Journal of Artificial Intelligence Research 75 (2022) 1401-1476 Submitted 02/2022; published 12/2022

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 by mathematically characterizing two key properties of non-stationarity, namely, the scope and driver non-stationarity. This offers a unified view of various formulations. Next, we review and present a taxonomy of continual RL approaches. 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. 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 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, Pineau, and Precup, 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 RL*. 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 online, 2) it learns behaviors or skills while solving presented tasks, 3) learning is task agnostic, 4) it learns 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, Kemker, Part, Kanan, and Wermter, 2019), (De Lange, Aljundi, Masana, Parisot, Jia, Leonardis, Slabaugh, and Tuytelaars, 2019), (Mundt, Hong, Pliushch, and Ramesh, 2020) and (Hadsell, Rao, Rusu, and Pascanu, 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 connections to neuroscience 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, and domain randomization, 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, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $p : \mathcal{S} \times \mathcal{A} \to Dist(\mathcal{S})$ 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 $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ with internal parameters $\theta \in \Theta$, and with probability $p(s'|s, a)$ enters next state s', receiving a numerical reward $r(s, a)$ 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*,

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

and one-step expected rewards,

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

for all $s, s' \in S$ and $a \in A$. $P^a_{ss'}$ and R^a_s 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, $\pi : S \times A \to [0, 1]$. 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:

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

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

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

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 as defined in equation (4). Similarly, the *optimal action-value function* $q_*(s, a)$ is the maximum action-value-function over all policies as defined in equation (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.

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 most 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 most 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 also possible that MDPs may vary as a

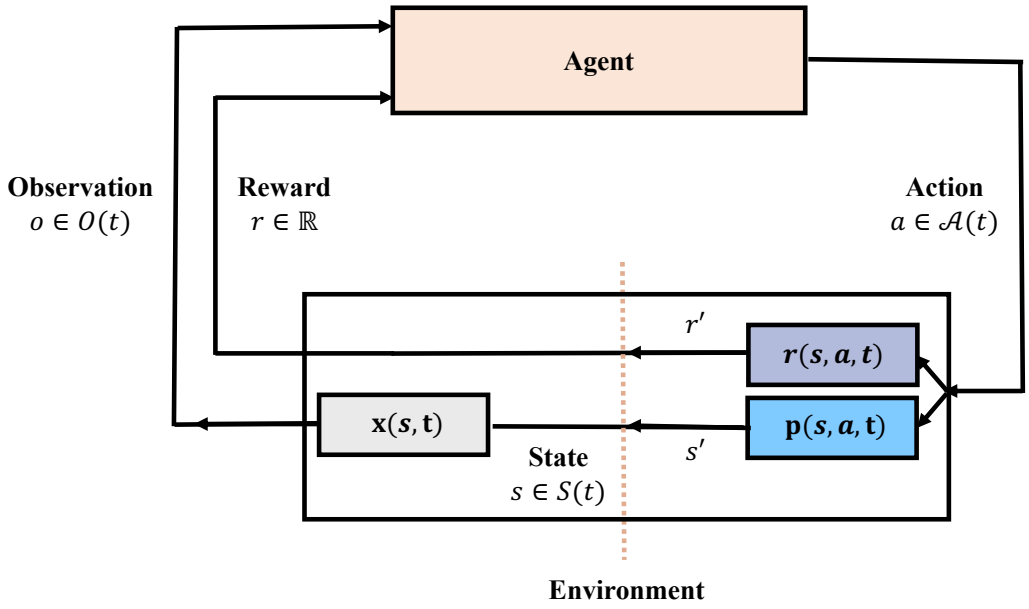

Figure 1: **Agent-Environment Interaction with Potentially Time Dependent Environment Components.** Extending Figure 3.1 of (Sutton and Barto, 1998) to highlight the agent-environment interaction in continual reinforcement learning.

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. Indeed, as explained in (Xie, Harrison, and Finn, 2021), non-stationarity chiefly adds a non-Markovian aspect to the learning setting that the POMDP framework helps directly address.

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 (Riemer, Raparthy, Cases, Subbaraj, Touzel, and Rish, 2022)).

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 Deployment Domains	Multiple Required Skills	Requires Online Learning	Requires Resource Efficiency & Sustainability	Non-stationary Task Evolution
Domain Adaptation	✓	X	X	X	X
Transfer Learning	✓	✓	X	X	✓
Meta-Training and Meta-Testing	✓	✓	X	X	X
Multi-task Learning	✓	✓	X	X	X
Continual (Lifelong) Learning	✓	✓	✓	✓	✓

Figure 2: **A Spectrum of Learning Settings**: For each setting we consider whether they typically involve multiple domains, multiple skills, online learning, resource efficiency/sustainability and a non-stationary evolution of the task distribution.

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. See Sec. 5.3.2 for a more in depth discussion of this setting.

Multi-task RL: More closely related to the aforementioned task incremental learning setting, is *multi-task reinforcement learning*. 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, Fern, Ray, and Tadepalli, 2007; Ammar, Eaton, Ruvolo, and Taylor, 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 behaviour policies and be made available in a batch generated beforehand in the *offline reinforcement learning* setting. We refer the reader to (Levine, Kumar, Tucker, and Fu, 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 maximizing the expected accumulated average return on all tasks, to learning a set of optimal policies for each task separately, to learning a set of shared skills leveraging 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. This difficulty is 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 policy over all 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 also 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, Machado, and Bowling, 2018) and game levels (Nichol, Pfau, Hesse, Klimov, and Schulman, 2018) for training and testing, by exposing internal parameters of various classic environments (Packer, Gao, Kos, Krähenbühl, Koltun, and Song, 2018), and by procedurally generating environments (Justesen, Torrado, Bontrager, Khalifa, Togelius, and Risi, 2018; Zhang, Vinyals, Munos, and Bengio, 2018a; Cobbe, Klimov, Hesse, Kim, and Schulman, 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, Kakade, Wang, and Yang, 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 often possible to study generalization without conflating its properties with the optimization difficulties specific to continual RL.

