Towards Continual Reinforcement Learning:
A Review and Perspectives

Khimya Khetarpal[*]   khimya.khetarpal@mail.mcgill.ca
Mila, McGill University

Matthew Riemer[*]   mdriemer@us.ibm.com
Mila, Université de Montréal, IBM Research

Irina Rish   irina.rish@mila.quebec
Mila, Université de Montréal

Doina Precup   dprecup@cs.mcgill.ca
Mila, McGill University, DeepMind

Abstract

In this article, we aim to provide a literature review of different formulations and approaches to continual reinforcement learning (RL), also known as lifelong or non-stationary RL. We begin by discussing our perspective on why RL is a natural fit for studying continual learning. We then provide a taxonomy of different continual RL formulations and mathematically characterize the non-stationary dynamics of each setting. We go on to discuss evaluation of continual RL agents, providing an overview of benchmarks used in the literature and important metrics for understanding agent performance. Finally, we highlight open problems and challenges in bridging the gap between the current state of continual RL and findings in neuroscience. While still in its early days, the study of continual RL has the promise to develop better incremental reinforcement learners that can function in increasingly realistic applications where non-stationarity plays a vital role. These include applications such as those in the fields of healthcare, education, logistics, and robotics.[1]

1. Introduction

Recent advances in deep RL have demonstrated superhuman performance by artificially intelligent (AI) agents on a variety of impressive tasks. However, current approaches for achieving these results center around an agent that primarily learns how to master a narrow task of interest. Meanwhile, untrained agents often need to play far more rounds of these games over their lifetime than their human competition and even after doing so, lack the ability to generalize to new variations even for simple RL problems (Bengio et al., 2020). In contrast, humans have a remarkable ability to continually learn and adapt to new scenarios over the duration of their lifetime. This ability is referred to as continual learning. *Continual learning (CL)* is the constant and incremental development of increasingly complex behaviors. This includes the process of building complicated behaviors on top of those already developed (Ring, 1997) while being able to reapply, adapt, and generalize previously learned abilities to new situations. CL is a rapidly growing area of modern machine learning and particularly so for the study of deep learning. It is also closely related to settings such as *lifelong learning*, *online learning* or *never-ending learning*. In this paper, we are concerned with *continual reinforcement learning*. This is a natural fit, as RL inherently provides an agent-environment interaction paradigm amenable to studying the topic of learning in a continual fashion.

A continual learner can be seen as an autonomous agent learning over an endless stream of tasks, which has the following desiderata in the context of learning: 1) it can learn context dependent tasks, 2) it learns behaviors while solving presented tasks, 3) learning is task agnostic, 4) it learns

---

[*]. The authors contributed equally to this work.
1. This survey is a *continual* work in progress, so please reach out if we have omitted any important references.

incrementally with no fixed training set, 5) it learns behaviors that can be built upon later, 6) it retains previously learned abilities i.e. it minimizes catastrophic forgetting and interference, and 7) it adapts efficiently to changes experienced over time and recovers quickly. In its most ambitious form, CL occurs at every moment with no bounded set of tasks or data sets and no clearly presented boundaries between tasks. The learning agent should be able to transfer and adapt what it has learned from previous experiences, data, or tasks to new situations and make use of more recent experiences to improve performance on capabilities learned earlier.

With the rapidly growing interest in continual learning research, there have been extensive reviews of a large body of work in supervised continual learning such as Parisi et al. (2019), De Lange et al. (2019), and (Hadsell et al., 2020). Most of this work considers the *task incremental learning setting*. In this setting, each task is received with its training data in the form of labelled samples of inputs and desired outputs $(\mathcal{X}, \mathcal{Y})$ that are randomly drawn from a distribution $\mathcal{D}$. Here, the goal is statistical risk minimization on all seen tasks given limited or no access to the data from previous tasks after initial learning. In contrast, continual RL involves a sequential decision making problem over a stream of tasks where each task can be considered a stationary *Markov Decision Process* (MDP) (Puterman, 1994).

Key Contributions. Due to the generality of the continual learning problem and the struggle to define its scope, researchers have often interchangeably used the terms *multi-task learning*, *lifelong learning*, and *continual learning* in the field of RL. One of the primary goals of this work is to provide a concrete taxonomy of the different formulations and approaches under the broad umbrella term continual RL. To this end, the key contributions of this work concern: 1) a taxonomy and review of relevant problem formulations, 2) a taxonomy and review of families of approaches considered, and 3) a discussion of evaluation metrics for assessing continual RL agents and how relevant benchmarks can be used to generate non-stationarity during learning. Finally, we discuss shared ideas across neuroscience and continual RL as well as perspectives on challenges and open problems in the field.

Scope and Overview of the Survey. In this work, we discuss the literature which addresses different perspectives on continual RL. This includes multi-task learning, meta-learning, never-ending learning, non-stationary RL, and lifelong learning in the context of RL, explicitly. We limit our scope to the aforementioned topics. While there are several related topics such as transfer learning, representation learning, domain adaptation, domain randomization, and generalization, we do not cover them in detail in our survey. We first introduce the RL paradigm (Sec. 2) and highlight research directions related to the study of continual RL in Sec. 2.2. We then proceed to discuss why reinforcement learning as a paradigm is a natural fit for studying continual learning in Sec. 3. To this end, we present a broad taxonomy of continual RL formalism and approaches in Sec. 4 and Sec. 5, respectively. We then consider current and potential directions for evaluation of continual RL agents in Sec. 6. Looking to the future, we conclude by presenting connections to findings in neuroscience and by discussing perspectives on challenges and open problems in the field in Sec. 7.

2. Background

Notation: In this survey capital letters are used for random variables, while lower case letters are used for the values of random variables and for scalar functions. For consistency with prior literature, we largely follow the notation of (Sutton and Barto, 1998).

Typically we formalize reinforcement learning based on a finite, discrete-time MDP (Puterman, 1994; Sutton and Barto, 1998), which is a tuple $M = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, $p : \mathcal{S} \times \mathcal{A} \rightarrow Dist(\mathcal{S})$ is the environment transition probability function, and $\gamma \in [0, 1)$ is the discount factor. At each time step, the learning agent perceives a state $s \in \mathcal{S}$, takes an action $a \in \mathcal{A}$ drawn from a policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ with internal parameters $\theta \in \Theta$, and with probability $p(s'|s, a)$ enters next state s', receiving a numerical reward following $r(s, a)$ from the environment. It is also possible that the environment is a partially

observable POMDP environment (Kaelbling, 1993). In this case, the environment also consists of a potentially stochastic function that generates observations o from the current state with probability $x(o|s)$. In these environments, agents must generate a belief about the current state based on their history of interactions and perform RL based on this belief. As such, we will mostly focus on reasoning with respect to the true environment state throughout this survey as extensions to partially observable settings are straightforward.

The environment's transition dynamics can be modeled by the one-step *state-transition probabilities*,

$$p(s'|s,a) \doteq P_{ss'}^a = Pr(S_{t+1} = s'|S_t = s, A_t = a) \tag{1}$$

and one-step expected rewards,

$$r(s,a) = R_s^a = E[R_{t+1}|S_t = s, A_t = a] \tag{2}$$

for all $s, s' \in S$ and $a \in A$. $P_{ss'}^a$ and R_s^a together form the one-step model of the environment.

The goal of the agent is to maximize the expected value of the accumulated discounted reward from time-step t. More precisely, the *return* G_t obtained from time step t is defined as:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{3}$$

where $0 \leq \gamma < 1$ is the *discount factor*. The agent's behavior is determined by its *policy* (stochastic and stationary), a mapping from states to probabilities of taking each of the admissible primitive actions, $\pi : S \times A \rightarrow [0,1]$. The value of being in a state is determined by the *state-value function* $v_\pi(s)$, defined as the expected return starting from state s, and then following policy π is defined as:

$$v_\pi(s) = \mathbb{E}_\pi\left[G_t \Big| S_t = s\right] = \sum_a \pi(a|s)\left[r(s,a) + \gamma \sum_{s'} p(s'|s,a)v_\pi(s')\right]$$

Analogous to the state value function, the *action-value function* $q_\pi(s,a)$ is defined as:

$$q_\pi(s,a) = \mathbb{E}_\pi\left[G_t \Big| S_t = s, A_t = a\right] = r(s,a) + \gamma \sum_{s'} p(s'|s,a)v_\pi(s')$$

For a finite MDP, there exists at least one deterministic policy, that is better than or equal to all other policies. This is an optimal policy π^*. The optimal policy π^* achieves the optimal state-value function or the optimal action-value function. The *optimal state-value function* $v_*(s)$ is the maximum value function over the class of stationary policies (see Eq (4)). Similarly, the *optimal action-value function* $q_*(s,a)$ is the maximum action-value-function over all policies as defined in Eq (5).

$$v_*(s) = \max_\pi v_\pi(s) \tag{4}$$

$$q_*(s,a) = \max_\pi q_\pi(s,a) \tag{5}$$

RL algorithms can be classified as *model-free* or *model-based* in nature. Knowing or estimating a model of the environment, and using this model to compute value functions and learn policies is called *planning* in RL. Algorithms such as policy-evaluation exploit the iterative Bellman expectation backup to find optimal solutions for prediction problems. However, in most practical situations, the model of the world is rarely known and many practical algorithms fall in the *model-free* regime.
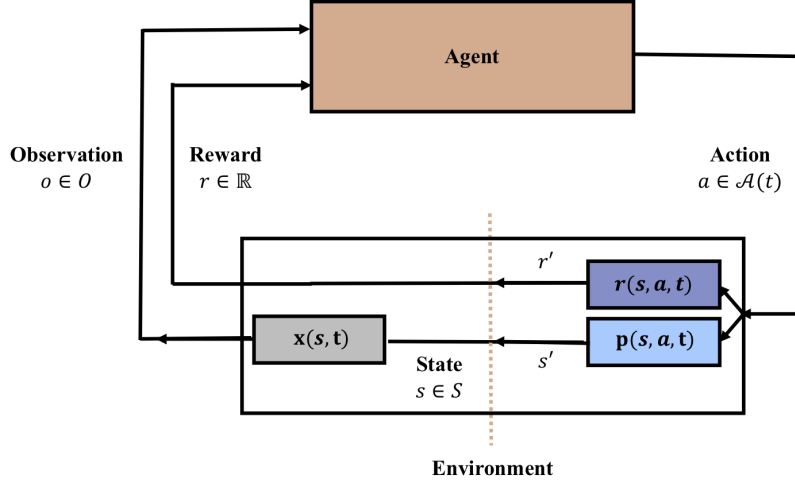
Figure 1: **Agent-Environment Interaction** with Potentially Time Dependent Environment Components. Extending Figure 3.1 of Sutton and Barto (1998) to highlight the agent-environment interaction in reinforcement learning.

2.1 Defining Tasks in RL

In the RL literature the concepts of tasks and non-stationarity have been defined in many different ways depending on the context. As such, this is a point of confusion in the literature that we hope to provide some clarity on here. From our perspective, there are two primary views that have been taken in past work. In the first view, we consider that actual components of the RL environment may exhibit some time dependence. We will position most of our survey in terms of this view as it is the most common perspective taken in the continual learning literature to date. However, another valid perspective to take is that the underlying physics of the world fundamentally exhibit stationary dynamics and that perceived non-stationarity is really only a consequence of unobserved phenomena that result in changed dynamics from an agent's own (potentially ignorant) perspective. We will take this view only when discussing approaches for *context detection* in Sec. 5.3.1 or approaches for multi-agent RL as this view is key to the underlying theory behind these techniques.

Non-stationary Function View: We highlight the view that fundamental components of the RL environment may exhibit time dependence in Figure 1. Indeed, in the extreme case it is possible that the transition function, reward function, observation function, and action space may all depend on time. In this setting, for the purposes of this paper, we will define a task z as constituting a stationary MDP $M^{(z)} = \langle \mathcal{S}^{(z)}, \mathcal{A}^{(z)}, p^{(z)}, r^{(z)}, \gamma^{(z)} \rangle$ with initial state distribution $p_0^{(z)}$. This implies that there is a discrete set of tasks. However, in the extreme case, this set may be of infinite size where no task is ever visited for more than a single time step. In principle, it is possible that MDPs may vary as a continuous function of time. That said, this distinction is only relevant for RL in continuous time MDPs, which is beyond the scope of this paper.

Partially Observable View: Another view on non-stationarity is that it is not a feature of the environment itself, but rather only the agent's perspective in that environment. This viewpoint certainly appears realistic when we make comparisons to human learning. It appears that most non-stationarity that humans experience in their lives is the result of interacting in a huge environment with a massive number of agents who are changing their behaviors for possibly unknown reasons over time. When we take this view in the context of RL, a task is defined as an unobserved component of the state that an agent must develop beliefs about to achieve optimal performance.

While these two views of tasks and non-stationarity may seem contradictory on the surface, we believe that they actually serve as complementary perspectives on the same problem. The non-stationary function view describes the agents perspective on the problem and is easier to reconcile with similar formulations in the context of supervised learning and bandit learning. Meanwhile, the partially observable view makes it easier to formalize convergence analysis. Moreover, the partially observable view puts emphasis on different aspects of the problem by focusing less on non-stationarity and more on related complications like non-Markovian observations and non-ergodic environments (or even ergodic environments with very high mixing times).

2.2 A Spectrum of Learning Settings

So far we have discussed the basics of RL in a stationary environment. However, in general, it is possible for each component of our environment to be non-stationary. In Figure 1 we illustrate the RL setup where the environment includes explicit time dependence in each component. This represents the most ambitious formalization of continual RL (as we will discuss in more detail in Sec. 4). Indeed, the agent must deal with potential non-stationarity in state transitions, the reward function, the way observations are produced, and even the availability of actions over time. As this setting presents a number of serious challenges, much of the work in the community has focused on less ambitious variations of the problem that are more targeted to highlight particular aspects of the difficulty encountered by agents as they continual learn. As such, in this section, we will highlight a spectrum of research directions of particular relevance to the study of continual RL.

Setting	Multiple Domains Of Deployment	Multiple Required Skills	Universal Master Policy	Non-stationary Evolution
Domain Adaptation	✓	X	X	X
Transfer Learning	✓	✓	X	✓
Meta-Training and Meta-Testing	✓	✓	X	X
Multi-task Learning	✓	✓	✓	X
Continual (Lifelong) Learning	✓	✓	✓	✓

Figure 2: **A Spectrum of Learning Settings Leveraging Prior Knowledge**: For each setting we consider whether they typically involve multiple domains, multiple skills, a universal master policy to solve all tasks, and a non-stationary evolution of the task distribution. Note that when discussing a universal master policy, we are suggesting that it is a requirement of the environment that the agent be able to produce a policy for each task or subtask when provided with the appropriate context or input. We are not specifying how this should accomplished, which is a property of the approach rather than the setting.

Domain Adaptation: As detailed in Figure 2, domain adaptation is the process of adapting a policy for a specific skill to a new domain. It generally involves building a separate policy for each domain and typically does not require algorithms to address environment non-stationary during training. A natural use case for domain adaptation in the context of reinforcement learning is *sim2real* transfer when an agent is trained in a simulation environment and then adapted as a result of interaction in the real world. Domain randomization is a related approach that deals with a transfer type of setting. The goal here is to be able to generalize learning from source domains to target domains. We refer the reader to Tobin (2019) for a comprehensive discussion on domain randomization and related topics such as domain adaptation.

Transfer in RL: Learning each task from scratch may require a huge amount of data to achieve adequate performance. Additionally, learning from scratch is computationally expensive and intractable for large scale problems such as everyday robotics. However, to learn about multiple potentially diverse tasks with limited data is also a challenging problem. A large body of work concerned with an agent's performance on more than one task has extensively studied the topic of *transfer learning*: training on *source* tasks to perform efficient policy modifications during training on a single *target* task drawn from a distribution of related tasks. We refer the reader to Taylor and Stone (2009) for an extensive review of work focused around transfer learning in RL. As highlighted in Figure 2, transfer learning settings generally include multiple domains and skills that have to be learned in the presence of non-stationarity. However, a key distinction with other settings of more interest to the field of continual RL is that in transfer learning settings it is generally assumed that a separate policy is learned for each task and that task boundaries are given. In transfer learning, the learning process is generally broken up into distinct phases such as *pre-training* and *fine-tuning*.

Meta-Training and Meta-Testing: A related setting is the *meta-training* and *meta-testing* protocol common in the meta-learning literature. In this setting, an agent first performs *meta-training* about how to learn to generalize efficiently on a distribution of tasks and this meta-learning model is transferred to a *meta-testing* distribution of tasks where the eventual performance of the policy of each task is used as a basis for comparison. As indicated in Figure 2, in some sense this meta-learning protocol results in an easier optimization process than we see in traditional transfer learning. This is because *meta-training* consists of drawing tasks to learn on from a stationary distribution and that distribution should theoretically be approximately stationary during *meta-testing* as well. As a result, the evolution of tasks is not really non-stationary as in more generic transfer learning. For a more in depth discussion on how this training paradigm works see Sec. 5.3.2.

Multi-task RL: More closely related to the aforementioned task incremental learning setting, is *multi-task reinforcement learning*. As pointed out in Figure 2, while multi-task learning no longer concerns itself with learning over a non-stationary task distribution, it generally includes learning a universal policy for all task even if that policy may be hierarchical in nature. In a commonly used formulation of multi-task RL, the agent is required to learn a series of sequential decision making tasks $M^{(1)}, ..., M^{(z_{max})}$ over its lifetime. The agent will learn the tasks consecutively, potentially acquiring multiple trajectories within each task before moving to the next. It is a common assumption that these tasks may be interleaved i.e. the agent might revisit earlier tasks, but the agent does not control the order of tasks. This setting has often been studied for *online reinforcement learning* (Wilson et al., 2007; Ammar et al., 2014). Alternatively the tasks do not necessarily arrive in a sequential fashion, and learning might not be considered in a fully online setting. In contrast, data could be generated by many different behaviours and be made available in a batch generated beforehand in the *offline reinforcement learning* setting. We refer the reader to Levine et al. (2020) for a tutorial and review of the offline setting. In the well posed case of multi-task reinforcement learning, the agent's overall objective is to maximize performance across the distribution of tasks being considered. It is desired that such an agent also perform well on out-of-distribution data usually not seen during training and generalize to similar or related tasks. How this is achieved could vary from learning a single universal policy whilst minimizing the expected accumulated average return on all tasks to learning a set of optimal policies for each task, to learning shared skills with a meta-controller.

Continual (Lifelong) RL: A prime challenge faced by continual RL agents is to be able to retrieve relevant information from a massive sensory data stream. A common approach entails compressing information in one way or another such as discovering information bottleneck states, subgoals, or state abstractions to name a few. Additionally, a continual learner has no direct access to all previous experiences and its memory is often limited. For such an agent, it is essential to properly assign credit to key events over the course of its lifetime commonly known as the *credit-assignment* problem. Additionally, while successful continual learning agents must have resistance to catastrophic forgetting, it is not immediately clear if agents must perform well on all previously seen tasks. As we note in Figure 2, continual learning adds the concern of learning over a non-stationary task

distribution to the complication of learning a universal policy over tasks. For instance, someone who learns to play tennis at a young age may not necessarily perform well on this previously seen task in later stages of life if it has not been rehearsed. As such, quick adaptation and building on relevant previously learned behaviors are central to the study of continual RL.

