by cybernetics and offering perhaps the opportunity to use learning theoretic methods to answer cybernetics-style questions. As the presence of the human species on earth becomes ever more impactful on our biosphere, and AI risks intensifying some of this effect by boosting technological reach, it will also become increasingly important to consider not only the AI models we train but their interaction with an environment that itself adapts. Perhaps some version of the above framing and questions can play a role in highlighting this interaction. This perspective also raises the possibility that the bitter lesson noted by Sutton may only truly hold when resources are infinite, and one can then ask if there is an inherent tension between that so-far highly successful principle in machine learning and long-term sustainability on a finite planet.

Acknowledgments

The authors thank the organizers for creating this workshop with its unusual and very inspiring theme. We also thank the reviewers for their very helpful comments. These will significantly help us to refine the ideas sketched here, and we hope to add simulations soon.

References

David Abel, Andre Barreto, Benjamin Van Roy, Doina Precup, Hado P van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50377–50407. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9d8cf1247786d6dfeefeeb53b8b5f6d7-Paper-Conference.pdf.

Maia Fraser and Vincent Létourneau. Inexperienced RL agents can't get it right: Lower bounds on regret at finite sample complexity. In Sarath Chandar, Razvan Pascanu, and Doina Precup (eds.),

Conference on Lifelong Learning Agents, CoLLAs 2022, 22-24 August 2022, McGill University, Montréal, Québec, Canada, volume 199 of *Proceedings of Machine Learning Research*, pp. 327–334. PMLR, 2022. URL https://proceedings.mlr.press/v199/fraser22a.html.

Karl Friston, Jean Daunizeau, and Stefan Kiebel. Reinforcement learning or active inference? *PloS one*, 4:e6421, 02 2009. doi: 10.1371/journal.pone.0006421.

Khurram Javed and Martha White. Meta-learning representations for continual learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f4dd765c12f2ef67f98f3558c282a9cd-Paper.pdf.

Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Continual reinforcement learning with complex synapses, 2018.

S. A. Kauffman. *The origins of order. Self-organization and selection in evolution.* Oxford University Press, 1993.

Tomaso Poggio and Maia Fraser. Compositional sparsity of learnable functions. *Bulletin of the AMS*, 07 2024. URL https://cbmm.mit.edu/publications/compositional-sparsity-learnable-functions.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, 1994, 2005.

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

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

Gido M. van de Ven and Andreas S. Tolias. Three scenarios for continual learning, 2019.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2024. doi: 10.1109/TPAMI.2024.3367329.

N. Wiener. *The Human Use Of Human Beings: Cybernetics And Society*. Houghton Mifflin, 1950. URL https://books.google.ca/books?id=l9l6zquHvZIC.

Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. The MIT Press, 10 1948. ISBN 9780262355902. doi: 10.7551/mitpress/11810.001.0001. URL https://doi.org/10.7551/mitpress/11810.001.0001.