3. RL: 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* as we will highlight in this section.

For RL in a continuing environment, we define an objective function $J_{\text{continuing}}$ which is to learn a policy π that maximizes its value function $v_\pi(s)$. We seek to maximize this infinite horizon 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 π 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 myopic 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. As a result, it becomes clear that 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, Pascanu, Rabinowitz, Veness, Desjardins, Rusu, Milan, Quan, Ramalho, Grabska-Barwinska, et al., 2017; Lopez-Paz and Ranzato, 2017; Riemer, Cases, Ajemian, Liu, Rish, Tu, and Tesauro, 2019; Chaudhry, Ranzato, Rohrbach, and Elhoseiny, 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.

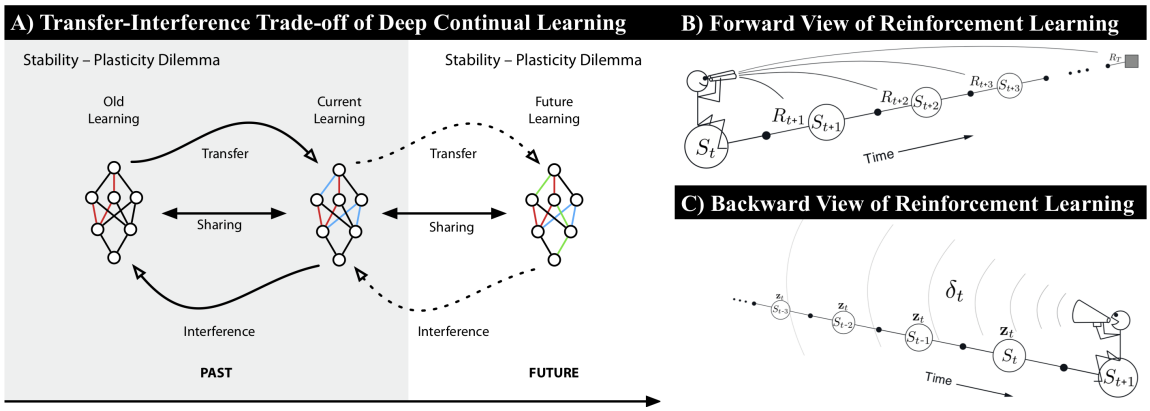

Figure 3: **Reinforcement Learning and the Stability-Plasticity Dilemma**: A) Depicts the stability-plasticity dilemma and its relation to both weight sharing and transfer dynamics over time (from (Riemer et al., 2019)). B) Depicts the forward view of RL where we evaluate the current state based on expected 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 and 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 previously seen data. However, it is counter intuitive to analyze this (as is typically done) without an explicit dependency on time. We should note 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 views 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

Due to the generality of the continual learning problem, formulations vary vastly in the literature as we also highlight in the former sections. In this section, we provide a taxonomy of the CRL problem formulations. Foremost, we consider the setting where all components of the problem can take a non-stationary functional form $f(i,t)$ as the most general continual reinforcement learning problem. Moreover, we detail some key additional assumptions on the functional form of non-stationarity that have been prominent in the literaure.

4.1 General CRL Problem

Formally, we understand a `General CRL Problem` $\mathcal{M}_{\texttt{CRL}}$ as:

Definition 1 (`General CRL Problem` $\mathcal{M}_{\texttt{CRL}}$): *Given a state space $\mathcal{S}$, action-space $\mathcal{A}$, an observation space $\mathcal{O}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathrm{R}$, a transition function $p : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, and an observation function $x : \mathcal{S} \to \mathcal{O}$, the most general continual reinforcement learning problem can be expressed as:*

$$\mathcal{M}_{\mathit{CRL}} \doteq \langle \mathcal{S}(t), \mathcal{A}(t), r(t), p(t), x(t), \mathcal{O}(t) \rangle, \tag{8}$$

where each component of the problem formulation can be considered as a non-stationary function of form $f(i,t)$ where i is the input specific to each component.

4.2 Common Non-stationary Functional Forms

Due to the very broad nature of the `General CRL Problem` in the definition above, concrete assumptions on the non-stationary functional form $f(i,t)$ are typical in the literature to help define structure in non-stationarity. This is indeed critical as an arbitrarily non-stationary environment gives no consistent signal to learn from, making learning impossible. As such, not making additional functional assumptions about the nature of non-stationarity can lead to a vacuous problem statement. The most common types of assumptions about non-stationarity in the literature include Lipschitz continuity and piecewise non-stationarity.

Definition 2 (`Non-stationary Functions With Lipschitz Continuity` $f(i,t)$): *A function is Lipschitz non-stationary if across the input space $i \in \mathcal{I}$ non-stationary can be bounded by a time independent constant C:*

$$\forall i \in \mathcal{I}, \forall t, t' \in \mathbb{R} \quad |f(i,t) - f(i,t')| \leq C|t - t'|. \tag{9}$$

This condition also ensures the function has bounded first derivatives in time $\partial f / \partial t$.

Definition 3 (`Piecewise Non-stationary Functions` $f(i,t)$): *A function is piecewise non-stationary if it can be seen as broken up into stationary functions $f_0(i), f_1(i), f_2(i), ...$ of the input space $i \in \mathcal{I}$ over intervals of time dictated by $t_0, t_1, t_2, ...$ such that:*

$$\forall i \in \mathcal{I} \quad f(i,t) = \begin{cases} f_0(i) & 0 \leq t < t_0 \\ f_1(i) & t_0 \leq t < t_1 \\ f_2(i) & t_1 \leq t < t_2 \\ ... \end{cases} \tag{10}$$

These are simply common examples of assumptions about the nature of non-stationarity. For example, another potentially interesting formulation is to assume arbitrary non-stationarity within a fixed variation budget as in (Mao, Zhang, Zhu, Simchi-Levi, and Basar, 2021).

4.3 Key Properties of Non-stationarity: Scope and Drivers

We would like to provide a taxonomy of formulations that includes prominent assumptions about non-stationarity that have been either explicitly or implicitly considered in the literature. Towards this end, we present a categorisation of non-stationarity along two primary dimensions, namely the **scope** and **driver** of non-stationarity.