2.3 Important Related Topics

When it comes to actually achieving good continual RL performance, it is not possible to totally disentangle the study of continual learning from other very important fields of study in RL.

Representation Learning: Learning good representations with minimal overlap is fundamental to much of the work related to multi-task deep RL. Considering representation learning is not unique to continual RL and is a common concern across literature on supervised continual learning, curriculum learning, transfer learning, multi-task learning, and more generally work towards broader AI. We refer the reader to discussions on encoder based lifelong learning by De Lange et al. (2019).

Generalization in RL: Due to the inherent difficulty of training and testing on the same environment in deep reinforcement learning, several efforts have been made in studying the generalization abilities of RL agents. Generalization in RL has been investigated by creating different game modes (Farebrother et al., 2018) and game levels (Nichol et al., 2018) for training and testing, by exposing internal parameters of various classic environments (Packer et al., 2018), and by procedurally generating environments (Justesen et al., 2018; Zhang et al., 2018b; Cobbe et al., 2018). There has indeed been surprising evidence to the extent that RL agents tend to overfit even in simple settings (Bengio et al., 2020). Recent theoretical work (Du et al., 2019a) suggests that perhaps for a class of tree-like MDPs, generalization might even be improbable. While robust generalization is a central capability for effective continual RL, most of the literature on generalization in RL studies it within the scope of a simple transfer learning setting. In this way, it is possible to study generalization without conflating its properties with the optimization difficulties specific to continual RL.

3. Reinforcement Learning: A Natural Fit For Studying Continual Learning

It has been well known for decades that the primary challenge for neural networks when learning over a non-stationary stream of data is balancing the *stability-plasticity dilemma* (Carpenter and Grossberg, 1987). This dilemma highlights the tension between prioritizing recent experiences and past experiences when training neural networks. A common failure case is the so called *catastrophic forgetting* problem (McCloskey and Cohen, 1989), where the network adapts to recent experiences while significantly deteriorating its capabilities on past experiences. However, in a certain sense, the formalism of RL in continuing environments actually provides us with means of directly studying the *stability-plasticity dilemma*.

For RL in a continuing environment, we define an objective function $J_{\text{continuing}}$ which is to learn a policy π that maximizes the value function $v_\pi(s)$. We seek to maximize this objective function at each point in time:

$$J_{\text{continuing}}(\pi) = v_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s \right] \tag{6}$$

Here our discounted objective with respect to the policy is to maximize the expected long-term discounted returns of this policy in the current state s. So, the objective we would like to maximize does not just concern itself with the current state, but rather the full expected future distribution of states as well. Some proportion of the expected future distribution of states is likely to be similar to states from the past. As such, this provides a general paradigm for addressing the prioritization

problem of the *stability-plasticity dilemma*. We care about the present and the past proportionally to how representative these distributions will be of the future. However, because the future is generally unknown, we will have to use our expectations about the future to guide our prioritization instead.

As RL in continuing environments is clearly a challenging problem, the vast majority of work has focused on easier learning problems such as RL in episodic environments and supervised learning. However, when we introduce non-stationary environment dynamics to either the episodic or supervised settings, naive approaches can be shown to produce biased updates that are overly focused on the current experience distribution rather than the expected future distribution. This is because non-stationary dynamics undermine the assumptions of many popular algorithms for these settings such as policy gradient based approaches in episodic reinforcement learning and stochastic gradient descent (SGD) in supervised learning. However, even settings with changing time correlated dynamics can be modeled in terms of the formulation of RL in continuing environments. This perspective can be illuminating and shed light on the pervasive *catastrophic forgetting* effects experienced by popular approaches when applied to non-stationary settings.

Catastrophic Forgetting in Continual Episodic Reinforcement Learning Much of the progress in deep reinforcement learning in recent years has been in application to so called *episodic* environments. These are environments that exhibit a clear decomposable structure into time windows that are drawn from a stationary distribution. This assumption about an environment can be very useful when applicable as it allows us to consider the performance of our policy during a single episode as a sample from a random variable representing our full objective:

$$
\begin{aligned}
J_{\text{episodic}}(\pi) = v_\pi(s) &= \mathbb{E}_\pi \left[\sum_{k=0}^{H-1} \gamma^k R_{t+k} \,\middle|\, S_t = s \right] \\
&= J_{\text{continuing}}(\pi) - \mathbb{E}_\pi \left[\sum_{k=H}^{\infty} \gamma^k R_{t+k} \,\middle|\, S_t = s \right]
\end{aligned}
\tag{7}
$$

The key difference between equation 6 and equation 7 is that the latter only optimizes over a future horizon H until the current episode terminates rather than until the end of the agent's lifetime. Unfortunately, approaches developed for this setting experience biased optimization when applied to continual episodic reinforcement learning settings where episodes are drawn from a non-stationary distribution that changes over time. In these settings, because the stationary episodic structure assumption is no longer valid, we must consider our objective in the continuing setting from equation 6. As such, it is clear that the standard episodic objective is biased to only optimize on the current episode distribution while disregarding the likely far more important expected future episode distribution over the agent's lifetime. Blindly applying episodic reinforcement learning approaches in this way to non-stationary settings leads to biased optimization, which is likely to cause catastrophic forgetting effects, for example, when we sample repeatedly from each task before changing task distributions.

This kind of situation appears naturally in many real-world applications of continual RL. For example, let us consider an agent learning in an environment whose dynamics can be naturally broken down into several modes of operation that repeat over the course of an hour, day, or week. One example would be a taxi driving agent that experiences very different traffic and demand patterns depending on the time of day (i.e. morning rush hour, lunch-time, evening rush hour, etc.). If updates to our agent's policy are greedy based on only the current mode of operation, optimization will become quite difficult as many learning steps may be taken for a single mode without regard for how an agent performs on the other modes of operation. Because these modes are temporally correlated, this greedy optimization will directly encourage catastrophic forgetting of other modes during the learning of the current mode. On the other hand, an agent that correctly identifies that each mode of operation always constitutes a constant proportion of its expected future distribution of modes

can perform balanced updates that do not overvalue the current mode of operation. Indeed, these balanced updates are a natural consequence of optimizing for the long-term objective in equation 6 rather than myopic short term optimization as in equation 7.

Catastrophic Forgetting in Continual Supervised Learning In applications of supervised learning to non-stationary data distributions, we actually see a very similar phenomenon that leads to catastrophic forgetting. Gradient based algorithms like SGD and popular variants of SGD either implicitly or explicitly assume that they are optimizing over a stationary i.i.d. distribution of data. When we optimize over a non-stationary distribution of data, we can see SGD as related to the deterministic policy gradient theorem with a differentiable reward function (i.e. the loss function). However, SGD simply optimizes on the current distribution of experiences without regard for the expected future distribution of experiences. As a result, when standard approaches to supervised learning are applied to non-stationary setting such as the popular *locally i.i.d.* setting (Kirkpatrick et al., 2017; Lopez-Paz and Ranzato, 2017; Riemer et al., 2019; Chaudhry et al., 2019), it is easy to see how the optimization is biased to only update for the current distribution while potentially catastrophically forgetting its capabilities on old data distributions that may once again be relevant in the future. See Appendix A for a more detailed description of how to view continual supervised learning as a special case of RL in a continuing environment.
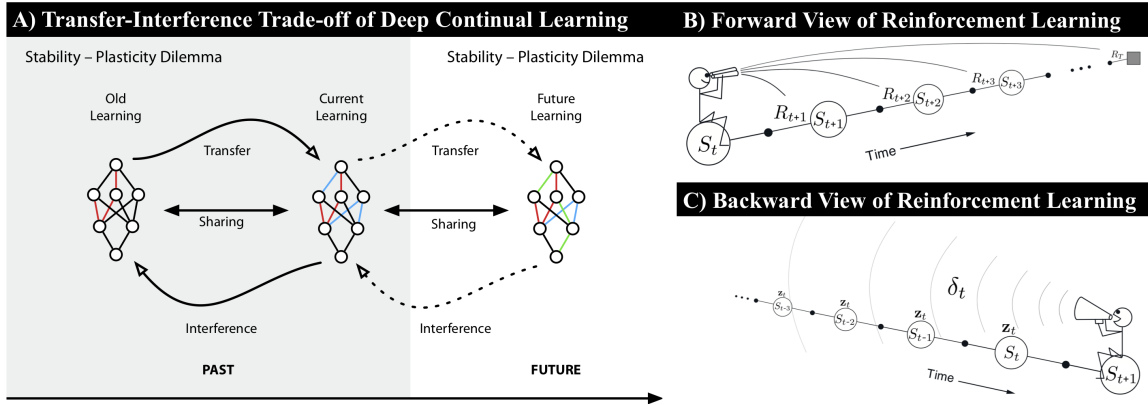

Figure 3: **Reinforcement Learning and the Stability-Plasticity Dilemma**: A) Depicts the stability-plasticity dilemma and its relation to both weight sharing and transfer dynamics over time (from (Riemer et al., 2019)). B) Depicts the forward view of RL where we evaluate the current state based on expected future rewards (from (Sutton and Barto, 1998)). C) Depicts the backward view of RL where we leverage recent states and rewards to correct our evaluations of past states (from (Sutton and Barto, 1998)).

Studying Time Correlations in CL with RL As we just discussed, a key assumption in supervised learning is that data is independently, identically distributed (i.i.d) and drawn from a fixed distribution. However, in reinforcement learning, data points are not only time correlated, but also, do not come from a fixed distribution. Considering that data points are not i.i.d, studying the learning dynamics within RL facilitates a deeper understanding of long-term memory. In particular, a common issue in supervised CL is to determine how newly seen data might be related to the previously seen data. However, it is counter intuitive to analyze this (as is typically done) without an explicit dependency on time. We should note, however, that one could always formulate a supervised learning problem within the RL framework as highlighted by (Barto and Dietterich, 2004).

Understanding the Forward and Backward view of CL with RL RL as a paradigm offers algorithms which are forward focusing or backward focusing (Sutton and Barto, 1998). For each state

visited, the forward view allows the agent to look forward in time to consider the future rewards and decide how best to combine them. Due to the unavailability of future states a more efficient incremental strategy makes use of backward-view computations. One could imagine that typical problems of balancing the stability-plasticity dilemma could potentially be more naturally understood with the forward and backward view toolkit readily available in RL. For example, in Figure 3 we highlight how the ideas of forward and backward transfer and interference in continual learning (Riemer et al., 2019) are naturally subsumed by RL's notion of a forward and backward view.

4. A Taxonomy of Continual RL Problem Formalisms

As we can see in the last section, a common problem plagues continual learning applications of both episodic RL and supervised learning. This key problem is that the deployment environment disobeys the assumptions of the underlying algorithm used to learn about it. Ideally, we would like to develop a theory about how to learn to act in environments with dynamics that are arbitrarily non-stationary. Unfortunately, this is not possible in general because arbitrary non-stationary dynamics may not provide any consistent signal to learn from. As a result, it is critical to make the right set of assumptions, that fully characterize the kinds of non-stationary environments we care about, in order to make progress. The result has been the development of a full spectrum of interesting environment settings in the recent literature. In this section, we attempt to provide a taxonomy of formulations; including prominent assumptions about non-stationarity that have been either explicitly or implicitly considered in the literature. We present a categorisation of the existing problem formulations in continual RL in Figure 4 along two primary dimensions: 1) the degree of non-stationarity, and 2) the drivers of non-stationarity. The degree of non-stationarity addresses what elements of the environment experience non-stationarity. Higher degrees indicate that non-stationarity is present in a greater number of the environment components highlighted in Figure 1. In contrast, the drivers of non-stationarity capture the causal assumptions that can be made about the nature of the evolution of non-stationary environment dynamics. We will highlight these dimensions of variation for problem formulations in the literature in more detail throughout this section.

4.1 The Degree of Non-stationarity in RL Environments

Revisiting the definition of a MDP as a tuple $M = \langle \mathcal{S}, \mathcal{A}, r, p, \gamma \rangle$, it is quite common to assume that there is non-stationarity in either rewards or transition dynamics. Meanwhile, it is also possible that the observation function x or action space $\mathcal{A}$ is non-stationary as highlighted in Figure 1. Very few researchers have actually explored this setting in a fully general manner with every component being potentially non-stationary to date (i.e. a degree of 4 in Figure 4). As we discussed earlier, this level of non-stationarity can be quite problematic for standard policy optimization approaches and lead to catastrophic optimization failures. As a result, we will begin by highlighting some prominent formulations with lower degrees of non-stationarity.

Non-stationarity in the reward function alone can be modelled in the problem formulation through many different approaches. For instance, *goal directed* learning without a task specific reward (see Sec. 5.2.4) could be interpreted as containing non-stationarity only in the reward function. These approaches allow the agent to learn about other stationary elements of the environment while being able to adapt as the reward function changes with changing goals. In fact, a family of related methods consider the *unsupervised RL* setup (see Sec. 5.2.4) with the aim of learning task-agnostic behaviors, wherein the agent is trained with no explicit rewards, but later evaluated on specific tasks. The shift from no reward to a task-specific reward distribution could be interpreted as a source of non-stationarity that the agent is explicitly trained to prepare for by learning as much as it can about the common stationary elements of the environment.

It is also somewhat common to consider environments where only the transition dynamics change and the reward function remains constant. Moreover, the most common case of 2nd degree non-
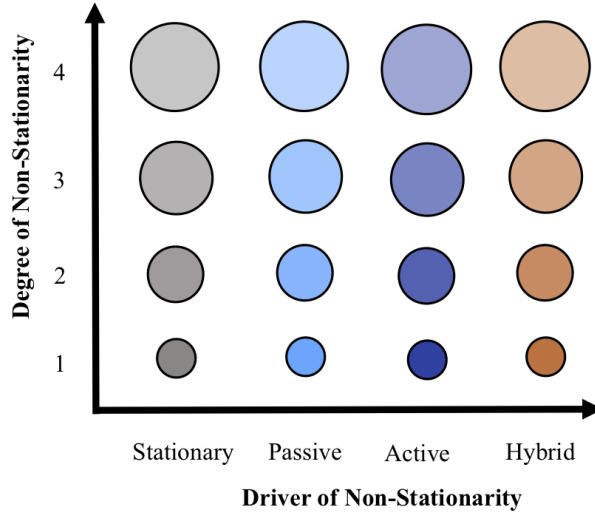
Figure 4: **Taxonomy of Continual RL Formalisms**: Problem formulations in continual RL can be categorized along two primary dimensions: 1) the degree of non-stationarity, and 2) the drivers of non-stationarity. Along the x-axis, we consider the drivers of non-stationarity i.e. how non-stationarity or task distribution evolution can be categorized by causal assumptions. Potential drivers include stationary task distributions with a fixed distribution over tasks, passive sources of non-stationarity where the evolution of the tasks does not depend on the agent's behavior, active sources of non-stationarity where the agent's behavior dictates the evolution of the tasks in the environment, and hybrid settings where both active and passive sources of non-stationarity are present. Along the y-axis, we consider the degree of non-stationarity with respect to the critical elements of the environment highlighted in Figure 1. First degree non-stationarity refers to settings where only one environment element from either p, r, x, or $\mathcal{A}$ changes. On the other extreme, fourth degree non-stationarity includes time dependence in each element of the environment.

stationarity is to assume that both the transition dynamics and reward function may change as is common in the meta-learning literature (see Sec. 5.3.2). This setting itself can be considered quite challenging and is probably the most ambitious formulation that is commonly attempted in the literature. It is to be noted, however, that most of these works do not explicitly consider the continual RL setting, but the concepts could be easily extended to study this more general form of agent deployment.

While less common, work on non-stationary RL has also considered the case of a changing action space (Chandak et al., 2019) and observation function (Zhang et al., 2020). These sources of non-stationarity may be very important for particular practical applications of continual RL. However, this remains an under explored problem setting in comparison to those focused on transition and reward dynamics. Moreover, to date, we are not aware of any approach that allows for the maximum of 4th degree non-stationarity. It remains an open question to what degree learning is still possible when we allow for non-stationary in each component of the environment and it is likely that additional assumptions about the scope of the problem will be necessary to develop agents that perform well in this very ambitious setting.

4.2 Causal Assumptions About the Non-stationary Environment Evolution

As highlighted in Figure 4, an important point of distinction between different potential formalism for continual RL is what assumption is made with regard to the agent's own causal role in influencing the non-stationary evolution of the environment. First, we should note that it is possible that the non-stationarity of the environment is drawn from a stationary distribution that the agent cannot influence. Additionally, non-stationary settings that are influenced by the agent's behavior can be said to be active while settings where the non-stationarity is independent of the agent's behavior can be considered passive. It is also possible that the environment may be non-stationary both based on the agent's behavior and based on causal factors beyond the agent's control.

4.2.1 STATIONARY TASK DISTRIBUTIONS: MULTI-TASK LEARNING

Perhaps the most common assumption in the literature is that while there may be multiple tasks z, which each have their own associated MDP $M^{(z)}$, these tasks are sampled from a unknown but *fixed* distribution $p(z)$. This assumption is common for learning in settings such as multi-task learning (Caruana, 1997; Wilson et al., 2007; Ammar et al., 2014). This setting can be far more difficult than single task learning as the environments may require diverse and possibly conflicting behaviors to achieve success. This setting can also be viewed with the lens of multi-objective optimization where potentially conflicting requirements make it important to consider concepts like *Pareto Optimality* of policies (Gábor et al., 1998) for cases where it is not possible to learn an optimal policy for each task with the same shared set of parameters. Additionally, when gradient based learning and function approximation are applied to multiple diverse tasks simultaneously, this is known to often result in interfering gradients (French, 1991; Caruana, 1997) across tasks. As noted by (Yu et al., 2020a), conflicting gradients across tasks can be particularly damaging to the learning process when tasks have differing gradient magnitudes and when the loss landscape exhibits high curvature.

A large body of work is focused on online multi-task learning where tasks are faced in a sequential fashion. A class of algorithms for multi-task RL in this setting use non parametric Bayesian models facilitating knowledge sharing across tasks. One approach is to model the distribution over tasks (i.e. MDPs) and use this distribution as a prior when a new task is seen (Wilson et al., 2007; Ammar et al., 2014). In the real world, data is often collected from experiences in many different environments and it is desirable to share knowledge learned from previous experiences (Li et al., 2009).

Recently popular approaches for "meta-learning" (Schmidhuber, 1987; Bengio et al., 1992) have come to leverage protocols for what is called *meta-training* and *meta-testing* that also align with this assumption of a stationary task distribution $p(z)$ (Wang et al., 2016; Duan et al., 2016; Finn et al., 2017; Mishra et al., 2017; Nichol and Schulman, 2018). See (Vilalta and Drissi, 2002; Hospedales et al., 2020) for an in-depth survey of meta-learning. In *meta-training* these meta-learning approaches learn a policy that can quickly adapt to tasks in the distribution $p(z)$. While it is generally assumed that during *meta-testing* different tasks are used for learning than those learned in *meta-training*, we only expect this generalization to be possible to the extent that there are shared commonalities between the *meta-testing* tasks and the *meta-training* distribution $p(z)$. In fact, there is often nothing done to explicitly prepare for out of distribution generalization or distributional shifts. See Sec. 5.3.2 for a more in depth discussion and examples of this problem setup.

4.2.2 PASSIVE NON-STATIONARITY

In passive non-stationary environments, we assume that the non-stationary behavior (i.e. the evolution of tasks) does not depend on the behavior of the agent itself when interacting with the environment. This allows us to model the evolution of tasks using the stochastic function $p(z'|z)$ as in (Choi et al., 2000) without having to consider the effects of our own changing policy on this distribution. While this setting is less human realistic in some sense, it is quite practical and describes the setting of the clear majority of experiments in the continual reinforcement learning literature.

An extension of this idea is to consider task transitions that happen at irregular multi-task intervals in the Semi-Markov Decision Process (Sutton et al., 1999) setting as in (Hadoux et al., 2014b). For example, this setting can model tasks that transition after the termination of each episode. Indeed, much of the work on learning to infer latent contexts (see 5.3.1) and learning to adapt (see 5.3.2) can be seen as part of this category of approaches that assume a passive source of non-stationarity.

One kind of popular assumption that makes solving passive non-stationary MDPs tenable is local consistency. Even mild assumptions such as restricting the change in an MDP to be slow or Lipschitz can provide a basis for robust worst case optimization (Lecarpentier and Rachelson, 2019). Another assumption that is common is that of piecewise stationarity (Padakandla et al., 2019) or local stationarity with change points. Generally approaches in these settings mentioned consider the problem of learning about the dynamics of $p(z'|z)$ to be too challenging and instead adopt a purely reactive philosophy with respect to environment changes. Alternatively, to account for both slow and abrupt changes in transition and reward dynamics, one approach is to quantify variation in the reward function and transition kernel over time in terms of their respective variation budgets Δ_r, Δ_p. In this framework, it then possible to consider assumptions around variation budgets in tabular (Cheung et al., 2020; Domingues et al., 2020) and linear function (Touati and Vincent, 2020) MDPs to improve learning in the face of non-stationarity.

4.2.3 Active Non-stationarity

In active non-stationary environments, we consider that our behavior may have an impact on the nature of the non-stationarity in the environment itself. This concept is foundational to work on intrinsic curiosity (Schmidhuber, 1991; Chentanez et al., 2005; Singh et al., 2009; Barto, 2013; Achiam and Sastry, 2017) and learning an agent's own curriculum (Schmidhuber, 2013; Justesen and Risi, 2018; Srinivasan et al., 2019). See Portelas et al. (2020) for a survey of automated curriculum learning approaches for deep RL. For example, agents have in some settings learned to generate their own imagined environments for future play (Wang et al., 2019), their own goals (Florensa et al., 2018; Racaniere et al., 2019), and their own opponents (Sukhbaatar et al., 2017). Agents in ambitious open-ended learning settings, where both the agent and the designer do not know the task or the domain ahead of time, could also experience active sources of non-stationarity. For example, these settings may experience a changing state and action space over the course of an agent's lifetime while continuously generating creative behaviors (Doncieux et al., 2020). In active non-stationary environments, we can now assume that environment dynamics may vary in a Markovian way following the function $p(z'|s, a, z)$ as in (Ong et al., 2010). See Sec. 5.3.3 for instances of approaches attempting to solve such a formulation.

4.2.4 Active and Passive Non-stationarity

It is also possible to consider settings where the non-stationarity can be controlled in some ways by changing the agent's behavior, but also influenced by causal mechanisms beyond the agent's control. For example, in multi-agent RL an agent may be able to influence the learned behavior of another agent by changing its own behavior and play an active role in shaping the non-stationarity of the RL problem from its perspective (Foerster et al., 2018a,b; Kim et al., 2020). However, certain aspects of the non-stationary evolution of the environment based on changes in multi-agent behavior is likely beyond the control of each individual agent, which is an important consideration when designing continual models. This hybrid non-stationary setting, while less common in the literature to date, seems like the most representative of many real-world applications. As such, designing agents that can function in this kind of non-stationary environment represents an emerging frontier for increasingly ambitious continual RL research that can tackle real-world problems.

5. A Taxonomy of Continual RL Approaches

In this section, we discuss in some detail a taxonomy of approaches for continual RL as highlighted in Figure 5. We will describe three high level clusters: those focused on explicit knowledge retention, those focused on leveraging shared structure, and those focused on learning to learn.
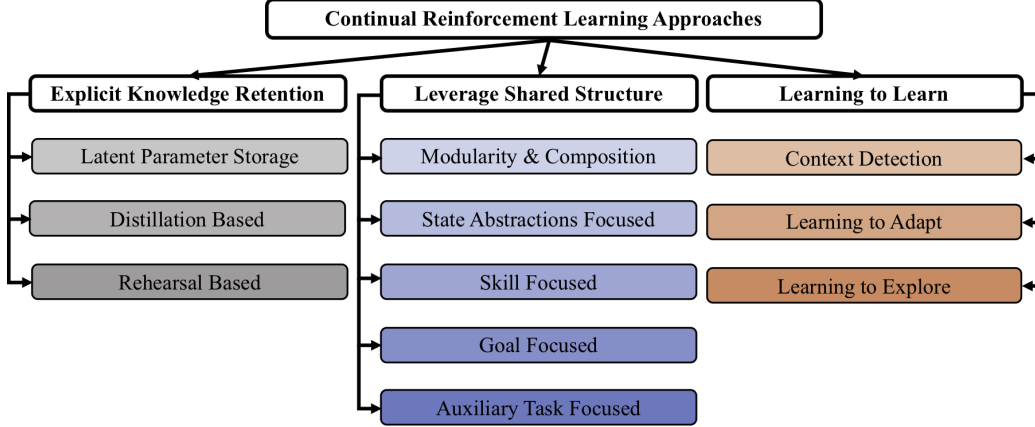

Figure 5: **Taxonomy of Continual RL Approaches**: A diagram illustrating different clusters of approaches for continual RL, highlighting prominent threads of research within each family.

5.1 Explicit Knowledge Retention

Due to concerns about agents experiencing catastrophic forgetting when they continually learn as detailed in Sect. 3, a number of approaches have been proposed for *stabilizing* learning based on a belief that straightforward optimization displays too much *plasticity* in this setting.

5.1.1 PARAMETER STORAGE BASED

The most obvious way to prevent catastrophic forgetting across tasks in continual learning is to save an independent model for each task. However, this is sub-optimal for a number of reasons. First of all, it requires a methodology of detecting the current task. It also requires significant storage and limits any ability to leverage relevant knowledge learned across tasks. One way to alternatively leverage knowledge is via shared latent components. Ammar et al. (2014) accomplish this by using a shared latent basis that captures reusable components of the learned policies. Leveraging the assumption of shared common structure between tasks, Borsa et al. (2016) build on the work of Calandriello et al. (2014) and explicitly model a shared abstraction of the state-action space for multi-task settings.

Another approach to leverage knowledge from previous tasks is to provide the representations of networks trained on previous tasks as inputs for subsequent tasks (Rusu et al., 2016). While this strategy directly avoids the problem of catastrophic forgetting, it also exacerbates the curse of input dimensionality and storage requirements. As the number of tasks seen grows, there is a larger space of past representations to sift through. As such, there is still some degradation in the efficacy of transfer in practice even if catastrophic forgetting is circumvented. One way to overcome these issues is by considering a single shared representation. For example, Maurer et al. (2016) extracted features for multiple tasks in a single low-dimensional shared representation. D'Eramo et al. (2020) further highlight the benefits of learning a shared representation as error propagation in approximate value iteration and policy iteration improves when learning multiple tasks jointly.

Alternatively, another popular approach is to store a prior about the extent of past usage of each parameter during learning in order to preserve important old knowledge (Kirkpatrick et al., 2017).

This approach decreases *plasticity* in areas of the network that were heavily used for past tasks, which can effectively prevent forgetting, but this *stability* may also unfortunately limit the potential for backward transfer in the process. A related approach achieves a similar goal by leveraging the concept of superposition where context information is stored for each task so that the weights can be explicitly decomposed into orthogonal sub-networks (Cheung et al., 2019; Wortsman et al., 2020). Minimizing representational overlap across tasks has the positive effect of minimizing forgetting, however, it also may minimize the potential for transfer similarly to the single task learning case. As a result, care must be taken when leveraging this kind of approach in a continual learning context.

5.1.2 Distillation Based

Another common way to encourage retention of knowledge from past tasks is by leveraging knowledge distillation (Buciluă et al., 2006; Hinton et al., 2015) from past tasks when learning a new task to prevent catastrophic forgetting as in (Rusu et al., 2015; Li and Hoiem, 2016; Riemer et al., 2016; Espeholt et al., 2018; Schwarz et al., 2018; Berseth et al., 2018; Kaplanis et al., 2019; Traoré et al., 2019; Tirumala et al., 2019). Knowledge distillation refers to the process of using one neural network as a target or soft target for another. Distillation can be used to augment experiences for training a network by providing a new auxiliary target for our current network being trained to match.

In the context of RL, this target could refer to either a policy or value function. In a continual learning context, distillation is a popular strategy for implementing conservative updates so that the agent is learning to ensure that its learning has *stability* in preserving important knowledge of past tasks whenever possible. An additional benefit of this distillation approach is that it ultimately learns a separate model for each task, which could help address issues of Pareto optimality when an agent must learn conflicting tasks. On the other hand, this in turn results in the need for a knowledge compression strategy for distillation approaches to scale to true many task learning settings.

5.1.3 Rehearsal Based

Another popular strategy for reinforcing the importance of experiences from the past distribution during continual RL is leveraging experience replay (Lin, 1992). As explained in (Pan et al., 2018), experience replay is closely related to model-based RL approaches like Dyna (Sutton, 1991) where a buffer is used to generate realistic samples from the past distribution of experiences. Replay approaches can thus help correct for the short-term bias in our objective function to the extent that the past is a good proxy for the future. As a result, replay has become a very successful approach for tackling continual RL (Isele and Cosgun, 2018; Riemer et al., 2019; Rolnick et al., 2019). However, replay may result in significant storage requirements as we scale to progressively more complex continual learning settings. This has led some to explore replacing replay buffers with *pseudo-rehearsals* sampled from a trained generative model (Robins, 1995; Atkinson et al., 2018).

Another promising strategy for improving the storage efficiency of replay buffers is to directly learn to compress experiences during learning, making the buffer much more efficient though now only consisting of approximate recollections (Riemer et al., 2017; Caccia et al., 2019). While they have been quite successful relative to other approaches to date, replay based strategies generally struggle to effectively leverage past data when the behavior of the current policy is significantly different as off-policy learning tends to be difficult in this setting (see Levine et al. (2020)). Additionally, it may not be necessary to perform replay to combat time correlation during learning (Kaplanis et al., 2018).

5.2 Leveraging Shared Structure

Continually learning agents must learn to solve problems in such a manner that later they are able to find structure in the world (Thrun and O'Sullivan, 1996) and thereby reuse aspects of the solutions to previously solved subproblems through function composition (Griffiths et al., 2019), by abstracting relevant meaningful information in the form of abstract concepts (Parr and Russell, 1998) or skills

(Thrun and Schwartz, 1995). Such an ability is seamless in humans; when performing a complex task, it is natural for us to automatically break the task into smaller subtasks, and be able to plan, learn, and reason with knowledge represented across multiple timescales. Enabling similar abilities in continual RL agents will be crucial for representing knowledge in a way that promotes retention and transfer across the lifetime of an agent. In this section, we will discuss approaches in the literature for leveraging this kind of shared structure.

5.2.1 Modularity and Composition Focused

A key idea explored in the literature is to explicitly formulate a family of tasks as compositionally-structured. The hope is that approaches in this setting can tackle the problem of how to build machines that are capable of compositional generalization i.e. leveraging prior experience to solve compositional perturbations of prior problems or even more complex problems than the agent has seen to this point. In this spirit, early work by Singh (1992) defined a class of composite tasks called sequential decision tasks which are expressed as a temporal concatenation of simpler tasks. Doya et al. (2002) decompose a non-stationary task into multiple domains in space and time to allow for predictable environment dynamics associated with each of the agent's learned modules. Their work further highlights and corroborates findings from the neuroscience literature (see Sec 7.1).

More recent work (Devin et al., 2017; Frans et al., 2017; Fernando et al., 2017; Rosenbaum et al., 2017; Meyerson and Miikkulainen, 2017; Ramachandran and Le, 2018; Chang et al., 2018; Liang et al., 2018; Alet et al., 2018; Cases et al., 2019; Yang et al., 2020) has focused on training neural network modules which can be composed for a family of related tasks, leveraging a combination of modules specialized to each task. This work is deeply related to RL methods for neural architecture search (Zoph and Le, 2016) and their one shot (Brock et al., 2017), multi-task learning (Pasunuru and Bansal, 2019), and continual learning (Xu and Zhu, 2018; Pasunuru and Bansal, 2019) variants.

While these models have increased power of composition and increased leverage in avoiding negative transfer by dividing information between modules, there are a number of challenges that make this idea difficult to implement in a continual learning setting. For example, one such problem is the "chicken and egg" problem of learning modules to combine and learning how to combine them. Here the issue becomes that even if the environment is stationary, the process of both learning modules and learning to combine them become non-stationary as modules are continually updated and leveraged in new combinations. See (Rosenbaum et al., 2019) for a detailed survey of the motivations and challenges of these modular and compositional approaches.

5.2.2 State Abstractions Focused

The tools of abstraction are central to the idea of capturing common structure across tasks and potentially facilitate positive forward transfer for related tasks. Li et al. (2006) provide a unified view on the theory of state abstractions. Given MDP M, with its abstraction version $\hat{M}$, the abstraction function $\phi : S \to \hat{S}$ maps the states in the ground MDP to the states in the abstract MDP. A useful abstraction preserves information that is crucial for the original MDP, or a family of MDPs.

One such abstraction is PAC state abstraction. A PAC state abstraction is defined such that it achieves correct clustering with high probability with respect to a distribution over learning problems (Abel et al., 2018). In the context of lifelong learning, proposed PAC state abstractions, are guaranteed to hold with respect to a distribution of tasks, albeit only in tabular RL settings. With rapid progress in deep RL, recent work (Zhang et al., 2018a; François-Lavet et al., 2019) has focused on building an abstract representation with a low-dimensional representation of relevant features for non-stationary settings. Learning task agnostic state abstractions has also been accomplished by identifying the causal states (Zhang et al., 2019) in POMDPs.

Another piece of important information that can be preserved while learning abstractions is the underlying reward and transition model resulting in a model-irrelevance abstraction (Li et al., 2006). ϕ_{model} is a model irrelevance abstractions if for any action a and any abstract state $\hat{s}$,

$\phi_{model}(s_1) = \phi_{model}(s_2)$ implies $R_{s_1}^a = R_{s_2}^a$ and $\sum_{s' \in \phi_{model}^{-1}(\hat{s})} P_{s_1,s'}^a = \sum_{s' \in \phi_{model}^{-1}(\hat{s})} P_{s_2,s'}^a$. In recent work, Zhang et al. (2020) connect invariant causal prediction to model-irrelevance state abstractions to learn invariant representation in the block MDP (Du et al., 2019b) setting. Specifically, the authors show that such state abstractions generalize to novel observations in the multi-environment setting. Analogous to the multi-task setting, a multi-environment setting is assumed in this work where the agent seeks to maximize return on a set of environments, but only some of which can be trained on. Akin to the idea of exploiting shared structure, in the multi-environment setting, Zhang et al. (2020) assume the presence of shared latent structure among all of the environments.

5.2.3 SKILL FOCUSED

Macro actions (Hauskrecht et al., 1998) or skills (Thrun and Schwartz, 1995) are approaches for learning that side step the requirement to make decisions at each individual time step. A MDP considers decision-making at each time-step. Humans on the other hand are able to plan and execute tasks across multiple time scales. A semi-Markov decision process (SMDP) (Puterman, 1994) provides a generalized framework in which the amount of time between two decision points is modeled as a random variable.

Consider an agent which is in state s and follows a policy π_i in the set of available policies Π. Let's say that the transit time for the agent to enter the next state s' is τ time steps; the state-transition probability from state s to s' could then be expressed as $p(S^\tau = s'|S^0 = s, \pi_i)$. The accumulated discounted reward under the policy π_i would then be denoted by $R_s^{\pi_i}$. The time for which an action persists is either a real or integer value, resulting in the SMDP model being continuous-time discrete event, or discrete-time respectively. The SMDP Bellman equations for an optimal state-value function and action-value function are then given by:

$$v_*(s) = \max_{\pi_i \in \Pi} \left[R_s^{\pi_i} + \sum_{\tau=1}^{\infty} \gamma^{\tau-1} \sum_{s'} p(S^\tau = s'|S^0 = s, \pi_i) v_*(s') \right] \tag{8}$$

$$q_*(s, \pi_i) = R_s^{\pi_i} + \sum_{\tau=1}^{\infty} \gamma^{\tau-1} \sum_{s'} p(S^\tau = s'|S^0 = s, \pi_i) \max_{\pi_i' \in \Pi} q_*(s', \pi_i') \tag{9}$$

Such an abstraction allows the agent to ignore irrelevant details and focus on learning across multiple scales of time and space. A sequence of actions forming a "macro" is one of the simplest kinds of abstraction. A macro can also be obtained as a sequence of other macros, which naturally results in a hierarchy in this architecture. The aim of hierarchical reinforcement learning is to find closed-loop policies at several levels of abstraction, also known as *temporally extended actions*. Temporally extended actions are usually defined over a subset of the state space, with the primary aim being to reduce the number of steps needed for the agent to solve a task. If agents can learn abstractions which are partial solutions to a task that could be reused for other tasks as well, discovering this structure in the world could potentially facilitate faster and more robust learning.

The options framework (Sutton et al., 1999) is a popular choice for temporal abstractions. An option is composed of a policy π, a termination condition $\beta : S \to [0, 1]$, and an initiation set $I \subseteq S$. A *Markov option* $\omega \in \Omega$ has a Markov internal policy and can be represented as a tuple $\langle I_\omega, \pi_\omega, \beta_\omega \rangle$. Options enable an MDP trajectory to be analyzed in either discrete-time transitions or SMDP-style transitions. In every state, a policy over options $\mu : S \times \Omega \to [0, 1]$ selects an option ω according to probability distribution $\mu(S_t, .)$. The option ω in turn determines actions which correspond to the core MDP, until ω terminates in S_{t+k}. Since the primitive action selection in a state S_τ between S_t and S_{t+k} depends not just on that time instance, but also on the option ω being followed at that time, the corresponding flat policy is considered a semi-Markov policy.

While much of the work on learning options has focused on single task learning (Bacon et al., 2017; Machado et al., 2017a,b; Harb et al., 2018; Khetarpal et al., 2020b; Riemer et al., 2020), work

on discovering options in a multi-task context has shown some promising potential both theoretically (Brunskill and Li, 2014) and empirically (Achiam et al., 2018; Riemer et al., 2018; Igl et al., 2019). Brunskill and Li (2014) derive sample complexity bounds for option discovery while learning a distribution over tasks. Mankowitz et al. (2016) proposed a framework to learn near-optional skills for a task and composes skills together to enable efficient multi-task learning.