Definition 4 (Scope of non-stationarity α): *defines what elements of the agent-environment interaction process experience non-stationarity:*

$$\alpha \subseteq \{\mathcal{S}, \mathcal{A}, r, p, x, \mathcal{O}\}, \tag{11}$$

where $p \in \alpha$ if $\exists t, t' \in \mathrm{R}, p(t)! = p(t')$, $r \in \alpha$ if $\exists t, t' \in \mathrm{R}, r(t)! = r(t')$, etc.

Definition 5 (Driver of non-stationarity β): *defines the causal assumptions that can be made about the nature of the evolution of non-stationary environment dynamics:*

$$\beta \in \{\texttt{stationary}, \texttt{passive}, \texttt{active}, \texttt{hybrid}\}, \tag{12}$$

where `stationary` $\implies \mathbb{E}[f(i,t)] = \mathbb{E}[f(i,t')] \;\; \forall t \in \mathbb{R}, \forall t' > t, \;\; \forall i \in \mathcal{I}$,
`passive` $\implies$ *if* $\mathbb{E}[f(i,t)] \neq \mathbb{E}[f(i,t')]$, *then* $|\mathbb{E}[f(i,t)] - \mathbb{E}[f(i,t')]| \perp\!\!\!\perp a \;\; \forall a \in \mathcal{A}, \;\; \forall t \in \mathbb{R}, \forall t' > t, \;\; \forall i \in \mathcal{I}$,
`active` $\implies$ *if* $\mathbb{E}[f(i,t)] \neq \mathbb{E}[f(i,t')]$, *then* $|\mathbb{E}[f(i,t)] - \mathbb{E}[f(i,t')]| \not\perp\!\!\!\perp a \;\; \forall a \in \mathcal{A}, \;\; \forall t \in \mathbb{R}, \forall t' > t, \;\; \forall i \in \mathcal{I}$, *and*
`hybrid` $\implies$ *if* $\mathbb{E}[f(i,t)] \neq \mathbb{E}[f(i,t')]$, *then* $|\mathbb{E}[f(i,t)] - \mathbb{E}[f(i,t')]| \perp\!\!\!\perp a \;\; \exists a \in \mathcal{A}, \;\; \exists t \in \mathbb{R}, \forall t' > t, \;\; \exists i \in \mathcal{I}$ *and* $|\mathbb{E}[f(i,t)] - \mathbb{E}[f(i,t')]| \not\perp\!\!\!\perp a \;\; \exists a \in \mathcal{A}, \;\; \exists t \in \mathbb{R}, \forall t' > t, \;\; \exists i \in \mathcal{I}$.

4.4 Examples of CRL Formulations

We now discuss how the proposed taxonomy can provide a lens to investigate existing formulations. Coupled with the consideration of the driver and scope of the non-stationarity,

the `General CRL Problem` $\mathcal{M}_{\texttt{CRL}}$ can be cast as **a family of MDPs with non-stationary functional forms** (See Proposition 1). Moreover, such a non-stationarity MDP can itself be recast as a **partially-observable MDP** (See Proposition 2) resulting in the following CRL problem formulations:

Proposition 1 (Non-stationary MDPs as `CRL Problems`)**.** *A **non-stationary MDP** is a special type of CRL problem where $\alpha \subseteq \{\mathcal{S}, \mathcal{A}, r, p\}$, the observation function is an appropriate identity matrix $x = \mathbb{I}$, and the observation space is the state space $\mathcal{O} = \mathcal{S}$.*

In this partially observable view, an arbitrarily non-stationary MDP is still unwieldy to solve as it corresponds to an infinite order history dependence. However, in many cases the partial observability view may be preferable for theoretical analysis and crafting more aggressive approaches. This is because this formulation allows for more aggressive solutions as epistemic uncertainty about the future reduces in comparison to assumptions such as piecewise and Lipschitz non-stationarity which always take a local view and never arrive at certainty about the nature of the global problem. It is also amenable to theoretical analysis as the greater system is considered stationary and has well defined long-term behavior.

Proposition 2 (Non-stationary MDP and POMDP Duality)**.** *Any non-stationary MDP $\mathcal{M}$ with $\alpha \subseteq \{p, r\}$ can dually be viewed as an equivalent POMDP $\hat{\mathcal{M}}$. This is because the potentially non-stationary transitions $p(s'|s, a, t)$ and rewards $r(s, a, t)$ of $\mathcal{M}$ can appear stationary with a simple change of variables to create $\hat{\mathcal{M}}$ with observation $\hat{o} = s$ and a full state also based on the unobserved time dependent variable $\hat{s} = [\hat{o}, t]$ so that transitions are $p(\hat{s}'|\hat{s}, a)$ and rewards are $r(\hat{s}, a)$. Furthermore, any POMDP can be seen as a non-stationary MDP because the combination of the observation and time variables always constitutes a uniquely identifying state representation.*

Multi-agent RL Example: As an example of an MDP that can be either viewed as non-stationary or partially observable depending on our perspective of the problem, we will briefly consider a multi-agent environment where each agent is constantly learning. Multi-agent environments are often characterized as a *Markov game* (Littman, 1994) in which a set of agents all interact in the environment and jointly impact the reward and transition dynamics. Formally, the reward dynamics $r(s, a^i, \boldsymbol{a}^{-i})$ and transition dynamics $p(s'|s, a^i, \boldsymbol{a}^{-i})$ depend on the global state of the environment s, a^i which denotes the action of some agent of focus, and $\boldsymbol{a}^{-i}$ denoting the vector of actions for all other agents in the environment. If the policies of the other agents remain constant, we can view this formulation as equivalent to a stationary single agent learning problem such that $r(s, a^i) = \sum_{\boldsymbol{a}^{-i}} \pi(\boldsymbol{a}^{-i}|s) r(s, a^i, \boldsymbol{a}^{-i})$ and $p(s'|s, a^i) = \sum_{\boldsymbol{a}^{-i}} \pi(\boldsymbol{a}^{-i}|s) p(s'|s, a^i, \boldsymbol{a}^{-i})$ without at all considering the role of the other agents in the environment. However, if their policies change over time, the rewards $r(s, a^i)$ and transitions $p(s'|s, a^i)$ will become non-stationary from the perspective of the focal agent. This apparent non-stationarity from the perspective of agent i was recently recast as partial observability using the formalism of an *active Markov game* (Kim, Riemer, Liu, Foerster, Everett, Sun, Tesauro, and How, 2022):

Proposition 3 (Active Markov Games as `CRL Problems`)**.** *Active Markov games define a set of problems that can be viewed as partially observable or time-dependent because of the dependence of p and r on the actions of other agents in the environment. The actions of other agents are not observed at the same time as the state, nor are their time-dependent parameters, nor their update functions.*

As discussed in (Kim et al., 2022), finding the optimal policy in this environment can be viewed as finding the optimal **stationary periodic distribution**. Moreover, if each agent finds its own optimal non-stationary policy, the result is called an **active equilibrium**. If each agent finds its optimal stationary policy, the result is called a **Nash equilibrium**.