Skills are an abstract concept that can generally be formalized as a special case of the options framework. In particular, they generally consist of partial policies which are usually intended to reaching critical states. While options explicitly consider where to initiate a skill, what actions to subscribe to, and where to terminate the skill, work that focuses on learning skills generally consider learning a skill policy alone. Additionally, the vast majority of the work on skill learning only implicitly shows *reusability* of skills in application to a low degree of non-stationarity as opposed to explicitly concerning themselves with a full blown continual RL formulation. For instance, Eysenbach et al. (2018) consider a latent-conditioned policy as a skill and aim to learn a diverse set of skills in the absence of rewards to prepare for changing rewards later. With similar motivations, (Campos et al., 2020) propose to learn a set of skills that are state-covering rather than simply diverse. Tessler et al. (2017) instead proposed a hierarchical, multi-skill distillation network explicitly allowing *knowledge retention* and *selective transfer* of skills. The Minecraft domain considered enables study of non-stationarity across both rewards and transitions when agents solve multiple composite tasks.

Researchers have also considered algorithms that combine learned skills, which is one of the core objectives of continual RL. (Sahni et al., 2017; Barreto et al., 2019) have focused on the natural combination of both skill and composition to allow for explicit reuse of previously acquired knowledge in the form of skills. Meanwhile, (Lu et al., 2020) have proposed a skill space planning framework for continual RL environments with no resets. Moreover, explicitly considering state and action abstraction as in (Abel et al., 2017, 2020) has been shown to be quite effective in theory for multi-task RL problems, albeit this theory is limited to tabular MDPs.

5.2.4 Goal Focused

Central to the mission of developing continual RL agents is the acquisition of universal knowledge encompassing a wide variety of tasks. The meaning of a "goal" is open to interpretation and can be formalized as states the agent wants to reach, a reward the agent must achieve, or a termination point of a skill. As such, a popular strategy for decomposing complex problems in RL is to focus on goal-based reasoning. The many goal RL setting is one way to study a non-stationary setting where changing goals are the causal source of non-stationarity in rewards. In some cases, one might even consider non-stationary settings where generalization across goals facilitates changes in both transition dynamics and rewards over time. The earliest work on goal-based RL dates back to Kaelbling (1993). Due to the generic nature of goals in RL, researchers have leveraged them in many different ways. One popular formulation, that is of particular interest in the context of continual RL, is to design algorithms with generalization abilities over goals rather than some other notion of tasks.

An ambitious approach then is to discover general purpose goals without any signal as in unsupervised RL. (Gregor et al., 2016; Hausman et al., 2018; Eysenbach et al., 2018) leverage principles of empowerment and discover goals based on information theoretic objectives, learning to represent the intrinsic control space of an agent in order to facilitate learning about more than a single task. Achiam et al. (2018) relate variational discovery of options to variational auto-encoders and pose it as an optimization problem. Veeriah et al. (2018) propose a deep RL adaptation of the work of Kaelbling (1993) and define a goal as a desired raw-pixel observation, demonstrating empirical benefits in unsupervised learning without a reward signal for improved agent performance on downstream tasks where goals can be treated as auxiliary tasks. Andreas et al. (2017); Oh et al. (2017) assume known information about high-level structural relationships among tasks and similarity between different subtasks respectively to enable fast learning across tasks, whereas, Shu et al. (2017) employ weak supervision from humans to define what skills should be learned in the form of language

instructions. (Florensa et al., 2017; Chen et al., 2019) extract meaningful and transferable goals for a family of related tasks in multi-task RL without any assumptions about prior knowledge.

Using a deep neural network as a value function approximator albeit conditioned on goals (Schaul et al., 2015a) allows reasoning for a multitude of tasks. Leveraging such a universal approximator with off-policy learning from many parallel streams of experience in a continual learning fashion is very beneficial in solving tasks with deep dependencies (Mankowitz et al., 2018). Moreover, Zhu et al. (2019) leverage an unsupervised diversity exploration method to address the problem of catastrophic forgetting by learning task-specific skills similar to Achiam et al. (2018) alongside an adversarial self-correction mechanism to learn knowledge by exploiting past experience. Another important area of research builds off goal conditioned value functions and policies while leveraging the concept of hindsight to learn off-policy about goals that were achieved even if they do not match the original goal of the policy (Andrychowicz et al., 2017; Rauber et al., 2017; Li et al., 2020). This kind of approach has been shown to greatly improve sample efficiency in sparse reward environments. Additionally, the concept of hindsight has been extended to both hierarchical (Levy et al., 2017) and modular (Colas et al., 2019) RL frameworks.

5.2.5 Auxiliary Tasks Focused

One of the vital requirements of a continual RL agent is to learn representations which capture task-agnostic underlying dynamics of the world. A common approach to enable learning about more than a specific task is to augment the loss function with auxiliary losses to provide denser training signals in RL (Li et al., 2015; Lample and Chaplot, 2017). The vast majority of work on this topic has considered hand engineered auxiliary tasks such as inferring the depth map from the RGB observation, detection of loop closures (Mirowski et al., 2016), pixel control (Jaderberg et al., 2016; Hessel et al., 2019), reward prediction (Jaderberg et al., 2016), inverse dynamics prediction Shelhamer et al. (2016), and latent observation prediction (Guo et al., 2020).

A natural research direction that follows is discovering useful and general purpose auxiliary tasks that may ease our agent's ability to learn tasks of interest. Recent work (Veeriah et al., 2019) addresses discovering useful auxiliary tasks in the form of *question functions* of a General Value Function (Sutton et al., 2017). A GVF question is a tuple $\langle \pi, c, \gamma \rangle$, conditioned on environment interaction history $h \in H$, composed of a policy $\pi : H \times A \to [0, \infty)$, cumulant $c : H \to \mathbb{R}$ and termination function $\gamma : H \to [0, 1]$. The answer to a GVF question is defined as the value function, $v : H \to \mathbb{R}$, which gives the expected, cumulative discounted cumulant from any history defined as:

$$v(h_t) = E[c(H_{t+1}) + \gamma(H_{t+1})v(H_{t+1})|H_t = h_t, A_{t+1} \sim \pi(.|h_t)] \qquad (10)$$

Almost none of the aforementioned works explicitly consider continual RL in its entirety and much of the advances in deep RL are yet to be fully explored in a continual non-stationary environments. Moreover, there might be implications of misalignment between auxiliary tasks and the main task at hand (Bellemare et al., 2019). Investigating and incorporating more principled ways of discovering auxiliary learning signals is an interesting direction particularly in the context of continual RL.

5.3 Learning to Learn

In this section, we will highlight applications of learning to learn or meta-learning in settings related to multi-task, lifelong, and continual RL. We will focus on three primary kinds of meta-learning. First, we will discuss approaches that learn about unknown parts of the state space. Next, we will discuss approaches that learn to improve their own ability to adapt and learn. Finally, we will highlight algorithms that learn how to explore and model curiosity driven behavior.

5.3.1 Context Detection

Very little of the non-stationarity people experience in their lives is from literal changes to the physics of the world. In fact, as mentioned in Sec. 2.1, the vast majority of what makes the world non-stationary is the changing behaviors of the many other agents in the world for reasons we don't understand. However, even a stationary multi-agent environment is non-stationary from the perspective of any single agent when the other agents are also constantly learning and adapting. As such, it is common to consider a problem formulation of continual RL in which an underlying MDP is assumed to be stationary, but with partially observable dynamics, namely an influential *task state* that cannot be determined sufficiently from a single observation. Note that this corresponds to the partially observable view of tasks that we mention in Sec. 2.1. This *task state* may have non-stationary characteristics and it is necessary to infer it correctly in order to act optimally based on physical observations. For a real-world example, we can consider the problem of inventory management. Here optimal behavior is dependent on the number of orders at the next step, which can be considered to be based on a not directly observed and constantly changing *task state* i.e. customer demand.

In Figure 6 we highlight a spectrum of assumptions about *task state* evolution in the literature and their connection to standard assumptions in POMDP settings. In this framework, it is generally assumed that the full state s of the environment is composed of a concatenation between a within task physical state that directly produces observations y and an unobserved *task state* z. One popular formulation, called Hidden Parameter MDPs (HiP-MDPs) (Doshi-Velez and Konidaris, 2013) models the environment as a single fixed unknown task that may, for example, be drawn from a stationary distribution over tasks at the beginning of each episode. While it is common to explore settings with a stationary distribution of tasks, there are straightforward assumptions that can be made to help better model non-stationary dynamics over tasks. For example, Hidden Mode MDPs (HM-MDPs) (Choi et al., 2000) consider this formulation for *task states* that evolve independently of the agent's behavior. As seen in Figure 6, the task evolution of a HM-MDP is assumed to follow the behavior of a simple Markov chain $p(z'|z)$.

A more realistic assumption in some cases may be that the tasks change at irregular and extended intervals as in the SMDP formalism (Hadoux et al., 2014b). For example, the recent work of (Xie et al., 2020) considers a formulation with Markovian transitions between tasks which change between episodes but stay constant within an episode. Alternatively, we can consider Mixed Observability MDPs (MOMDPs) (Ong et al., 2010), which are a more general formulation than passive settings where the agent itself can potentially impact the way that the *task state* evolves. In this case, the task evolution follows some stochastic function $p(z'|y, a, z)$.
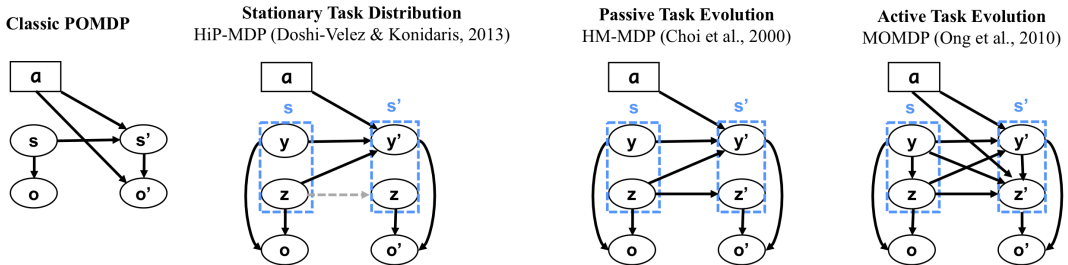

Figure 6: **Useful Classes of Assumptions for POMDPs**: We include from left to right a classic POMDP, a HiP-MDP with stationary *task state* z, a HM-MDP with passive Markovian *task state* evolution, and a MOMDP with an active Markovian *task state* evolution.

Work on context detection (i.e. discovering *task states*) has often assumed access to a change point oracle. This has a very practical benefit as it limits the length of applicable interaction

history. Current work, also typically assumes a stationary *task state* distribution for *meta-learning* (Doshi-Velez and Konidaris, 2013; Wang et al., 2016; Duan et al., 2016; Zintgraf et al., 2018; Rakelly et al., 2019; Zintgraf et al., 2019; Humplik et al., 2019; Fakoor et al., 2019; Ortega et al., 2019; Perez et al., 2020). However, this framework has also has been readily applicable to more challenging multi-agent learning settings (Da Silva et al., 2006; Amato et al., 2013; Marinescu et al., 2017; Vezhnevets et al., 2019). These approaches have even recently been extended to learn encodings for policies themselves based on limited environment interaction (Harb et al., 2020; Raileanu et al., 2020). Overall, context detection is a promising approach for learning about task relatedness, which we are likely to see applied to increasingly challenging continual RL settings in the future.

Changepoint Detection: As discussed previously, one core issue in addressing settings with evolving *task states* is being able to detect change points or boundaries between significant switches without an oracle as in (Da Silva et al., 2006; Rosman and Ramamoorthy, 2012; Hadoux et al., 2014a; Padakandla et al., 2019; Li et al., 2019). However, these approaches generally tend to be reactive to a changing distribution rather than proactive about anticipated changes in the future. This kind of approach generally can converge to the optimal policy eventually for each task, but may not be able to exploit cross task dependencies effectively to improve the sample efficiency of learning.

Bayesian RL: On the other hand, a more ambitious approach would be to try to learn a belief of the unobserved *task state* directly from the history of environment interactions (Li et al., 2009; Hernandez-Leal et al., 2017; Majeed and Hutter, 2018). Bayesian Adaptive MDPs (Martin, 1967; Duff, 2002) take this a step further by directly modeling uncertainty in the space of *task states*. In Bayesian RL we thus generally seek to find a *Bayesian Optimal* policy. This is a policy that acts optimally over time commensurate with its uncertainty about the current task.

5.3.2 Learning to Adapt

A key requirement of continual RL is to acquire new capabilities in a sample efficient manner. Meta-learning is a data driven approach for improving an agent's learning efficiency. The agent attempts to learn to alter its own optimization process based on historical successes and failures of learning. To the extent that these modifications to an agent's learning algorithm generalize into the future, meta-learning should provide an inductive bias to learning that improves an agent's sample efficiency in acquiring new behaviors. Self-modifying policies (Schmidhuber et al., 1998) provide a general framework for meta-learning in continuing RL environments. One early promising example was the Success Story Algorithm (SSA) (Schmidhuber et al., 1997), which leveraged backtracking for improvements that maximize a long-term average reward per step objective. SSA was also successfully applied to handle more complex multi-agent learning settings (Schmidhuber, 1999). Self-modifying policies represent a very ambitious form of meta-learning, which has unfortunately not yet been scaled in application to deep RL with modern neural networks.

On the other hand, there has been significant adoption of a less ambitious form of meta-optimization in the deep RL literature, which relies of a notion of what is called *meta-training* and *meta-testing*. We highlight some of the key details of this framework in Figure 7. As depicted in the left diagram, the recursive process of learning to improve an agent's own learning is decomposed into two separate learning processes called the *inner loop* and *outer loop* (Bengio et al., 1990, 1992; Schmidhuber, 1992). The *inner loop* is responsible for fast time-scale learning and the *outer loop* is responsible for the slower process of learning about learning. For example, in the popular Model Agnostic Meta-Learning (MAML) framework (Finn et al., 2017) the *inner loop* is standard gradient based learning and the *outer loop* computes a gradient to improve the performance of the *inner loop* learning process. The *meta-training* process consists of *meta-updates* where we learn using the *inner loop* learning process on *meta-train training* experiences and based on this learning take an *outer loop* learning step on the network with the purpose of making the *inner loop* more effective on *meta-train testing* experiences. After we have completed *meta-training*, agents are deployed in a *meta-testing* setting. In this setting, we evaluate the improvements to the *inner loop* learning made
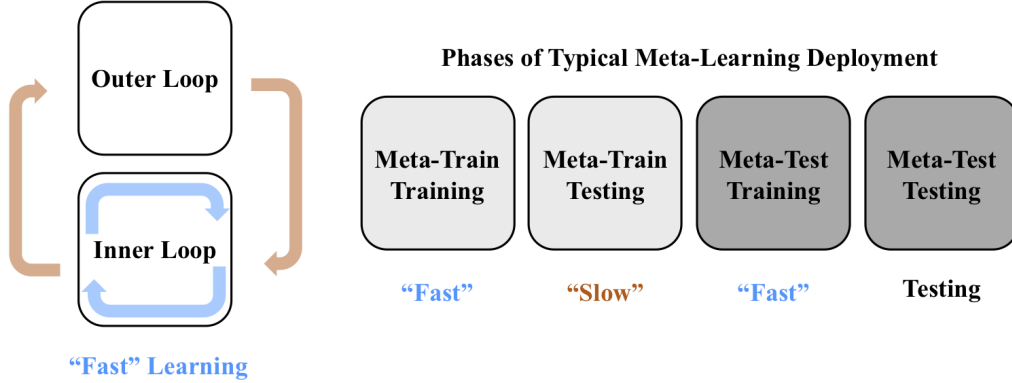
Figure 7: **Phases of Meta-Training & Meta-Testing**. Left: we depict the *inner loop* learning process and the *outer loop* learning process that aims to improve it. Right: we detail the phases of typical meta-learning, including which learning process is used in each phase.

in *meta-training* by providing some new *meta-test training* experiences for *inner loop* learning and comparing performance on held out *meta-test testing* experiences. As such, because of the disconnect between the *meta-training* and *meta-testing* phases of learning, these approaches can optimize for the evaluation objective despite approximating the recursive process of learning to learn by a process with two distinct steps. Indeed, this phased learning setting can even be adopted for continual learning deployment at *meta-test* time by training the *inner loop* continually over a sequence of tasks (Javed and White, 2019; Spigler, 2019; Beaulieu et al., 2020; Caccia et al., 2020; Kim et al., 2020).

There have been a number of recent improvements on this topic of meta-optimization in multi-task RL (Nichol and Schulman, 2018; Kim et al., 2018; Rothfuss et al., 2018; Flennerhag et al., 2018; Nagabandi et al., 2018; Mendonca et al., 2019; Finn et al., 2019; Lin et al., 2020) and multi-agent RL (Al-Shedivat et al., 2017; Foerster et al., 2018a,b; Kim et al., 2020). Moreover, another group of recent approaches focuses on learning a meta-critic, which explicitly guides updates to the agent's policy rather than simply guiding its actions (Sung et al., 2017; Xu et al., 2019; Harb et al., 2020). While most of this work has been confined to the *meta-training* and *meta-testing* setting, some recent approaches have taken a first step towards applying these ideas of meta-optimization (Riemer et al., 2019) and meta-critics (Zhou et al., 2020) in a true continual RL setting closer to the spirit of self-modifying policies. For example, an approach by (Chandak et al., 2020b) leverages the approximate performance of a proposed policy on the time series of past episodes to forecast its expected performance into the future and improve an agent's ability to adapt in the presence of smooth and passive non-stationarity. Additionally, a related set of approaches leverage the idea of online cross-validation (Sutton, 1992), which can be deployed in episodic environments leveraging successive episodes for *meta-training* and *meta-testing* (Xu et al., 2018b; Zahavy et al., 2020; Xu et al., 2020). As these meta-learning techniques become less and less reliant on leveraging explicit training phases to segment learning, meta-learning should be able to make an impact for an even wider variety of continual RL use cases moving forward.

5.3.3 Learning to Explore

Another core problem of continual RL that meta-learning can potentially help with is exploration. Meta-learning has been repeatedly used in the recent literature to learn functions for intrinsic motivation and improved exploration (Baranes and Oudeyer, 2009; Zheng et al., 2018; Stadie et al., 2018; Xu et al., 2018a; Houthooft et al., 2018; Yang et al., 2019; Zou et al., 2019; Zheng et al., 2019).

There are also a number of successful approaches that consider the concept of artificial curiosity. This is generally achieved by defining a heuristic based on some measurement of surprise, information gain, compression, or empowerment (Schmidhuber, 1991; Klyubin et al., 2005; Kaplan and Oudeyer, 2006; Schmidhuber, 2008; Frank et al., 2014; Mohamed and Rezende, 2015; Bellemare et al., 2016; Houthooft et al., 2016; Pathak et al., 2017; Karl et al., 2017; Achiam and Sastry, 2017; Burda et al., 2018; Shyam et al., 2019; Liu et al., 2019; Ecoffet et al., 2019; Sekar et al., 2020).

These approaches can even be useful for learning in a single stationary environment, especially one with sparse rewards. However, there is a related curiosity problem unique to continual RL when you have some level of agency over the order in which tasks are encountered. For example, the PowerPlay framework (Schmidhuber, 2013) is a methodology for deciding on the optimal task to solve next in a lifelong learning setting. There has been some limited recent work on deciding on the next task to learn on in deep multi-task RL (Sharma et al., 2017) and on learning to illicit valuable information to learn from in non-stationary multi-agent settings (Shu et al., 2018). However, work that considers an active and agent driven setting for exploring tasks in continual RL still remains scarce to date. This will be an interesting area to keep an eye on as it is a natural fit for RL and may be an exciting focus for future research.

6. On Evaluation of Continual RL Agents: Benchmarks & Metrics

The proper procedure and protocol for evaluating continually learning agents remains as an open research question. While there has been tremendous progress in the field chasing the state-of-art on widely acknowledged benchmarks (Bellemare et al., 2013; Brockman et al., 2016) and achieving super human performance (Silver et al., 2018), it is not immediately clear if the aforementioned benchmarks have the sufficient characteristics of desired environments for lifelong RL agents. Schaul et al. (2018) discuss a long list of open problems in continual RL. We here expand on their categorization of *benchmarks* and *metrics* presenting a discussion on the evaluation of continual RL agents.

Benchmarks. Arguably, one of the primary roadblocks in the study and rapid progress of continual RL has been the lack of well suited environments to evaluate agents in this setting. Existing and widely used benchmarks such as classical small MDP environments (Mountain Car, Cartpole, Taxi), OpenAI's Gym (Brockman et al., 2016), and the Arcade Learning Environment (ALE) (Bellemare et al., 2013) have been instrumental to the progress made in reinforcement learning over the years.

Specific environments have provided specific dimensions to study and measure agents on. For instance, Taxi (Dietterich, 2000) and fourooms (Sutton et al., 1999) have exploitable structure and have been thoroughly used to study abstractions. Similarly Mujoco and the DM-control suite were designed for continuous control, ALE for deep RL with image processing, and DeepMind Lab (Beattie et al., 2016) for rich 3D navigation tasks. For more complex agents which can play long-horizon, strategy games accompanied with rich visual observations, several algorithms have also explored Minecraft (Duncan, 2011) and VizDoom (Kempka et al., 2016).

Much of the work related to continual RL has hand engineered customization to the aforementioned environments to facilitate measurements for continual learning performance. In fact, often times, researchers studying continual RL end up designing tasks suitable for the specific questions they are trying to address. This might result in inherent bias in the design of experiments and tasks, which can potentially lead to unintended consequences. For instance, due to inherent determinism in some of these domains (e.g ALE), agents have been shown to often resort to memorization of state-action sequences as opposed to achieving true generalization capabilities (Machado et al., 2018).

Generating non-stationarity in environments in a principled way is vital for understanding the changes in learning over time. Henderson et al. (2017a) takes a step in this direction for multi-task RL and proposed a benchmark that parameterizes different variants of OpenAI Gym tasks, making it easier to generate novel unseen variants by modifying specific internal environment parameters in order to capture variation in transition and reward dynamics. Additionally, in recent years, a lot of

progress has been made on developing benchmarks to train and test RL agents to better highlight generalization abilities. Benchmarks such as Coinrun (Cobbe et al., 2019b) which is part of the larger Procgen (Cobbe et al., 2019a) suite of games leverage procedural generation to create a large set of train and test environments with subtle differences. Additionally, Osband et al. (2019) proposed bsuite, a framework with a simple set of environments aimed at specifically measuring different capabilities of an RL algorithm such as exploration, credit-assignment, and memory. Finally, Yu et al. (2020b) proposed Meta-world, a benchmark of 50 unique continuous control tasks for training and testing RL agents where each task can be associated with a number of environment configurations.

The recent advancements of RL algorithms is in large parts due to these environments with an increased focus on testing generalization. However, most environments require a clear distinction between train-test boundaries, dependence on a well defined notion of task, often structured across episodic settings, with limitations on how and when non-stationarity is introduced. As a result, there is still a significant need to make standard benchmarks that are tailor made for continual RL specifically. In particular, we need benchmarks that provide rich streams of data that are configurable for varying degrees of complexity.

A potentially promising direction would be to develop continual RL domains that allow for a range of non-stationary settings as discussed in our taxonomy of formalism (see Sec. 4). Additionally, desirable characteristics of such benchmarks would include the ability to: 1) train in a progressive incremental fashion, 2) facilitate discovery and composition of skills, 3) facilitate understanding of real-world dynamics such as physical rules governing the world, and 4) learning causal relationships including affordances associated with objects. Ahmed et al. (2020) have recently proposed Causal-World, which is a promising benchmark fulfilling many of these criteria in application to robotic manipulation tasks. Furthermore, research suggests that human intelligence is grounded in learning via embodiment (Kiefer and Trumpp, 2012; Kiela et al., 2016; Lake et al., 2017; Bisk et al., 2020). Embodied experience grounds learning in intuitive physics and causal reasoning (Lake et al., 2017). Khetarpal et al. (2018b) discussed similar recommendations for moving towards environments for lifelong agents and also highlighted the benefits of embodied cognition.

Recently, Jelly Bean World (Platanios et al., 2020) was introduced as a test-bed for never-ending learning. In particular, Jelly Bean World supports a variety of non-stationary environment configurations including multi-task, multi-agent, multi-modal, and curriculum learning settings. Although designed for never ending learning, the framework has considerable overlap with lifelong/continual RL, making it suitable for evaluating continual learning in its purest form.[2] In the context of non-stationarity in real-world applications, researchers have also explored large scale recommender systems (Chandak et al., 2019) and diabetes treatment (Chandak et al., 2020a) as case studies where the nature of RL problems have inbuilt non-stationarity to account for.

Metrics. The longstanding tradition in reinforcement learning has been to measure an agent's performance by recording its average expected accumulated rewards over time (Sutton and Barto, 1998). While performance is a good quantitative measure of an agent's performance, measuring expected accumulated returns by the agent may not suffice for fully understanding the abilities of a continually learning agent. The choice of metrics in RL is often tightly coupled with the choice of the problem formulation. Consider an RL algorithm designed to solve many sequential tasks which are related to each other through proper skill discovery (see Sec. 5.2.3). While accumulated returns is a strong primary indicator of learning performance, it does not give us any insight into the robustness of skills learned, or if the skills learned in one task are leveraged across the lifetime of an agent, etc. Furthermore, for a continual learner, it is vital to measure core metrics such as forward and backward transfer (or interference/forgetting), skill reusability, and skill composition to name a few.

---

2. Never ending learning as posed in the Jelly Bean World environment suggests that the notion of a task or subtask naturally emerges and the distinction between tasks is not always as sharp as in literature on lifelong learning or continual learning. Never-ending learning can also be formulated withing our taxonomy of formalism where the nature of the non-stationarity allows these lines to be blurry.

A promising approach is to incorporate the idea of *auxiliary metrics* (Schaul et al., 2018) to measure an agent's intelligence via *probe questions* akin to auxiliary losses. Incorporating auxiliary tasks (see Sec. 5.2.5) has been shown to improve the representations learned by an agent. Moreover, auxiliary evaluation metrics can further our understanding of an agent's abilities and overall intelligence. Specifically, the core desired capabilities of a continual learner can be tested with probe questions.

Towards Broader Evaluation Criteria for Continual RL. At the intersection of metrics (what to measure) and domains (how to measure)[3] we recommend that bsuite (Osband et al., 2019) is a promising example of the type of framework needed for training and evaluating agents in a continual fashion. In Figure 8 we highlight some important evaluation criteria to consider to better understand the performance of continual RL agents. For a given degree and nature of non-stationarity (see Sec. 4), researchers should generate a set of carefully designed experiments with a carefully chosen complexity to train and evaluate the continual RL agent on. Ideally, proper empirical analysis would result in a measure of the behavior along different dimensions of probe-metrics as shown in the Fig. 8.
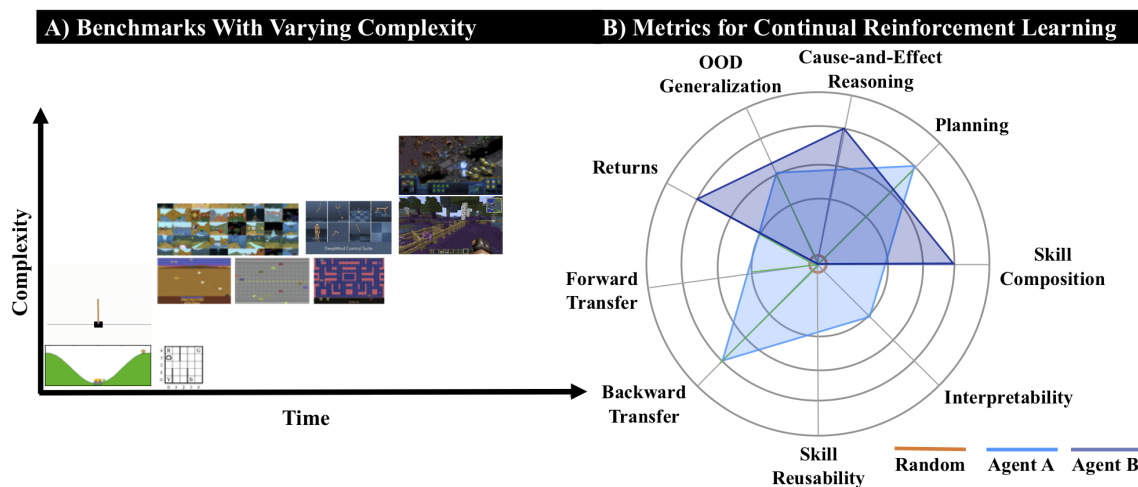

Figure 8: **Evaluating Continual Reinforcement Learning Agents**. A) Depicts the evolution of domains and benchmarks over time commonly used in RL. B) Depicts key metrics for evaluating continual RL agents in the style of bsuite. Such a framework should also offer a knob controlling the degree and nature of non-stationarity that agents experience (see Figure 4). For a given degree of non-stationarity, a set of carefully designed experiments to test different *probe questions* would help foster more rapid progress in the field.

To this end, it is important to consider the following capabilities as probe questions (i.e. auxiliary metrics) in addition to measuring the accumulated returns over time.

1. *Catastrophic Forgetting (Forward and Backward Transfer)*: It is desired for our agents to be able to effectively use previously acquired knowledge in new related situations that they might encounter (i.e. forward transfer). Moreover, if an agent has seen a situation before and encounters a similar experience, it should be able to perform backward transfer to improve its previously learned capabilities. When an agent's current learning greatly interferes with its ability on previous experiences, the agent is experiencing catastrophic forgetting.

---

3. We acknowledge that evaluation of deep RL agents faces several challenges pertaining to reproducibility (see Henderson et al. (2017b); Khetarpal et al. (2018a)). This is even more reason for the community to move towards standardized evaluation benchmarks for continual RL.

2. *Skill Reusability*: Just as humans acquire skills and build on them to solve increasingly complex tasks, a continual learning agent must be able to reuse previously learned skills in new unseen situations. This is an important ability especially when new skills can be created on-the-fly in new situations. Measuring skill reusability is challenging because this might include adjustments and tweaks to the previously learned representations or explicit skills (options). Much of the existing work uses qualitative analysis as a tool to measure reusability or adaptability. See Sec. 5.2.3 for a detailed discussion on skill focused approaches.

3. *Qualitative Analysis (Interpretability)*: Understanding and interpreting different aspects of behaviors through qualitative analysis is undervalued in the field while the heavy emphasis is on improving scores and learning curves. We posit that qualitative analysis that provides clarity about the type of representations learned, the kind of behaviors acquired, the landscape of the value functions and policy, the changes made to previously learned knowledge, and model predictions will all be key to furthering our understanding of continual RL methods.

4. *Skill Composition*: A common desiderata from continual RL agents is to effectively leverage shared structure over the data that an agent sees. See Sec. 5.2 for a detailed discussion. Composing previously learned behaviors to perform new ones is an important capability to consider as this enables agents to exploit what was learned before with greater efficacy.

5. *Planning*: A lifelong learning agent must be able to effectively plan for the future by leveraging its acquired knowledge. Planning can be explicitly measured by evaluating an agent's explicit or implicit plans over time on different tasks. For instance, if an agent is building models of the world, measuring the value function through approximate dynamic programming can give an estimate of how well the agent can plan. Such a procedure should be adaptable for planning in both observation spaces and latent representation spaces including single time-step and multi-time-steps extrapolation.

6. *Cause and Effect Reasoning*: To ensure that continual RL agents are indeed learning underlying rules of an environment, and objects in that environment, designing probes for cause-and-effect analysis will allow a quantitative measure of how well this ability is acquired by agents. One concrete way to measure this is to design object-oriented probes associated with well grounded perturbations on objects to model *interventions* and test an agent's causal understanding. For example, this kind of evaluation is possible leveraging CausalWorld (Ahmed et al., 2020).

7. *Out of Distribution (OOD) Generalization*: Another probe-metric to test agents on over the course of their lifetime would be to measure an agent's performance when we situate it in environments that lie outside of its prior training distribution. In these cases, it would be interesting to evaluate an agent's generalization performance by measuring both zero-shot expected returns and its sample complexity when learning new capabilities.

7. Looking Forward

In this section we conclude our survey by looking forward at the frontiers of continual RL research. Specifically, we will first discuss potential connections between continual RL research and findings in the neuroscience community. We will then discuss challenges and open problems that continual RL research will have to address in order to make progress.

7.1 Bridging the Gap Between Continual RL and Neuroscience

Findings in the neuroscience literature have often served as a guiding light towards developing more human-like continual learners. How humans have the innate ability to perform long-term decision making without a task reward or rather even delayed spare rewards remains largely a mystery. In

fact, it is unclear even what utility (Samuelson, 1937) humans optimize for. Nonetheless, there has still been a good deal of interesting work at the intersection of neuroscience and reinforcement learning. For example, Hassabis et al. (2017) provides a survey on neuroscience inspired artificial intelligence, and Niv (2009) details evidence that reinforcement learning happens in the human brain. We will now briefly discuss some areas of interest to continual RL where it may be potentially useful to draw insights from research on the human mind.

Balancing Stability and Plasticity. The tension between prioritizing recent experiences and past experiences, also known as the *stability-plasticity dilemma*, is often encountered when training neural networks and has also been demonstrated in the human brain (Carpenter and Grossberg, 1987). This additionally closely ties in with the inherent problem of credit-assignment in RL. Humans have somewhat seamless mechanisms in place to assign credit even for events which happen far after the relevant actions that caused them. Inspired by these findings in neuroscience, the recent work of Hung et al. (2019) introduces a temporal value transport algorithm where the agent uses specific memories to credit past actions.

Moreover, the neuroscience literature can give us more insight into the nature of task interference that humans experience. For example, Detre et al. (2013) demonstrates that human memories that are highly activated during continual learning are less likely to be forgotten later. Additionally, Sagiv et al. (2020) theorize that the inability of humans to effectively perform multiple tasks at the same time results from a trade-off related to the human tendency to share network architectures between tasks in order to enable quicker learning. While AI agents suffer a great deal from *catastrophic forgetting*, evidence in the literature (McClelland et al., 1995) suggests that the mammalian brain may avoid catastrophic forgetting by protecting previously acquired knowledge in neocortical circuits through interaction with the hippocampus. Leveraging these findings (Kirkpatrick et al., 2017) discuss how continual learning in the neocortex relies on task-specific synaptic consolidation, whereby knowledge is durably encoded by rendering a proportion of synapses less plastic and therefore stable over long timescales. Finally, the work of Ajemian et al. (2013) has demonstrated the role that noisy computations in the human brain may have in avoiding catastrophic forgetting.

Nature of Human Rewards. One place where inspiration from neuroscience may be particularly helpful for designing AI systems is in understanding more about the origins of reward. It has been suggested that an average reward per step objective is more consistent with human studies than the discounted cumulative reward objective that has become more popular for RL research (Daw and Touretzky, 2000). There is also evidence that humans prefer to acquire more knowledge even when it lacks predictive value (Niv and Chan, 2011). Indeed, there has been a significant body of research trying to understand the role of the dopamine system as a reward signal for humans (Samson et al., 2010; Starkweather et al., 2017). It has also been theorized that the dopamine system is directly responsible for a kind of meta-learning by training the prefrontal cortex, which then can function as its own standalone learning system (Wang et al., 2018). Moreover, it has been demonstrated that stress can play a big factor in selectively modulating the reward signal for humans (Berghorst et al., 2013). Finally, it has been recently suggested that many sophisticated aspects of human learning cannot be explained by RL, and that these aspects of behavior may be supported by the brain's executive functions (Rmus et al., 2020).

Leveraging Memory. Another important aspect of continual RL that perspectives from the neuroscience literature can help provide direction for is with respect to the interplay of learning and memory formation. For example, Collins and Frank (2012) study how RL and working memory complement each other in the human brain. Additionally, there is strong evidence in the literature of hippocampal replay facilitating the model-based planning process for human RL (Mattar and Daw, 2018; Momennejad et al., 2018; Vikbladh et al., 2019; Momennejad, 2020). Interestingly, the close connection between experience replay and planning was also recently highlighted in the RL literature (Pan et al., 2018; Eysenbach et al., 2019). Moreover, it has been suggested that humans perform a form

of pseudo-replay or pseudo-rehearsals that is likely particularly active and useful for consolidation of knowledge during human sleep (Robins, 1995, 1996; Frean and Robins, 1999). Schuck and Niv (2019) also suggests that hippocampal replay may be important for building representations of complex, abstract tasks elsewhere in the brain. Likewise, it has been suggested that the hippocampus may be central to the superior value generalization capabilities that humans demonstrate (Wimmer et al., 2012). In fact, there is evidence that it also may be critical for computing value over complex state spaces, learning with little data, and performing long-term credit assignment (Gershman, 2017). Finally, it has been demonstrated that reward prediction errors also play a key role in the memory forming process (Jang et al., 2019), which also mirrors findings in deep RL (Schaul et al., 2015b).

Balancing Model-Based and Model-Free Learning: There has also been significant research on the interplay between model-free and model-based learning in the human brain that could provide guidance in designing sample efficient continual RL agents. Research suggests that a combination of model-free and model-based computations are performed in the human brain (Gläscher et al., 2010; Daw et al., 2011; Doll et al., 2012). In fact, Langdon et al. (2018) even highlight the influence of model-based computations rather than only model-free ones in dopamine responses. It is believed that humans primarily rely on model-based learning for dealing with tasks that have both high volatility and low noise (Simon and Daw, 2011). Additionally, model-based reasoning has been shown to prevent humans from forming habits (Gillan et al., 2015). Moreover, recent work has demonstrated that the successor representation model from the RL literature (Dayan, 1993) may provide a more accurate model to reflect the way humans balance their use of both model-based and model-free learning (Momennejad et al., 2017).

Exploiting Modular Structure.   Doya et al. (2002) proposed a modular control architecture using multiple prediction models based on the computational model of the cerebellum proposed by Wolpert et al. (1998). Their simulation results corroborate findings associated with fMRI data (Imamizu, 1997; Imamizu et al., 2000) suggesting that when a new task is introduced, many modules initially compete to learn it. However, after repetitive learning, only a subset of modules are specialized and recruited for the new task. Related work by (Daw et al., 2005) proposes a Bayesian uncertainty based model for arbitrating competition among important subsystems of the human brain. Indeed, studies of the human brain are a fruitful reference point for understanding what may be good paradigms for exploiting a modular architecture in the context of continual RL.