4.5 A Unified View

A `General CRL Problem` $\mathcal{M}_{\texttt{CRL}}$ broadly captures existing problem formulations in the literature. The two primary dimensions, **scope** and **driver** of non-stationarity, provide a taxonomy that can characterize different CRL problem formalims. This view results in CRL as a strict generalization of existing settings and therefore offers a unified formulation.

Proposition 4 (CRL as Strict Generalization)**.** *The `CRL Problem` as stated in Definition 1 is a strict generalization of existing categories of problem settings.*[a]

1. *Multi-task MDPs are special CRL problems where $\beta \in \{\texttt{stationary}\}$.*
2. *HiP-MDPs (Doshi-Velez and Konidaris, 2013) are special CRL problems where $\alpha \subseteq \{p, r\}$ and $\beta \in \{\texttt{stationary}\}$.*
3. *HM-MDPs (Choi, Yeung, and Zhang, 2000) and DP-MDPs (Xie et al., 2021) are special CRL problems where $\alpha \subseteq \{p, r\}$ and $\beta \in \{\texttt{stationary}, \texttt{passive}\}$.*
4. *MOMDPs (Ong, Png, Hsu, and Lee, 2010) and active Markov games (Kim et al., 2022) are both special CRL problems where $\alpha \subseteq \{p, r\}$ and $\beta \in \{\texttt{stationary}, \texttt{passive}, \texttt{active}, \texttt{hybrid}\}$.*

a. This view allows us to define categories of settings and not necessarily concrete setting descriptions.

4.6 Assumptions About Shared Structure

Next, it is common to assume shared structure across different components of the `CRL Problem` $\mathcal{M}_{\texttt{CRL}}$. This is because generic MDPs with no assumed structure have a minimax T-step regret bound of $\Omega(\sqrt{|\mathcal{S}||\mathcal{A}|T})$ and sample complexity bound of $\Omega(|\mathcal{S}||\mathcal{A}|)$ (Jaksch, Ortner, and Auer, 2010; Azar, Osband, and Munos, 2017), which implies that no proveable generalization to novel state and action pairs is possible in this setting.

Luckily, many real world applications also possess shared structure either in the observation space, state space, transition dynamics or reward dynamics. For instance, objects or abstract concepts emerge from shared structure across different regions of the world we live

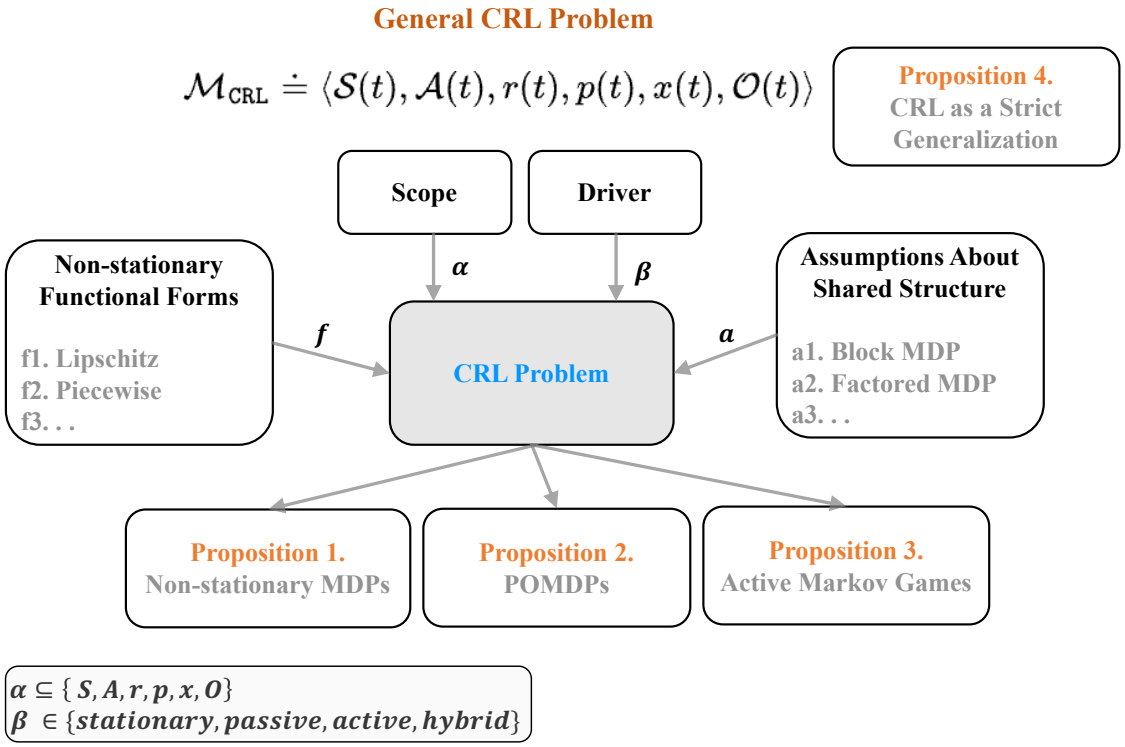

Figure 4: **Taxonomy of Continual RL Formalisms**: Problem formulations in continual RL can be categorized along two primary dimensions: 1) the **scope** of the non-stationarity α and 2) the **driver** of non-stationarity β. Coupled with the scope and the driver of the non-stationarity, assumptions about the **non-stationary functional forms** (f) and **shared structure** (a) can result in different CRL formulations (Propositions 1, 2, and 3). This view offers a unified perspective resulting in continual reinforcement learning as a strict generalization of most of the existing formulations in the literature (Proposition 4).