Results In Line With Current Trends.   Finally, there is a growing body of evidence that justifies many currently popular trends in the continual RL literature based on human comparisons. For example, there is significant evidence in the literature that human reinforcement learning is either uncertainty aware or Bayesian in nature (Niv et al., 2005; Fleming and Daw, 2017; Babayan et al., 2018; Lowet et al., 2020). There is also significant evidence that substantial hierarchical structure is present in the human reinforcement learning process (Botvinick et al., 2009; Gershman et al., 2009; Frank and Badre, 2012; Badre and Frank, 2012; Diuk et al., 2013; Botvinick and Weinstein, 2014; Solway et al., 2014; Ribas-Fernandes et al., 2019). Moreover, (Niv et al., 2015) highlights the role of attention in the human brain in supporting the reinforcement learning process to address the curse of dimensionality. Interesting work by (Niv, 2019) also suggests that the orbitofrontal cortex may be used to represent *task-states* that are deployed for decision making and learning elsewhere in the brain. Additionally, it has been demonstrated that human reinforcement learning attempts to identify the causal structure in the task at hand (Gershman et al., 2015; Gershman, 2017).

7.2 Challenges and Open Problems

As we have repeatedly seen in this review, due to the lack of a concrete definition, and the extremely general nature of continual reinforcement learning, it still has a lot of unexplored potential. Moreover,

there remain a number of open problems which come with their own challenges. In this section, we will discuss some of these fundamental challenges in more detail.[4]

Inductive Biases. There are without doubt many different perspectives within the community of AI researchers on how much and to what degree we should embed inductive biases in our agents even with regard to the quest of achieving general intelligence. However, leveraging priors from the human learning process has enabled the field to make significant progress over time. Many real-world applications such as robot navigation and autonomous driving are tangible now only as a result of making assumptions around such priors.

In the context of a continual RL agent, this remains an open question and of greater significance. What utility an agent optimizes for over its lifetime is not immediately clear even for human learning. Sutton (2019) argues that the most promise lies in leveraging computation, as opposed to leveraging human knowledge and inductive biases coming from this human knowledge. "*We have to learn the bitter lesson that building in how we think we think does not work in the long run.*" (Sutton, 2019). To this end, potential directions aligned with this process of discovery through computation include: self-tuning approaches (Xu et al., 2018b; Zahavy et al., 2020), end-to-end skill discovery (Bacon et al., 2017), discovering RL algorithms (Oh et al., 2020), learning what objective to learn (Xu et al., 2020), and other approaches moving more and more towards open-ended learning (Wang et al., 2020).

Task Specification. Expressing *task specification*, which often carries subtle assumptions as to how tasks are related, is often left to the interpretation of the designer. Although a task can be specified as its own MDP, this definition is very broad and thus limited in that sense. To overcome the extremely general nature of this definition, researchers have proposed several MDP variations of relevance to continual RL such as HM-MDPs, MoMDPs, Block-MDPs, HiP-MDPs, and Factored MDPs to name a few. Moreover, where tasks and rewards actually come from is another open and somewhat philosophical problem that is a long-standing question for researchers. Indeed, hand engineering of rewards and therefore tasks has been under scrutiny for decades without much progress in making agents less dependent on this level of human intervention.

The Agent-Environment Boundary. Traditionally MDPs serve as a framework for formalizing agent-environment interaction and the corresponding boundary between the two. Considering the role of other agents in the learning of real-world decision makers, the correct way to view the agent-environment boundary remains an open question of high significance. Jiang (2019) and Harutyunyan (2020) offer fresh perspectives on this discussion. The taxonomy presented in this work mostly included literature that is concerned with the traditional view of a single decision making agent with everything else delineated as the environment. On the other hand, multi-agent formulations explicitly consider the presence of other agents and model learning with this alternative view. In the context of non-stationarity, the agent-environment boundary is an open concept because the space of affordances (Khetarpal et al., 2020a) might evolve over time (Pezzulo and Cisek, 2016) and may only emerge as a result of agent-environment interaction (Gibson, 1977; Heft, 1989; Chemero, 2003).

Experiment Design and Evaluation. As we discussed earlier, studying the full non-stationary setting in its entirety is a significant challenge for continually learning agents. Designing and engineering such experiments is not only laborious, but also, subject to the scale at which one can study such a setting. We have discussed possible benchmarks with a unification of domains and metrics (see Sec. 6) to generate systematic experiments for training continual RL agents. However, another related open question in continual RL has to do with the nature of evaluation itself. Indeed, it seems somewhat unnatural to have separate validation or testing phases in a true continual RL setting. In some sense, the notion of separate phases for evaluation that prohibit learning can be seen as idealistic as it is not generally possible to really test humans in such a way. Moreover, disconnects

---

4. We note that here we only touch upon some open problems and refer the reader to Schaul et al. (2018) for more discussion on open problems in continual RL. We plan to revisit this discussion and incorporate more detailed discussion on those as well in later revisions of this draft.

between training and testing settings can be problematic for optimization in the context of continual RL. For example, it is popular to test continual learning agents in the so called *one pass* setting where an agent is trained on each task in succession while never revisiting old tasks and then tested on its retention over the distribution of old tasks. This setting would be problematic for any continual RL agent without a pre-specified prior towards remembering old tasks. This is because a purely data driven agent has seen no evidence that past tasks will reoccur and thus has no reason to maintain performance on these tasks that is conveyed to it through its rewards during training. As a result, even basic questions about how to evaluate agents can be quite tricky for continual RL and can be considered a challenge that the community will still have to understand more deeply moving forward.

Interpreting Discovery. The advances in image classification tasks with the advent of large scale labeled datasets such as Imagenet (Krizhevsky et al., 2012; Russakovsky et al., 2015) was followed by a plethora of research on visualizing (Zeiler and Fergus, 2014) and understanding (Szegedy et al., 2013) the *what* and *hows* of the deep convolutional neural networks. A large body of research in continual RL which address *the discovery problem* makes an overarching promise for a certain kind of solution (e.g. those that are general and adaptive), but most of these claims are really hard to verify. A natural technique researchers adopt is qualitative analysis which is often subject to interpretation. To advance research towards the overarching goal of artificial general intelligence, we need to develop tools to understand and introspect our agents based on more than just rewards.

Learning at Scale. Much of the recent revolution within machine learning and in deep learning in particular is in large part due to advancements in the hardware Hooker (2020). It is arguable that continual RL agents might overcome potential issues such as forgetting if learning is performed at scale. Recently, we have seen similar findings in the natural language processing community where language models such as GPT-3 (Brown et al., 2020) are able to perform much better due to learning at scale. This is not to say that throwing a lot of data and compute will solve everything, and is the one solution we need. However, when connections to evolutionary learning in the human brain are made, we do need to really appreciate the scale at which such learning takes place. Recent work has looked at the topic of *scaling laws* in deep supervised learning (Kaplan et al., 2020; Sharma and Kaplan, 2020; Henighan et al., 2020). Exploring the relevant *scaling laws* for continual RL will be a potentially fruitful topic to understand more clearly moving forward.

8. Conclusion

To conclude this review, we would like to highlight some final considerations. While finding the right set of assumptions is an important meta challenge of continual reinforcement learning, we would like to make fewer assumptions in an effort to push the continual RL research community towards increasingly realistic settings. Findings about the human brain, psychology, and cognitive behaviors, including animal learning, have laid strong foundations for the field of reinforcement learning. Sec. 7.1 further highlights that perhaps bridging the gap between AI and computational neuroscience has promising potential to help make rapid advancements in the field of continual RL. We should aim to address these open problems of continual RL with the aspiration to deploy agents in challenging real-world applications where continual RL has potentially promising use cases. Supervised continual learning has seen some success in the same spirit; Carlson et al. (2010) is one such successful example deployed in a true never ending continual learning fashion, albeit in a very controlled setting. While largely a theoretical field to date, continual RL marks a shift towards a more robust style of learning that, when successful, will greatly broaden the applicability of RL for real-world use cases. We hope that this survey can serve as a reference point for the continual RL research community to build off that will help us collectively understand and eventually achieve this lofty goal.

9. Acknowledgements

References

D. Abel, D. Arumugam, L. Lehnert, and M. L. Littman. Toward good abstractions for lifelong learning. In *NIPS Workshop on Hierarchical Reinforcement Learning*, 2017.

D. Abel, D. Arumugam, L. Lehnert, and M. L. Littman. State abstractions for lifelong reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2018.

D. Abel, N. Umbanhowar, K. Khetarpal, D. Arumugam, D. Precup, and M. Littman. Value preserving state-action abstractions. In *International Conference on Artificial Intelligence and Statistics*, pages 1639–1650. PMLR, 2020.

J. Achiam and S. Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*, 2017.

J. Achiam, H. Edwards, D. Amodei, and P. Abbeel. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*, 2018.

O. Ahmed, F. Träuble, A. Goyal, A. Neitz, M. Wüthrich, Y. Bengio, B. Schölkopf, and S. Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

R. Ajemian, A. D'Ausilio, H. Moorman, and E. Bizzi. A theory for how sensorimotor skills are learned and retained in noisy and nonstationary neural circuits. *Proceedings of the National Academy of Sciences*, 110(52):E5078–E5087, 2013.

M. Al-Shedivat, T. Bansal, Y. Burda, I. Sutskever, I. Mordatch, and P. Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. *arXiv preprint arXiv:1710.03641*, 2017.

F. Alet, T. Lozano-Pérez, and L. P. Kaelbling. Modular meta-learning. *arXiv preprint arXiv:1806.10166*, 2018.

C. Amato, F. A. Oliehoek, and E. Shyu. Scalable bayesian reinforcement learning for multiagent pomdps. Citeseer, 2013.

H. B. Ammar, E. Eaton, P. Ruvolo, and M. Taylor. Online multi-task learning for policy gradient methods. In *International Conference on Machine Learning*, pages 1206–1214, 2014.

J. Andreas, D. Klein, and S. Levine. Modular multitask reinforcement learning with policy sketches. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 166–175. JMLR. org, 2017.

M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 5048–5058, 2017.

C. Atkinson, B. McCane, L. Szymanski, and A. Robins. Pseudo-rehearsal: Achieving deep reinforcement learning without catastrophic forgetting. *arXiv preprint arXiv:1812.02464*, 2018.

B. M. Babayan, N. Uchida, and S. J. Gershman. Belief state representation in the dopamine system. *Nature communications*, 9(1):1–10, 2018.

P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

D. Badre and M. J. Frank. Mechanisms of hierarchical reinforcement learning in cortico–striatal circuits 2: Evidence from fmri. *Cerebral cortex*, 22(3):527–536, 2012.

A. Baranes and P.-Y. Oudeyer. R-iac: Robust intrinsically motivated exploration and active learning. *IEEE Transactions on Autonomous Mental Development*, 1(3):155–169, 2009.

A. Barreto, D. Borsa, S. Hou, G. Comanici, E. Aygün, P. Hamel, D. Toyama, S. Mourad, D. Silver, D. Precup, et al. The option keyboard: Combining skills in reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 13031–13041, 2019.

A. G. Barto. Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems*, pages 17–47. Springer, 2013.

A. G. Barto and T. G. Dietterich. Reinforcement learning and its relationship to supervised learning. *Handbook of learning and approximate dynamic programming*, 10:9780470544785, 2004.

C. Beattie, J. Z. Leibo, D. Teplyashin, T. Ward, M. Wainwright, H. Küttler, A. Lefrancq, S. Green, V. Valdés, A. Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016.

S. Beaulieu, L. Frati, T. Miconi, J. Lehman, K. O. Stanley, J. Clune, and N. Cheney. Learning to continually learn. *arXiv preprint arXiv:2002.09571*, 2020. URL https://arxiv.org/abs/2002.09571.

M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in neural information processing systems*, pages 1471–1479, 2016.

M. Bellemare, W. Dabney, R. Dadashi, A. A. Taiga, P. S. Castro, N. Le Roux, D. Schuurmans, T. Lattimore, and C. Lyle. A geometric perspective on optimal representations for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4360–4371, 2019.

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

E. Bengio, J. Pineau, and D. Precup. Interference and generalization in temporal difference learning. *arXiv preprint arXiv:2003.06350*, 2020.

S. Bengio, Y. Bengio, J. Cloutier, and J. Gecsei. On the optimization of a synaptic learning rule. In *Preprints Conf. Optimality in Artificial and Biological Neural Networks*, volume 2. Univ. of Texas, 1992.

Y. Bengio, S. Bengio, and J. Cloutier. *Learning a synaptic learning rule*. Citeseer, 1990.

L. H. Berghorst, R. Bogdan, M. J. Frank, and D. A. Pizzagalli. Acute stress selectively reduces reward sensitivity. *Frontiers in human neuroscience*, 7:133, 2013.

G. Berseth, C. Xie, P. Cernek, and M. Van de Panne. Progressive reinforcement learning with distillation for multi-skilled motion control. *arXiv preprint arXiv:1802.04765*, 2018.

Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

D. Borsa, T. Graepel, and J. Shawe-Taylor. Learning shared representations in multi-task reinforcement learning. *arXiv preprint arXiv:1603.02041*, 2016.

M. Botvinick and A. Weinstein. Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130480, 2014.

M. M. Botvinick, Y. Niv, and A. G. Barto. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, 113(3):262–280, 2009.

A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Smash: one-shot model architecture search through hypernetworks. *arXiv preprint arXiv:1708.05344*, 2017.

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

E. Brunskill and L. Li. Pac-inspired option discovery in lifelong reinforcement learning. In *International Conference on Machine Learning*, pages 316–324, 2014.

C. Buciluă, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541. ACM, 2006.

Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

L. Caccia, E. Belilovsky, M. Caccia, and J. Pineau. Online learned continual compression with stacked quantization module. *arXiv preprint arXiv:1911.08019*, 2019.

M. Caccia, P. Rodriguez, O. Ostapenko, F. Normandin, M. Lin, L. Caccia, I. Laradji, I. Rish, A. Lacoste, D. Vazquez, et al. Online fast adaptation and knowledge accumulation: a new approach to continual learning. *Advances in Neural Information Processing Systems*, 2020.

D. Calandriello, A. Lazaric, and M. Restelli. Sparse multi-task reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 819–827, 2014.

V. Campos, A. Trott, C. Xiong, R. Socher, X. Giro-i Nieto, and J. Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. *arXiv preprint arXiv:2002.03647*, 2020.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

G. A. Carpenter and S. Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1):54–115, 1987.

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. doi: 10.1023/A:1007379606734. URL http://dx.doi.org/10.1023/A:1007379606734.

I. Cases, C. Rosenbaum, M. Riemer, A. Geiger, T. Klinger, A. Tamkin, O. Li, S. Agarwal, J. D. Greene, D. Jurafsky, et al. Recursive routing networks: Learning to compose modules for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3631–3648, 2019.

Y. Chandak, G. Theocharous, J. Kostas, S. Jordan, and P. S. Thomas. Learning action representations for reinforcement learning. *arXiv preprint arXiv:1902.00183*, 2019.

Y. Chandak, S. M. Jordan, G. Theocharous, M. White, and P. S. Thomas. Towards safe policy improvement for non-stationary mdps, 2020a.

Y. Chandak, G. Theocharous, S. Shankar, S. Mahadevan, M. White, and P. S. Thomas. Optimizing for the future in non-stationary mdps. *arXiv preprint arXiv:2005.08158*, 2020b.

M. B. Chang, A. Gupta, S. Levine, and T. L. Griffiths. Automatically composing representation transformations as a means for generalization. *arXiv preprint arXiv:1807.04640*, 2018.

A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny. Efficient lifelong learning with a-gem. *ICLR*, 2019.

A. Chemero. An outline of a theory of affordances. *Ecological psychology*, 15(2):181–195, 2003.

D. Chen, Q. Yan, S. Guo, Z. Yang, X. Su, and F. Chen. Learning effective subgoals with multi-task hierarchical reinforcement learning. 2019.

N. Chentanez, A. G. Barto, and S. P. Singh. Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 1281–1288, 2005.

B. Cheung, A. Terekhov, Y. Chen, P. Agrawal, and B. Olshausen. Superposition of many models into one. In *Advances in Neural Information Processing Systems*, pages 10868–10877, 2019.

W. C. Cheung, D. Simchi-Levi, and R. Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, pages 1843–1854. PMLR, 2020.

S. P. Choi, D.-Y. Yeung, and N. L. Zhang. Hidden-mode markov decision processes for nonstationary sequential decision making. In *Sequence Learning*, pages 264–287. Springer, 2000.

K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman. Quantifying generalization in reinforcement learning. *arXiv preprint arXiv:1812.02341*, 2018.

K. Cobbe, C. Hesse, J. Hilton, and J. Schulman. Leveraging procedural generation to benchmark reinforcement learning. *arXiv preprint arXiv:1912.01588*, 2019a.

K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, pages 1282–1289. PMLR, 2019b.

C. Colas, P. Fournier, M. Chetouani, O. Sigaud, and P.-Y. Oudeyer. Curious: intrinsically motivated modular multi-goal reinforcement learning. In *International conference on machine learning*, pages 1331–1340, 2019.

A. G. Collins and M. J. Frank. How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7):1024–1035, 2012.

B. C. Da Silva, E. W. Basso, A. L. Bazzan, and P. M. Engel. Dealing with non-stationary environments using context detection. In *Proceedings of the 23rd international conference on Machine learning*, pages 217–224, 2006.

N. D. Daw and D. S. Touretzky. Behavioral considerations suggest an average reward td model of the dopamine system. *Neurocomputing*, 32:679–684, 2000.

N. D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.

N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

P. Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2019.

C. D'Eramo, D. Tateo, A. Bonarini, M. Restelli, and J. Peters. Sharing knowledge in multi-task deep reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgpv2VFvr.

G. J. Detre, A. Natarajan, S. J. Gershman, and K. A. Norman. Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia*, 51(12):2371–2388, 2013.

C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine. Learning modular neural network policies for multi-task and multi-robot transfer. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2169–2176. IEEE, 2017.

T. G. Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

C. Diuk, A. Schapiro, N. Córdova, J. Ribas-Fernandes, Y. Niv, and M. Botvinick. Divide and conquer: hierarchical reinforcement learning and task decomposition in humans. In *Computational and robotic models of the hierarchical organization of behavior*, pages 271–291. Springer, 2013.

B. B. Doll, D. A. Simon, and N. D. Daw. The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6):1075–1081, 2012.

O. D. Domingues, P. Ménard, M. Pirotta, E. Kaufmann, and M. Valko. A kernel-based approach to non-stationary reinforcement learning in metric spaces. *arXiv preprint arXiv:2007.05078*, 2020.

S. Doncieux, N. Bredeche, L. L. Goff, B. Girard, A. Coninx, O. Sigaud, M. Khamassi, N. Díaz-Rodríguez, D. Filliat, T. Hospedales, et al. Dream architecture: a developmental approach to open-ended learning in robotics. *arXiv preprint arXiv:2005.06223*, 2020.

F. Doshi-Velez and G. Konidaris. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. *arXiv preprint arXiv:1308.3513*, 2013.

K. Doya, K. Samejima, K.-i. Katagiri, and M. Kawato. Multiple model-based reinforcement learning. *Neural computation*, 14(6):1347–1369, 2002.

S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019a.

S. S. Du, A. Krishnamurthy, N. Jiang, A. Agarwal, M. Dudík, and J. Langford. Provably efficient rl with rich observations via latent state decoding. *arXiv preprint arXiv:1901.09018*, 2019b.

Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. Rl2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

M. O. Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts at Amherst, 2002.

S. C. Duncan. Minecraft, beyond construction and survival. *Well Played: a journal on video games, value and meaning*, 1(1):1–22, 2011.

A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

B. Eysenbach, R. R. Salakhutdinov, and S. Levine. Search on the replay buffer: Bridging planning and reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 15246–15257, 2019.

R. Fakoor, P. Chaudhari, S. Soatto, and A. J. Smola. Meta-q-learning. *arXiv preprint arXiv:1910.00125*, 2019.

J. Farebrother, M. C. Machado, and M. Bowling. Generalization and regularization in dqn. *arXiv preprint arXiv:1810.00123*, 2018.

C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

C. Finn, A. Rajeswaran, S. Kakade, and S. Levine. Online meta-learning. *arXiv preprint arXiv:1902.08438*, 2019.

S. M. Fleming and N. D. Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1):91, 2017.

S. Flennerhag, P. G. Moreno, N. D. Lawrence, and A. Damianou. Transferring knowledge across learning processes. *arXiv preprint arXiv:1812.01054*, 2018.

C. Florensa, D. Held, X. Geng, and P. Abbeel. Automatic goal generation for reinforcement learning agents. *arXiv preprint arXiv:1705.06366*, 2017.

C. Florensa, D. Held, X. Geng, and P. Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pages 1515–1528, 2018.

J. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems, 2018a.

J. Foerster, G. Farquhar, M. Al-Shedivat, T. Rocktäschel, E. P. Xing, and S. Whiteson. Dice: The infinitely differentiable monte-carlo estimator. *arXiv preprint arXiv:1802.05098*, 2018b.

V. François-Lavet, Y. Bengio, D. Precup, and J. Pineau. Combined reinforcement learning via abstract representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3582–3589, 2019.

M. Frank, J. Leitner, M. Stollenga, A. Förster, and J. Schmidhuber. Curiosity driven reinforcement learning for motion planning on humanoids. *Frontiers in neurorobotics*, 7:25, 2014.

M. J. Frank and D. Badre. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cerebral cortex*, 22(3):509–526, 2012.

K. Frans, J. Ho, X. Chen, P. Abbeel, and J. Schulman. Meta learning shared hierarchies. *arXiv preprint arXiv:1710.09767*, 2017.

M. Frean and A. Robins. Catastrophic forgetting in simple networks: an analysis of the pseudorehearsal solution. *Network: Computation in Neural Systems*, 10(3):227–236, 1999.

R. M. French. Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. In *Proceedings of the 13th annual cognitive science society conference*, volume 1, pages 173–178, 1991.

Z. Gábor, Z. Kalmár, and C. Szepesvári. Multi-criteria reinforcement learning. In *ICML*, volume 98, pages 197–205, 1998.

S. J. Gershman. Reinforcement learning and causal models. In *The Oxford handbook of causal reasoning*, page 295. Oxford University Press, 2017.

S. J. Gershman, B. Pesaran, and N. D. Daw. Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience*, 29(43):13524–13531, 2009.

S. J. Gershman, K. A. Norman, and Y. Niv. Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, 5:43–50, 2015.

J. J. Gibson. The theory of affordances. *Hilldale, USA*, 1(2), 1977.

C. M. Gillan, A. R. Otto, E. A. Phelps, and N. D. Daw. Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, 15(3):523–536, 2015.

J. Gläscher, N. Daw, P. Dayan, and J. P. O'Doherty. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66 (4):585–595, 2010.

K. Gregor, D. J. Rezende, and D. Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

T. L. Griffiths, F. Callaway, M. B. Chang, E. Grant, P. M. Krueger, and F. Lieder. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29:24–30, 2019.

D. Guo, B. A. Pires, B. Piot, J.-b. Grill, F. Altché, R. Munos, and M. G. Azar. Bootstrap latent-predictive representations for multitask reinforcement learning. *arXiv preprint arXiv:2004.14646*, 2020.

E. Hadoux, A. Beynier, and P. Weng. Sequential decision-making under non-stationary environments via sequential change-point detection. 2014a.

E. Hadoux, A. Beynier, and P. Weng. Solving hidden-semi-markov-mode markov decision problems. In *International Conference on Scalable Uncertainty Management*, pages 176–189. Springer, 2014b.

R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028 – 1040, 2020. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2020.09.004. URL http://www.sciencedirect.com/science/article/pii/S1364661320302199.

J. Harb, P.-L. Bacon, M. Klissarov, and D. Precup. When waiting is not an option: Learning options with a deliberation cost. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

J. Harb, T. Schaul, D. Precup, and P.-L. Bacon. Policy evaluation networks. *arXiv preprint arXiv:2002.11833*, 2020.

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

D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245 – 258, 2017. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2017.06.011. URL http://www.sciencedirect.com/science/article/pii/S0896627317305093.

M. Hauskrecht, N. Meuleau, L. P. Kaelbling, T. Dean, and C. Boutilier. Hierarchical solution of markov decision processes using macro-actions. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998.

K. Hausman, J. T. Springenberg, Z. Wang, N. Heess, and M. Riedmiller. Learning an embedding space for transferable robot skills. 2018.

H. Heft. Affordances and the body: An intentional analysis of gibson's ecological approach to visual perception. *Journal for the theory of social behaviour*, 19(1):1–30, 1989.

P. Henderson, W.-D. Chang, F. Shkurti, J. Hansen, D. Meger, and G. Dudek. Benchmark environments for multitask learning in continuous domains. *arXiv preprint arXiv:1708.04352*, 2017a.

P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. *arXiv preprint arXiv:1709.06560*, 2017b.

T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

P. Hernandez-Leal, Y. Zhan, M. E. Taylor, L. E. Sucar, and E. M. de Cote. An exploration strategy for non-stationary opponents. *Autonomous Agents and Multi-Agent Systems*, 31(5):971–1002, 2017.

M. Hessel, H. Soyer, L. Espeholt, W. Czarnecki, S. Schmitt, and H. van Hasselt. Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3796–3803, 2019.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

S. Hooker. The hardware lottery. *arXiv preprint arXiv:2009.06489*, 2020.

T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

R. Houthooft, Y. Chen, P. Isola, B. Stadie, F. Wolski, O. J. Ho, and P. Abbeel. Evolved policy gradients. In *Advances in Neural Information Processing Systems*, pages 5400–5409, 2018.

J. Humplik, A. Galashov, L. Hasenclever, P. A. Ortega, Y. W. Teh, and N. Heess. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*, 2019.

C.-C. Hung, T. Lillicrap, J. Abramson, Y. Wu, M. Mirza, F. Carnevale, A. Ahuja, and G. Wayne. Optimizing agent behavior over long time scales by transporting value. *Nature communications*, 10(1):1–12, 2019.

M. Igl, A. Gambardella, J. He, N. Nardelli, N. Siddharth, W. Böhmer, and S. Whiteson. Multitask soft option learning. *arXiv preprint arXiv:1904.01033*, 2019.

H. Imamizu. Separated modules for visuomotor control and learning in the cerebellum: a functional mri study. In *NeuroImage: Third International Conference on Functional Mapping of the Human Brain*, volume 5, page S598. Academic Press, 1997.

H. Imamizu, S. Miyauchi, T. Tamada, Y. Sasaki, R. Takino, B. PuÈtz, T. Yoshioka, and M. Kawato. Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature*, 403(6766): 192–195, 2000.

D. Isele and A. Cosgun. Selective experience replay for lifelong learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

A. I. Jang, M. R. Nassar, D. G. Dillon, and M. J. Frank. Positive reward prediction errors during decision-making strengthen memory encoding. *Nature human behaviour*, 3(7):719–732, 2019.

K. Javed and M. White. Meta-learning representations for continual learning. pages 1818–1828. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8458-meta-learning-representations-for-continual-learning.pdf.

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

N. Justesen and S. Risi. Automated curriculum learning by rewarding temporally rare events. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2018.

N. Justesen, R. R. Torrado, P. Bontrager, A. Khalifa, J. Togelius, and S. Risi. Illuminating generalization in deep reinforcement learning through procedural level generation. *arXiv preprint arXiv:1806.10729*, 2018.

L. P. Kaelbling. Learning to achieve goals. In *IJCAI*, pages 1094–1099. Citeseer, 1993.

F. Kaplan and P.-Y. Oudeyer. Curiosity-driven development. In *Proceedings of the International Workshop on Synergistic Intelligence Dynamics*. Citeseer, 2006.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

C. Kaplanis, M. Shanahan, and C. Clopath. Continual reinforcement learning with complex synapses. *arXiv preprint arXiv:1802.07239*, 2018.

C. Kaplanis, M. Shanahan, and C. Clopath. Policy consolidation for continual reinforcement learning. *arXiv preprint arXiv:1902.00255*, 2019.

M. Karl, M. Soelch, P. Becker-Ehmck, D. Benbouzid, P. van der Smagt, and J. Bayer. Unsupervised real-time control through variational empowerment. *arXiv preprint arXiv:1710.05101*, 2017.

M. Kempka, M. Wydmuch, G. Runc, J. Toczek, and W. Jaśkowski. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *Computational Intelligence and Games (CIG), 2016 IEEE Conference on*, pages 1–8. IEEE, 2016.

K. Khetarpal, Z. Ahmed, A. Cianflone, R. Islam, and J. Pineau. Re-evaluate: Reproducibility in evaluating reinforcement learning algorithms. 2018a.

K. Khetarpal, S. Sodhani, S. Chandar, and D. Precup. Environments for lifelong reinforcement learning. *arXiv preprint arXiv:1811.10732*, 2018b.

K. Khetarpal, Z. Ahmed, G. Comanici, D. Abel, and D. Precup. What can i do here? a theory of affordances in reinforcement learning. *International Conference on Machine Learning*, 2020a.

K. Khetarpal, M. Klissarov, M. Chevalier-Boisvert, P.-L. Bacon, and D. Precup. Options of interest: Temporal abstraction with interest functions. *arXiv preprint arXiv:2001.00271*, 2020b.

M. Kiefer and N. M. Trumpp. Embodiment theory and education: The foundations of cognition in perception and action. *Trends in Neuroscience and Education*, 1(1):15–20, 2012.

D. Kiela, L. Bulat, A. L. Vero, and S. Clark. Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *CoRR*, abs/1610.07432, 2016. URL http://arxiv.org/abs/1610.07432.

D.-K. Kim, M. Liu, M. Riemer, C. Sun, M. Abdulhai, G. Habibi, S. Lopez-Cot, G. Tesauro, and J. P. How. A policy gradient algorithm for learning to learn in multiagent reinforcement learning, 2020.

T. Kim, J. Yoon, O. Dia, S. Kim, Y. Bengio, and S. Ahn. Bayesian model-agnostic meta-learning. *arXiv preprint arXiv:1806.03836*, 2018.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

A. S. Klyubin, D. Polani, and C. L. Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135. IEEE, 2005.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Ingredients of intelligence: From classic debates to an engineering roadmap. *Behavioral and Brain Sciences*, 40, 2017.

G. Lample and D. S. Chaplot. Playing fps games with deep reinforcement learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

A. J. Langdon, M. J. Sharpe, G. Schoenbaum, and Y. Niv. Model-based predictions for dopamine. *Current Opinion in Neurobiology*, 49:1–7, 2018.

E. Lecarpentier and E. Rachelson. Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 7214–7223, 2019.

S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

A. Levy, G. Konidaris, R. Platt, and K. Saenko. Learning multi-level hierarchies with hindsight. *arXiv preprint arXiv:1712.00948*, 2017.

A. C. Li, L. Pinto, and P. Abbeel. Generalized hindsight for reinforcement learning. *arXiv preprint arXiv:2002.11708*, 2020.

H. Li, X. Liao, and L. Carin. Multi-task reinforcement learning in partially observable stochastic environments. *Journal of Machine Learning Research*, 10(May):1131–1186, 2009.

L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for mdps. In *ISAIM*, 2006.

S. Li, F. Gu, G. Zhu, and C. Zhang. Context-aware policy reuse. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 989–997. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

X. Li, L. Li, J. Gao, X. He, J. Chen, L. Deng, and J. He. Recurrent reinforcement learning: a hybrid approach. *arXiv preprint arXiv:1509.03044*, 2015.

Z. Li and D. Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pages 614–629. Springer, 2016.

J. Liang, E. Meyerson, and R. Miikkulainen. Evolutionary architecture search for deep multitask networks. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 466–473, 2018.

L.-J. Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3-4):293–321, 1992.

Z. Lin, G. Thomas, G. Yang, and T. Ma. Model-based adversarial meta-reinforcement learning. *arXiv preprint arXiv:2006.08875*, 2020.

H. Liu, A. Trott, R. Socher, and C. Xiong. Competitive experience replay. *arXiv preprint arXiv:1902.00528*, 2019.

D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continuum learning. *Advances in Neural Information Processing Systems*, 2017.

A. S. Lowet, Q. Zheng, S. Matias, J. Drugowitsch, and N. Uchida. Distributional reinforcement learning in the brain. *Trends in Neurosciences*, 2020.

K. Lu, A. Grover, P. Abbeel, and I. Mordatch. Reset-free lifelong learning with skill-space planning. *NeurIPS Workshop on Deep Reinforcement Learning*, 2020.

M. C. Machado, M. G. Bellemare, and M. Bowling. A laplacian framework for option discovery in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2295–2304. JMLR. org, 2017a.

M. C. Machado, C. Rosenbaum, X. Guo, M. Liu, G. Tesauro, and M. Campbell. Eigenoption discovery through the deep successor representation. *arXiv preprint arXiv:1710.11089*, 2017b.

M. C. Machado, M. G. Bellemare, E. Talvitie, J. Veness, M. Hausknecht, and M. Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

S. J. Majeed and M. Hutter. On q-learning convergence for non-markov decision processes. In *IJCAI*, pages 2546–2552, 2018.

D. J. Mankowitz, T. A. Mann, and S. Mannor. Adaptive skills adaptive partitions (asap). In *Advances in Neural Information Processing Systems*, pages 1588–1596, 2016.

D. J. Mankowitz, A. Žídek, A. Barreto, D. Horgan, M. Hessel, J. Quan, J. Oh, H. van Hasselt, D. Silver, and T. Schaul. Unicorn: Continual learning with a universal, off-policy agent. *arXiv preprint arXiv:1802.08294*, 2018.

A. Marinescu, I. Dusparic, and S. Clarke. Prediction-based multi-agent reinforcement learning in inherently non-stationary environments. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 12(2):1–23, 2017.

J. J. Martin. *Bayesian decision problems and Markov chains*. Wiley, 1967.

M. G. Mattar and N. D. Daw. Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, 21(11):1609–1617, 2018.

A. Maurer, M. Pontil, and B. Romera-Paredes. The Benefit of Multitask Representation Learning. page 32, 2016.

J. L. McClelland, B. L. McNaughton, and R. C. O'reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165, 1989.

R. Mendonca, A. Gupta, R. Kralev, P. Abbeel, S. Levine, and C. Finn. Guided meta-policy search. In *Advances in Neural Information Processing Systems*, pages 9653–9664, 2019.

E. Meyerson and R. Miikkulainen. Beyond shared hierarchies: Deep multitask learning through soft layer ordering. *arXiv preprint arXiv:1711.00108*, 2017.

P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

S. Mohamed and D. J. Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 2125–2133, 2015.

I. Momennejad. Learning structures: Predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, 32:155–166, 2020.

I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, and S. J. Gershman. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680–692, 2017.

I. Momennejad, A. R. Otto, N. D. Daw, and K. A. Norman. Offline replay supports planning in human reinforcement learning. *Elife*, 7:e32548, 2018.

A. Nagabandi, C. Finn, and S. Levine. Deep online learning via meta-learning: Continual adaptation for model-based rl. *arXiv preprint arXiv:1812.07671*, 2018.

A. Nichol and J. Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2018.

A. Nichol, V. Pfau, C. Hesse, O. Klimov, and J. Schulman. Gotta learn fast: A new benchmark for generalization in rl. *arXiv preprint arXiv:1804.03720*, 2018.

Y. Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, 2009.

Y. Niv. Learning task-state representations. *Nature neuroscience*, 22(10):1544–1553, 2019.

Y. Niv and S. Chan. On the value of information and other rewards. *Nature neuroscience*, 14(9): 1095–1097, 2011.

Y. Niv, M. O. Duff, and P. Dayan. Dopamine, uncertainty and td learning. *Behavioral and brain Functions*, 1(1):6, 2005.

Y. Niv, R. Daniel, A. Geana, S. J. Gershman, Y. C. Leong, A. Radulescu, and R. C. Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21):8145–8157, 2015.

J. Oh, S. Singh, H. Lee, and P. Kohli. Zero-shot task generalization with multi-task deep reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2661–2670. JMLR. org, 2017.

J. Oh, M. Hessel, W. M. Czarnecki, Z. Xu, H. P. van Hasselt, S. Singh, and D. Silver. Discovering reinforcement learning algorithms. *Advances in Neural Information Processing Systems*, 33, 2020.

S. C. Ong, S. W. Png, D. Hsu, and W. S. Lee. Planning under uncertainty for robotic tasks with mixed observability. *The International Journal of Robotics Research*, 29(8):1053–1068, 2010.

P. A. Ortega, J. X. Wang, M. Rowland, T. Genewein, Z. Kurth-Nelson, R. Pascanu, N. Heess, J. Veness, A. Pritzel, P. Sprechmann, et al. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

I. Osband, Y. Doron, M. Hessel, J. Aslanides, E. Sezener, A. Saraiva, K. McKinney, T. Lattimore, C. Szepezvari, S. Singh, et al. Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*, 2019.

C. Packer, K. Gao, J. Kos, P. Krähenbühl, V. Koltun, and D. Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.

S. Padakandla, S. Bhatnagar, et al. Reinforcement learning in non-stationary environments. *arXiv preprint arXiv:1905.03970*, 2019.

Y. Pan, M. Zaheer, A. White, A. Patterson, and M. White. Organizing experience: a deeper look at replay mechanisms for sample-based planning in continuous state domains. *arXiv preprint arXiv:1806.04624*, 2018.

G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

R. Parr and S. J. Russell. Reinforcement learning with hierarchies of machines. In *Advances in neural information processing systems*, pages 1043–1049, 1998.

R. Pasunuru and M. Bansal. Continual and multi-task architecture search. *arXiv preprint arXiv:1906.05226*, 2019.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

C. F. Perez, F. P. Such, and T. Karaletsos. Generalized hidden parameter mdps transferable model-based rl in a handful of trials. *arXiv preprint arXiv:2002.03072*, 2020.

G. Pezzulo and P. Cisek. Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends in cognitive sciences*, 20(6):414–424, 2016.

E. A. Platanios, A. Saparov, and T. Mitchell. Jelly bean world: A testbed for never-ending learning. *arXiv preprint arXiv:2002.06306*, 2020.

R. Portelas, C. Colas, L. Weng, K. Hofmann, and P.-Y. Oudeyer. Automatic curriculum learning for deep rl: A short survey. *arXiv preprint arXiv:2003.04664*, 2020.

M. Puterman. Markov decision processes. 1994. *Jhon Wiley & Sons, New Jersey*, 1994.

S. Racaniere, A. K. Lampinen, A. Santoro, D. P. Reichert, V. Firoiu, and T. P. Lillicrap. Automated curricula through setter-solver interactions. *arXiv preprint arXiv:1909.12892*, 2019.

R. Raileanu, M. Goldstein, A. Szlam, and R. Fergus. Fast adaptation via policy-dynamics value functions. *arXiv preprint arXiv:2007.02879*, 2020.