in. Similarly, laws of physics governing motion, force, or how objects interact are shared across different tasks and applications. We now detail a few popular assumptions that have emerged in the literature. This includes Block MDPs (Du, Krishnamurthy, Jiang, Agarwal, Dudík, and Langford, 2019b):

Definition 6 (Block MDP Assumption): *Each observation o uniquely determines its generating state s. That is, the observation space $\mathcal{O}$ can be partitioned into disjoint blocks $\mathcal{O}_s$, each containing the support of the conditional distribution $x(\cdot|s)$.*

Moreover, in many real-world settings a slowly changing i.e. Lipschitz assumption may not be practical as changes may be large. However, it may still be the case that changes are limited to only a small subset of the causal variables of the underlying MDP (Gumbsch, Butz, and Martius, 2021). These cases can be formalized through the Factored MDP assumption (Kearns and Koller, 1999; Boutilier, Dearden, and Goldszmidt, 2000):

Definition 7 (Factored MDP Assumption): *Each state is composed of n factors $s := s^1, s^2, ..., s^n \in \mathcal{S} \subseteq \mathcal{S}^1 \times \mathcal{S}^2 \times ... \times \mathcal{S}^n$ and each action can be composed of m components $a := a^1, a^2, ..., a^m \in \mathcal{A} \subseteq \mathcal{A}^1 \times \mathcal{A}^2 \times ... \times \mathcal{A}^m$ with sparse causal interactions such that $p(s'|s,a) = \prod_{i=1}^{n} p_i(s^{i'}|Par_i(s,a))$ where Par_i sub-selects from the set of state and action components $\{1, ..., n+m\}$ based on the causal Bayesian network structure. The reward function may also be factored such that $r(s,a) := (1/\ell) \sum_{j=1}^{\ell} r_j(Par_j(s,a))$.*

Moreover, there are a number of other assumptions about share structure that have proven useful in the literature. For example, POMDPs can be restricted to those displaying linear structure as in Cai, Yang, and Wang (2022). Additionally, assumptions about low-rank structure enable a basis for provable generalization while allowing for more complexity than simpler Block MDP models (Agarwal, Kakade, Krishnamurthy, and Sun, 2020; Uehara, Zhang, and Sun, 2021; Zhang, He, Zhou, Zhang, and Gu, 2021a). Similarly, recent work has achieved proveable generalization for problems with low Bellman Eluder dimensions (Jin, Liu, and Miryoosefi, 2021) and low dimensional underlying causal structure (Huang, Feng, Lu, Magliacane, and Zhang, 2021). It is also indeed possible to combine the Block MDP and Factored MDP assumptions into a single common formulation (Misra, Liu, Jin, and Langford, 2021) to relax Block MDPs to handle non-stationary environments (Katt, Oliehoek, and Amato, 2019; Sodhani, Meier, Pineau, and Zhang, 2022). Additionally, Block MDPs have been formalized in settings with multiple domains (Han, Zheng, Chan, Paster, Zhang, and Ba, 2021). Furthermore, a number of settings exist that refine Factored MDPs even further in order to provide additional opportunities for generalization including relational MDPs (Guestrin, Koller, Gearhart, and Kanodia, 2003), first-order MDPs (Boutilier, Reiter, and Price, 2001), object-oriented MDPs (Diuk, Cohen, and Littman, 2008), and MDPs described by linear temporal logic (Vaezipoor, Li, Icarte, and Mcilraith, 2021; Jiang, Bharadwaj, Wu, Shah, Topcu, and Stone, 2021a).

4.7 Literature Review: The Scope of Non-stationarity α

We now review the literature through the lens of the scope of the non-stationarity. 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 i.e. $\alpha \subseteq \{p, r\}$. Meanwhile, it is also possible that the observation function x or action space $\mathcal{A}$ is non-stationary as highlighted in Figure 1 i.e. $\alpha \subseteq \{\mathcal{S}, \mathcal{A}, r, p, x, \mathcal{O}\}$. Very few researchers have actually explored this setting in a fully general manner with every component being potentially non-stationary to date (i.e. the scope). 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 in the literature with a smaller scope 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 including two elements of non-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 should be noted that most of these papers do not explicitly consider the continual RL setting, but the concepts could be 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, Theocharous, Kostas, Jordan, and Thomas, 2019; Langlois and Everitt, 2021; Jain, Kosaka, Kim, and Lim, 2021; Trabucco, Phielipp, and Berseth, 2022) and observation function (Trabucco et al., 2022; Zhang, Lyle, Sodhani, Filos, Kwiatkowska, Pineau, Gal, and Precup, 2020). These sources of non-stationarity may be very important for particular practical applications of continual RL. However, these remain an under explored problem settings in comparison to those focused on transition and reward dynamics. Moreover, to date, we are not aware of any approach that allows for every element of the problem formulation to be non-stationary. It remains an open question to what degree learning is still possible when we allow for non-stationarity in each component of the environment and it is likely that additional assumptions about the nature of the problem will be necessary to develop agents that perform well in this very ambitious setting.

4.8 Literature Review: The Drivers of Non-stationarity

An important point of distinction between different 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.

4.8.1 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 (Wilson et al., 2007; Ammar et al., 2014; Caruana, 1997). 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* (Gábor, Kalmár, and Szepesvári, 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 (Caruana, 1997; French, 1991) across tasks. As noted by (Yu, Kumar, Gupta, Levine, Hausman, and Finn, 2020a), conflicting gradients 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 (Wilson et al., 2007; Ammar et al., 2014) is to model the distribution over tasks and use this distribution as a prior when a new task is seen. 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, Liao, and Carin, 2009).

Recently popular approaches for "meta-learning" (Schmidhuber, 1987; Bengio, Bengio, Cloutier, and Gecsei, 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, Kurth-Nelson, Tirumala, Soyer, Leibo, Munos, Blundell, Kumaran, and Botvinick, 2016; Duan, Schulman, Chen, Bartlett, Sutskever, and Abbeel, 2016; Finn, Abbeel, and Levine, 2017; Mishra, Rohaninejad, Chen, and Abbeel, 2017; Nichol and Schulman, 2018). See (Vilalta and Drissi, 2002; Hospedales, Antoniou, Micaelli, and Storkey, 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 of this problem setup.

4.8.2 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 RL 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, Precup, and Singh, 1999) setting as in (Hadoux, Beynier, and Weng, 2014a). 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; Li, Wang, Jin, Sheng, and Zha, 2021). Another assumption that is common is that of piecewise stationarity (Padakandla, Bhatnagar, 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, Simchi-Levi, and Zhu, 2020; Domingues, Ménard, Pirotta, Kaufmann, and Valko, 2020) and linear function (Touati and Vincent, 2020) MDPs to improve learning in the face of non-stationarity.

4.8.3 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, Barto, and Singh, 2005; Singh, Lewis, and Barto, 2009; Barto, 2013; Achiam and Sastry, 2017) and learning an agent's own curriculum (Schmidhuber, 2013; Justesen and Risi, 2018; Srinivasan, Bahdanau, Chevalier-Boisvert, and Bengio, 2019). See (Portelas, Colas, Weng, Hofmann, and Oudeyer, 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 or scenarios for future play (Wang, Lehman, Clune, and Stanley, 2019; Raileanu, Goldstein, Yarats, Kostrikov, and Fergus, 2021; Chen, Zhang, Xu, Ma, Yang, Song, Wang, and Wu, 2021; Lee and Chung, 2021), their own goals (Florensa, Held, Geng, and Abbeel, 2018; Racaniere, Lampinen, Santoro, Reichert, Firoiu, and Lillicrap, 2019; Sharma, Gupta, Levine, Hausman, and Finn, 2021), relevant information from their past learning (Klink, D'Eramo, Peters, and Pajarinen, 2021; Jiang, Dennis, Parker-Holder, Foerster, Grefenstette, and Rocktäschel, 2021b), and their own opponents (Sukhbaatar, Lin, Kostrikov, Synnaeve, Szlam, and Fergus, 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, Bredeche, Goff, Girard, Coninx, Sigaud, Khamassi, Díaz-Rodríguez, Filliat, Hospedales, 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.

4.8.4 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 (Kim et al., 2022; Foerster, Chen, Al-Shedivat, Whiteson, Abbeel, and Mordatch, 2018a; Foerster, Farquhar, Al-Shedivat, Rocktäschel, Xing, and Whiteson, 2018b; Kim, Liu, Riemer, Sun, Abdulhai, Habibi, Lopez-Cot, Tesauro, and How, 2021a). 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 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.

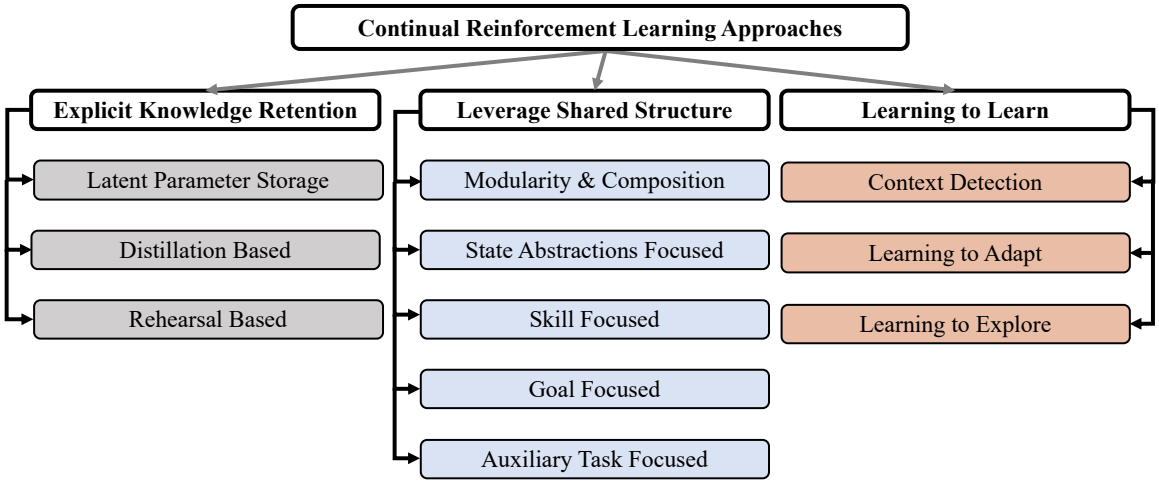

Figure 5: **Taxonomy of Continual RL Approaches**: A diagram illustrating different clusters of approaches for continual RL, highlighting prominent threads of research within each family. Though these categories are not mutually exclusive, we examine each separately for the purpose of this paper.

Figure 6 depicts the distribution of subcategories of approaches for continual RL. The taxonomy of approaches broadly consists of three key families: explicit knowledge retention (18.5%), leveraging shared structure (40.8%), and learning to learn (40.7%). Across subcategories, we note that, unsurprisingly, learning to adapt is predominant. On the other hand, techniques based on latent parameter storage, distillation, state abstraction, and auxiliary tasks constitute only a small fraction of the approaches covered.