K. Rakelly, A. Zhou, D. Quillen, C. Finn, and S. Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. *arXiv preprint arXiv:1903.08254*, 2019.

P. Ramachandran and Q. V. Le. Diversity and depth in per-example routing models. 2018.

P. Rauber, A. Ummadisingu, F. Mutz, and J. Schmidhuber. Hindsight policy gradients. *arXiv preprint arXiv:1711.06006*, 2017.

J. J. Ribas-Fernandes, D. Shahnazian, C. B. Holroyd, and M. M. Botvinick. Subgoal-and goal-related reward prediction errors in medial prefrontal cortex. *Journal of cognitive neuroscience*, 31(1):8–23, 2019.

M. Riemer, E. Khabiri, and R. Goodwin. Representation stability as a regularizer for improved text analytics transfer learning. *arXiv preprint arXiv:1704.03617*, 2016.

M. Riemer, T. Klinger, D. Bouneffouf, and M. Franceschini. Scalable recollections for continual lifelong learning. *arXiv preprint arXiv:1711.06761*, 2017.

M. Riemer, M. Liu, and G. Tesauro. Learning abstract options. *NIPS*, 2018.

M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *ICLR*, 2019.

M. Riemer, I. Cases, C. Rosenbaum, M. Liu, and G. Tesauro. On the role of weight sharing during deep option learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5519–5526, 2020.

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

M. Rmus, S. McDougle, and A. Collins. The role of executive function in shaping reinforcement learning. 2020.

A. Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

A. Robins. Consolidation in neural networks and in the sleeping brain. *Connection Science*, 8(2): 259–276, 1996.

D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, pages 348–358, 2019.

C. Rosenbaum, T. Klinger, and M. Riemer. Routing networks: Adaptive selection of non-linear functions for multi-task learning. *arXiv preprint arXiv:1711.01239*, 2017.

C. Rosenbaum, I. Cases, M. Riemer, and T. Klinger. Routing networks and the challenges of modular and compositional computation. *arXiv preprint arXiv:1904.12774*, 2019.

B. S. Rosman and S. Ramamoorthy. A multitask representation using reusable local policy templates. In *2012 AAAI Spring Symposium Series*, 2012.

J. Rothfuss, D. Lee, I. Clavera, T. Asfour, and P. Abbeel. Promp: Proximal meta-policy search. *arXiv preprint arXiv:1810.06784*, 2018.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

A. A. Rusu, S. G. Colmenarejo, C. Gulcehre, G. Desjardins, J. Kirkpatrick, R. Pascanu, V. Mnih, K. Kavukcuoglu, and R. Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Y. Sagiv, S. Musslick, Y. Niv, and J. D. Cohen. Efficiency of learning vs. processing: Towards a normative theory of multitasking. *arXiv preprint arXiv:2007.03124*, 2020.

H. Sahni, S. Kumar, F. Tejani, and C. Isbell. Learning to compose skills. *arXiv preprint arXiv:1711.11289*, 2017.

R. Samson, M. Frank, and J.-M. Fellous. Computational models of reinforcement learning: the role of dopamine as a reward signal. *Cognitive neurodynamics*, 4(2):91–105, 2010.

P. A. Samuelson. A note on measurement of utility. *The review of economic studies*, 4(2):155–161, 1937.

T. Schaul, D. Horgan, K. Gregor, and D. Silver. Universal value function approximators. In *International conference on machine learning*, pages 1312–1320, 2015a.

T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015b.

T. Schaul, H. van Hasselt, J. Modayil, M. White, A. White, P.-L. Bacon, J. Harb, S. Mourad, M. Bellemare, and D. Precup. The barbados 2018 list of open issues in continual learning. *arXiv preprint arXiv:1811.07004*, 2018.

J. Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent n eu al networks for dynamic reinforcement learning and planning in non-stationary environm nts.

J. Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

J. Schmidhuber. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pages 1458–1463, 1991.

J. Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

J. Schmidhuber. A general method for incremental self-improvement and multi-agent learning. In *Evolutionary Computation: Theory and Applications*, pages 81–123. World Scientific, 1999.

J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Workshop on anticipatory behavior in adaptive learning systems*, pages 48–76. Springer, 2008.

J. Schmidhuber. Powerplay: Training an increasingly general problem solver by continually searching for the simplest still unsolvable problem. *Frontiers in psychology*, 4:313, 2013.

J. Schmidhuber, J. Zhao, and M. Wiering. Shifting inductive bias with success-story algorithm, adaptive levin search, and incremental self-improvement. *Machine Learning*, 28(1):105–130, 1997.

J. Schmidhuber, J. Zhao, and N. N. Schraudolph. Reinforcement learning with self-modifying policies. In *Learning to learn*, pages 293–309. Springer, 1998.

N. W. Schuck and Y. Niv. Sequential replay of nonspatial task states in the human hippocampus. *Science*, 364(6447):eaaw5181, 2019.

J. Schwarz, J. Luketina, W. M. Czarnecki, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell. Progress & compress: A scalable framework for continual learning. *arXiv preprint arXiv:1805.06370*, 2018.

R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. Planning to explore via self-supervised world models. *arXiv preprint arXiv:2005.05960*, 2020.

S. Sharma, A. Jha, P. Hegde, and B. Ravindran. Learning to multi-task by active sampling. *arXiv preprint arXiv:1702.06053*, 2017.

U. Sharma and J. Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

E. Shelhamer, P. Mahmoudieh, M. Argus, and T. Darrell. Loss is its own reward: Self-supervision for reinforcement learning. *arXiv preprint arXiv:1612.07307*, 2016.

T. Shu, C. Xiong, and R. Socher. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. *arXiv preprint arXiv:1712.07294*, 2017.

T. Shu, C. Xiong, Y. N. Wu, and S.-C. Zhu. Interactive agent modeling by learning to probe. *arXiv preprint arXiv:1810.00510*, 2018.

P. Shyam, W. Jaśkowski, and F. Gomez. Model-based active exploration. In *International Conference on Machine Learning*, pages 5779–5788, 2019.

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

D. A. Simon and N. D. Daw. Environmental statistics and the trade-off between model-based and td learning in humans. In *Advances in neural information processing systems*, pages 127–135, 2011.

S. Singh, R. L. Lewis, and A. G. Barto. Where do rewards come from. In *Proceedings of the annual conference of the cognitive science society*, pages 2601–2606. Cognitive Science Society, 2009.

S. P. Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, 8(3-4):323–339, 1992.

A. Solway, C. Diuk, N. Córdova, D. Yee, A. G. Barto, Y. Niv, and M. M. Botvinick. Optimal behavioral hierarchy. *PLOS Comput Biol*, 10(8):e1003779, 2014.

G. Spigler. Meta-learnt priors slow down catastrophic forgetting in neural networks. *arXiv preprint arXiv:1909.04170*, 2019. URL https://arxiv.org/pdf/1909.04170.pdf.

A. Srinivasan, D. Bahdanau, M. Chevalier-Boisvert, and Y. Bengio. Automated curriculum generation for policy gradients from demonstrations. *arXiv preprint arXiv:1912.00444*, 2019.

B. C. Stadie, G. Yang, R. Houthooft, X. Chen, Y. Duan, Y. Wu, P. Abbeel, and I. Sutskever. Some considerations on learning to explore via meta-reinforcement learning. *arXiv preprint arXiv:1803.01118*, 2018.

C. K. Starkweather, B. M. Babayan, N. Uchida, and S. J. Gershman. Dopamine reward prediction errors reflect hidden-state inference across time. *Nature neuroscience*, 20(4):581–589, 2017.

S. Sukhbaatar, Z. Lin, I. Kostrikov, G. Synnaeve, A. Szlam, and R. Fergus. Intrinsic motivation and automatic curricula via asymmetric self-play. *arXiv preprint arXiv:1703.05407*, 2017.

F. Sung, L. Zhang, T. Xiang, T. Hospedales, and Y. Yang. Learning to learn: Meta-critic networks for sample efficient learning. *arXiv preprint arXiv:1706.09529*, 2017.

R. Sutton. Adapting bias by gradient descent: an incremental version of the delta-bar-delta. In *Tenth National Conference on Artificial Intelligence*. MIT Press, 1992.

R. Sutton. The bitter lesson. *Incomplete Ideas (blog), March*, 13:12, 2019.

R. S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991.

R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

R. S. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

R. S. Sutton, J. Modayil, M. D. T. Degris, P. M. Pilarski, and A. White. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. 2017.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.

C. Tessler, S. Givony, T. Zahavy, D. J. Mankowitz, and S. Mannor. A deep hierarchical approach to lifelong learning in minecraft. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

S. Thrun and J. O'Sullivan. Discovering structure in multiple learning tasks: The tc algorithm. In *ICML*, volume 96, pages 489–497, 1996.

S. Thrun and A. Schwartz. Finding structure in reinforcement learning. In *Advances in neural information processing systems*, pages 385–392, 1995.

D. Tirumala, H. Noh, A. Galashov, L. Hasenclever, A. Ahuja, G. Wayne, R. Pascanu, Y. W. Teh, and N. Heess. Exploiting hierarchy for learning and transfer in kl-regularized rl. *arXiv preprint arXiv:1903.07438*, 2019.

J. Tobin. Beyond domain randomization. *Slides*, 2019. URL http://josh-tobin.com/assets/pdf/BeyondDomainRandomization_Tobin_RSS19.pdf.

A. Touati and P. Vincent. Efficient learning in non-stationary linear markov decision processes, 2020.

R. Traoré, H. Caselles-Dupré, T. Lesort, T. Sun, G. Cai, N. Díaz-Rodríguez, and D. Filliat. Discorl: Continual reinforcement learning via policy distillation. *arXiv preprint arXiv:1907.05855*, 2019.

V. Veeriah, J. Oh, and S. Singh. Many-goals reinforcement learning. *arXiv preprint arXiv:1806.09605*, 2018.

V. Veeriah, M. Hessel, Z. Xu, J. Rajendran, R. L. Lewis, J. Oh, H. P. van Hasselt, D. Silver, and S. Singh. Discovery of useful questions as auxiliary tasks. In *Advances in Neural Information Processing Systems*, pages 9306–9317, 2019.

A. S. Vezhnevets, Y. Wu, R. Leblond, and J. Leibo. Options as responses: Grounding behavioural hierarchies in multi-agent rl. *arXiv preprint arXiv:1906.01470*, 2019.

O. M. Vikbladh, M. R. Meager, J. King, K. Blackmon, O. Devinsky, D. Shohamy, N. Burgess, and N. D. Daw. Hippocampal contributions to model-based planning and spatial memory. *Neuron*, 102 (3):683–693, 2019.

R. Vilalta and Y. Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

J. X. Wang, Z. Kurth-Nelson, D. Kumaran, D. Tirumala, H. Soyer, J. Z. Leibo, D. Hassabis, and M. Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21 (6):860–868, 2018.

R. Wang, J. Lehman, J. Clune, and K. O. Stanley. Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions. *arXiv preprint arXiv:1901.01753*, 2019.

R. Wang, J. Lehman, A. Rawal, J. Zhi, Y. Li, J. Clune, and K. O. Stanley. Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions. *arXiv preprint arXiv:2003.08536*, 2020.

A. Wilson, A. Fern, S. Ray, and P. Tadepalli. Multi-task reinforcement learning: a hierarchical bayesian approach. In *Proceedings of the 24th international conference on Machine learning*, pages 1015–1022. ACM, 2007.

G. E. Wimmer, N. D. Daw, and D. Shohamy. Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience*, 35(7):1092–1104, 2012.

D. M. Wolpert, R. C. Miall, and M. Kawato. Internal models in the cerebellum. *Trends in cognitive sciences*, 2(9):338–347, 1998.

M. Wortsman, V. Ramanujan, R. Liu, A. Kembhavi, M. Rastegari, J. Yosinski, and A. Farhadi. Supermasks in superposition. *arXiv preprint arXiv:2006.14769*, 2020.

A. Xie, J. Harrison, and C. Finn. Deep reinforcement learning amidst lifelong non-stationarity. *arXiv preprint arXiv:2006.10701*, 2020.

J. Xu and Z. Zhu. Reinforced continual learning. In *Advances in Neural Information Processing Systems*, pages 899–908, 2018.

T. Xu, Q. Liu, L. Zhao, and J. Peng. Learning to explore with meta-policy gradient. *arXiv preprint arXiv:1803.05044*, 2018a.

Z. Xu, H. P. van Hasselt, and D. Silver. Meta-gradient reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2402–2413, 2018b.

Z. Xu, L. Cao, and X. Chen. Learning to learn: Hierarchical meta-critic networks. *IEEE Access*, 7: 57069–57077, 2019.

Z. Xu, H. P. van Hasselt, M. Hessel, J. Oh, S. Singh, and D. Silver. Meta-gradient reinforcement learning with an objective discovered online. *Advances in Neural Information Processing Systems*, 33, 2020.

R. Yang, H. Xu, Y. Wu, and X. Wang. Multi-task reinforcement learning with soft modularization. *arXiv preprint arXiv:2003.13661*, 2020.

Y. Yang, K. Caluwaerts, A. Iscen, J. Tan, and C. Finn. Norml: No-reward meta learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 323–331. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020a.

T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pages 1094–1100, 2020b.

T. Zahavy, Z. Xu, V. Veeriah, M. Hessel, J. Oh, H. van Hasselt, D. Silver, and S. Singh. Self-tuning deep reinforcement learning. *arXiv preprint arXiv:2002.12928*, 2020.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

A. Zhang, H. Satija, and J. Pineau. Decoupling dynamics and reward for transfer learning. *arXiv preprint arXiv:1804.10689*, 2018a.

A. Zhang, Z. C. Lipton, L. Pineda, K. Azizzadenesheli, A. Anandkumar, L. Itti, J. Pineau, and T. Furlanello. Learning causal state representations of partially observable environments. *arXiv preprint arXiv:1906.10437*, 2019.

A. Zhang, C. Lyle, S. Sodhani, A. Filos, M. Kwiatkowska, J. Pineau, Y. Gal, and D. Precup. Invariant causal prediction for block mdps. *arXiv preprint arXiv:2003.06016*, 2020.

C. Zhang, O. Vinyals, R. Munos, and S. Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018b.

Z. Zheng, J. Oh, and S. Singh. On learning intrinsic rewards for policy gradient methods. In *Advances in Neural Information Processing Systems*, pages 4644–4654, 2018.

Z. Zheng, J. Oh, M. Hessel, Z. Xu, M. Kroiss, H. van Hasselt, D. Silver, and S. Singh. What can learned intrinsic rewards capture? *arXiv preprint arXiv:1912.05500*, 2019.

W. Zhou, Y. Li, Y. Yang, H. Wang, and T. M. Hospedales. Online meta-critic learning for off-policy actor-critic methods. *arXiv preprint arXiv:2003.05334*, 2020.

F. Zhu, X. Chang, R. Zeng, and M. Tan. Continual reinforcement learning with diversity exploration and adversarial self-correction. *arXiv preprint arXiv:1906.09205*, 2019.

L. Zintgraf, K. Shiarlis, M. Igl, S. Schulze, Y. Gal, K. Hofmann, and S. Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

L. M. Zintgraf, K. Shiarlis, V. Kurin, K. Hofmann, and S. Whiteson. Fast context adaptation via meta-learning. *arXiv preprint arXiv:1810.03642*, 2018.

B. Zoph and Q. V. Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

H. Zou, T. Ren, D. Yan, H. Su, and J. Zhu. Reward shaping via meta-learning. *arXiv preprint arXiv:1901.09330*, 2019.

Appendix A. The Relationship Between Continual RL and Continual Supervised Learning

As noted in Barto and Dietterich (2004), supervised learning can be cast as a special case of RL in continuing environments. In this section, we consider a particular set of assumptions that simplify the RL framework to better match the typical supervised learning setting:

1. **Deterministic Policies:** Instead of a potentially stochastic policy $a \sim \pi(s; \theta)$ with parameters θ, we consider a deterministic function $\hat{y} = f(x; \theta)$ of the input x where the predicted output can be interpreted as a sampled action $A_t = \hat{Y}_t$ associated with sampled input X_t.

2. **Decomposed State Space:** The state space must provide enough information to compute the reward of a given action, so we consider the full state space of supervised learning to include both the input x and optimal output y^* i.e. $S_t = (X_t, Y_t^*)$. In this formulation, the supervised learning problem is partially observable (if Y_t^* is provided the problem becomes trivial).

3. **Differentiable Reward Function:** In supervised learning we generally consider a differentiable loss function in lieu of a reward function i.e. $r(s, a) = -\ell(y^*, \hat{y})$.

4. **Action Invariant Transitions:** We also assume that the transition dynamics of the environment $p(s'|s, a)$ do not depend on the agent's behavior. The data distribution is is rather drawn from the joint distribution of inputs and optimal outputs $p(x, y^*)$.

Following from these assumptions and considering the limit of the undiscounted problem i.e. $\gamma \to 1$ during a lifetime T, the continuing environment objective function from equation 6 can be extended to the supervised learning setting as:

$$J_{\text{continuing}}(\theta) = -\mathbb{E}_{p(x,y^*)}\left[\sum_{k=0}^{T} \ell(Y_{t+k}^*, \hat{Y}_{t+k} = f(X_{t+k}; \theta)) \right] \tag{11}$$

Notice that if we assume that the incoming data distribution is stationary, we can consider the linearity of expectations over random variables and bring the sum outside of the expectation:

$$J_{\text{stationary}}(\theta) = -\sum_{k=0}^{T} \mathbb{E}_{p(x,y^*)}\left[\ell(Y_{t+k}^*, \hat{Y}_{t+k} = f(X_{t+k}; \theta)) \right] \tag{12}$$

If $p(x, y^*)$ is assumed to be i.i.d., samples from this sum are unbiased, drawing an equivalence with the standard stochastic gradient descent (SGD) supervised learning objective:

$$J_{\text{SGD}}(\theta) = -\mathbb{E}_{p(x,y^*)}\left[\ell(Y_t^*, \hat{Y}_t = f(X_t; \theta)) \right] \tag{13}$$

However, in continual supervised learning, we generally consider the data to be following some non-stationary pattern $p(x, y^*, t)$. Unfortunately, the linearity of expectations only applies over random variables, so it becomes clear that the SGD objective is biased towards the current experience without considering the long-term effects of parameter changes:

$$
\begin{aligned}
J_{\text{non-stationary}}(\theta) &= -\mathbb{E}_{p(x,y^*,t)}\left[\sum_{k=0}^{T} \ell(Y_{t+k}^*, \hat{Y}_{t+k} = f(X_{t+k}; \theta)) \right] \\
&= -\mathbb{E}_{p(x,y^*,t)}\left[\ell(Y_t^*, \hat{Y}_t = f(X_t; \theta)) + \sum_{k=1}^{T} \ell(Y_{t+k}^*, \hat{Y}_{t+k} = f(X_{t+k}; \theta)) \right]
\end{aligned}
\tag{14}
$$

When experiences from the past data distribution are expected to re-occur in the future distribution but not the current distribution, the biased objective of SGD may thus naturally lead to catastrophic forgetting of this past knowledge. Notice that the SGD objective has originally been shown to converge in the stationary and i.i.d. setting, so outside of that setting we must use another valid framework to model learning. From this analysis, it is clear that the RL framework is general enough to capture many of the important features of continual supervised learning in non-stationary settings.