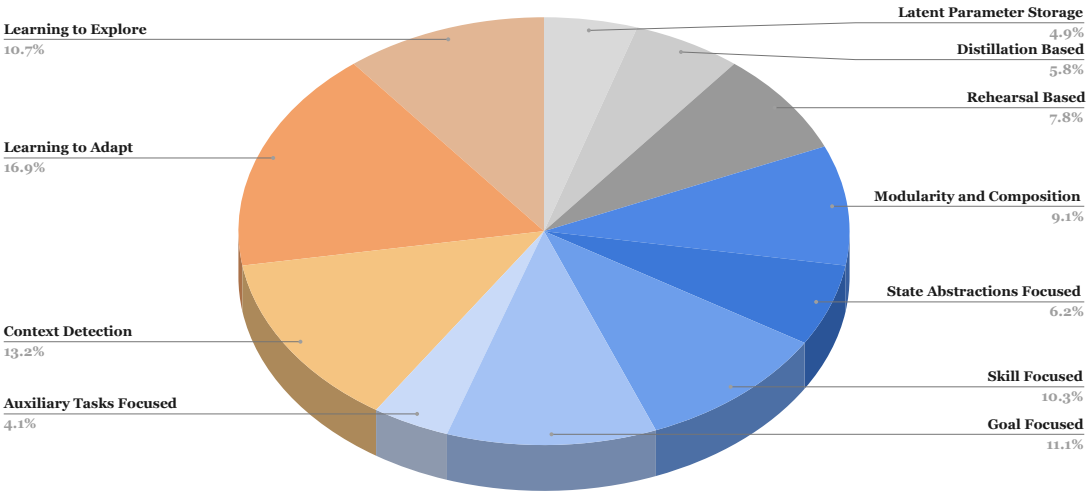

Figure 6: **Popularity of Subcategories of Continual RL Approaches**. This chart tallies the citations in our paper by subcategory. This is not meant to be exhaustive or exact, but rather provide a high-level idea of the relative popularity of these topics in the continual RL literature to date.

5.1 Explicit Knowledge Retention

Due to concerns about agents experiencing catastrophic forgetting when they continually learn as detailed in Sec. 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 for detecting the current task. It also requires significant storage, and lastly, it limits any ability to leverage relevant knowledge learned across tasks. One alternative way to leverage knowledge is through the use of 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, Graepel, and Shawe-Taylor (2016) build on the work of Calandriello, Lazaric, and Restelli (2014) and explicitly model a shared abstraction of the state-action space.

Another approach for leveraging knowledge from previous tasks is to provide the representations of networks trained on previous tasks as inputs for subsequent tasks (Rusu,

Rabinowitz, Desjardins, Soyer, Kirkpatrick, Kavukcuoglu, Pascanu, and Hadsell, 2016). While this strategy directly avoids the problem of catastrophic forgetting, it 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, Pontil, and Romera-Paredes (2016) extracted features for multiple tasks in a single low-dimensional shared representation. D'Eramo, Tateo, Bonarini, Restelli, and Peters (2020); Shi, Azizzadenesheli, O'Connell, Chung, and Yue (2021) 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, a 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; Yu, Kumar, Gupta, Levine, Hausman, and Finn, 2020b; Liu, Liu, Jin, Stone, and Liu, 2021a). This kind of approach generally decreases *plasticity* in areas of the network that were heavily used for past tasks, which can effectively prevent forgetting. Unfortunately, this *stability* may also 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 subnetworks (Cheung, Terekhov, Chen, Agrawal, and Olshausen, 2019; Wortsman, Ramanujan, Liu, Kembhavi, Rastegari, Yosinski, and Farhadi, 2020). Minimizing representational overlap has the positive effect of minimizing forgetting, but it may also minimize the potential for transfer similarly to the single task learning case. As a result, care must be taken while 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 (Buciluǎ, Caruana, and Niculescu-Mizil, 2006; Hinton, Vinyals, and Dean, 2015) from past tasks when learning a new task to prevent catastrophic forgetting as in (Rusu, Colmenarejo, Gulcehre, Desjardins, Kirkpatrick, Pascanu, Mnih, Kavukcuoglu, and Hadsell, 2015; Li and Hoiem, 2016; Riemer, Khabiri, and Goodwin, 2016; Espeholt, Soyer, Munos, Simonyan, Mnih, Ward, Doron, Firoiu, Harley, Dunning, et al., 2018; Schwarz, Luketina, Czarnecki, Grabska-Barwinska, Teh, Pascanu, and Hadsell, 2018; Berseth, Xie, Cernek, and Van de Panne, 2018; Kaplanis, Shanahan, and Clopath, 2019; Traoré, Caselles-Dupré, Lesort, Sun, Cai, Díaz-Rodríguez, and Filliat, 2019; Tirumala, Noh, Galashov, Hasenclever, Ahuja, Wayne, Pascanu, Teh, and Heess, 2019; Igl, Farquhar, Luketina, Böhmer, and Whiteson, 2021; Lan, Pan, Luo, and Mahmood, 2022; Zhang, Wang, Liang, and Yuan, 2022). 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 the 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's 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 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, Zaheer, White, Patterson, and White, 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 (Riemer et al., 2019; Isele and Cosgun, 2018; Rolnick, Ahuja, Schwarz, Lillicrap, and Wayne, 2019; Oh, Shin, Yang, and Hwang, 2021; Henning, Cervera, D'Angelo, Von Oswald, Traber, Ehret, Kobayashi, Grewe, and Sacramento, 2021; Liu, Xue, Pang, Jiang, Xu, and Yu, 2021b; Queeney, Paschalidis, and Cassandras, 2021; Chandak, Niekum, da Silva, Learned-Miller, Brunskill, and Thomas, 2021; Lampinen, Chan, Banino, and Hill, 2021; Venuto, Lau, Precup, and Nachum, 2021; Liotet, Vidaich, Metelli, and Restelli, 2022). 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 generative model (Robins, 1995; Atkinson, McCane, Szymanski, and Robins, 2018; Daniels, Raghavan, Hostetler, Rahman, Sur, Piacentino, and Divakaran, 2022).

Another promising strategy for improving the storage efficiency of replay buffers is to directly learn to compress experiences during learning. This makes the buffer much more efficient, though now only consisting of approximate recollections (Riemer, Klinger, Bouneffouf, and Franceschini, 2017; Caccia, Belilovsky, Caccia, and Pineau, 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, Shanahan, and Clopath, 2018).

5.2 Leveraging Shared Structure

Continually learning agents must learn to solve problems so that they are able to find structure in the world that will benefit them later (Thrun and O'Sullivan, 1996). To achieve this, continual learning agents should reuse aspects of the solutions to previously solved subproblems through function composition (Griffiths, Callaway, Chang, Grant, Krueger, and Lieder, 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. Compositional generalization can be understood as 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, Samejima, Katagiri, and Kawato (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, Gupta, Darrell, Abbeel, and Levine, 2017; Frans, Ho, Chen, Abbeel, and Schulman, 2017; Fernando, Banarse, Blundell, Zwols, Ha, Rusu, Pritzel, and Wierstra, 2017; Rosenbaum, Klinger, and Riemer, 2017; Meyerson and Miikkulainen, 2017; Kirsch, Kunze, and Barber, 2018; Ramachandran and Le, 2018; Chang, Gupta, Levine, and Griffiths, 2018; Liang, Meyerson, and Miikkulainen, 2018; Alet, Lozano-Pérez, and Kaelbling, 2018; Cases, Rosenbaum, Riemer, Geiger, Klinger, Tamkin, Li, Agarwal, Greene, Jurafsky, et al., 2019; Yang, Xu, Wu, and Wang, 2020; Lee, Behpour, and Eaton, 2021; Tseng, Lin, Feng, and Sun, 2021; Mendez and Eaton, 2022; Mendez, van Seijen, and Eaton, 2022; Gaya, Doan, Caccia, Soulier, Denoyer, and Raileanu, 2022) 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, Lim, Ritchie, and Weston, 2017), multi-task learning (Pasunuru and Bansal, 2019), and continual learning (Pasunuru and Bansal, 2019; Xu and Zhu, 2018) 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, Cases, Riemer, and Klinger, 2019) for a detailed survey of the motivations and challenges of these modular and compositional approaches.

5.2.2 State Abstractions Focused

State abstraction (or aggregation) is central to the idea of capturing common structure within various tasks and potentially facilitating positive forward transfer across related tasks. Li, Walsh, and Littman (2006) provide a unified view on the theory of state abstraction and

consider a variety of metrics that can be used to help develop these abstractions. Given MDP M, with its abstracted version $\hat{M}$, the abstraction function $\phi : S \to \hat{S}$ maps states in the ground MDP to 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 state abstractions based on the PAC framework. A PAC state abstraction is defined such that it achieves correct clustering with high probability with respect to a distribution over learning problems (Abel, Arumugam, Lehnert, and Littman, 2018). In the context of lifelong learning, PAC state abstractions are guaranteed to hold with respect to a distribution of tasks, albeit only in tabular settings. With rapid progress in deep RL, recent work (Zhang, Satija, and Pineau, 2018b; François-Lavet, Bengio, Precup, and Pineau, 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, Lipton, Pineda, Azizzadenesheli, Anandkumar, Itti, Pineau, and Furlanello, 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 abstraction if for any action a and any abstract state $\hat{s}$, $\phi_{model}(s_1) = \phi_{model}(s_2)$ implies $R^a_{s_1} = R^a_{s_2}$ and $\sum_{s' \in \phi^{-1}_{model}(\hat{s})} P^a_{s_1,s'} = \sum_{s' \in \phi^{-1}_{model}(\hat{s})} P^a_{s_2,s'}$. In recent work, Zhang et al. (2020); Hansen-Estruch, Zhang, Nair, Yin, and Levine (2022); Ashcraft, Stoler, Ewulum, and Agarwala (2022) connect invariant causal prediction to model-irrelevance state abstractions to learn invariant representations in the Block MDP setting (Du et al., 2019b). These kinds of abstractions are even possible without explicitly modelling rewards (Allen, Parikh, Gottesman, and Konidaris, 2021). Additionally, state abstractions could also be formed on the basis of value equivalence (Grimm, Barreto, Singh, and Silver, 2020; Sokota, Ho, Ahmad, and Kolter, 2021; Cui, Chow, and Ghavamzadeh, 2021) or for improved planning (Curtis, Silver, Tenenbaum, Lozano-Pérez, and Kaelbling, 2022). Furthermore, in large environments with underlying Factored MDP structure, powerful abstractions have also been formed by leveraging context specific independencies that only hold for a subset of states (Abdulhai, Kim, Riemer, Liu, Tesauro, and How, 2022).

5.2.3 Skill Focused

Macro actions (Hauskrecht, Meuleau, Kaelbling, Dean, and Boutilier, 1998) or skills (Thrun and Schwartz, 1995) are approaches for learning that side step the requirement to make decisions at each individual time step. An 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 $R^{\pi_i}_s$. 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:

$$v_*(s) = \max_{\pi_i \in \Pi} \left[R_s^{\pi_i} + \sum_{\tau=1}^{\infty} \gamma^{\tau-1} \sum_{s'} p(S^\tau = s' | S^0 = s, \pi_i) v_*(s') \right] \tag{13}$$

$$q_*(s, \pi_i) = R_s^{\pi_i} + \sum_{\tau=1}^{\infty} \gamma^{\tau-1} \sum_{s'} p(S^\tau = s' | S^0 = s, \pi_i) \max_{\pi_i' \in \Pi} q_*(s', \pi_i') \tag{14}$$

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, 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 determines actions 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 only on that time instance but also on the option ω being followed, 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, Harb, and Precup, 2017; Machado, Bellemare, and Bowling, 2017a; Machado, Rosenbaum, Guo, Liu, Tesauro, and Campbell, 2017b; Harb, Bacon, Klissarov, and Precup, 2018; Khetarpal, Klissarov, Chevalier-Boisvert, Bacon, and Precup, 2020a; Riemer, Cases, Rosenbaum, Liu, and Tesauro, 2020; Klissarov and Precup, 2021), work on discovering options in a multi-task context has shown some promising potential both theoretically (Brunskill and Li, 2014) and empirically (Achiam, Edwards, Amodei, and Abbeel, 2018; Riemer, Liu, and Tesauro, 2018; Igl, Gambardella, He, Nardelli, Siddharth, Böhmer, and Whiteson, 2019). Brunskill and Li (2014) derive sample complexity bounds for option discovery over a distribution of tasks. Mankowitz, Mann, and Mannor (2016) proposed a framework to learn near-optimal skills for a task, composing these 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 for reaching critical states. While options explicitly consider where to initiate a skill, which 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, Gupta, Ibarz, and Levine (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, Trott, Xiong, Socher, Giro-i Nieto, and Torres (2020) propose to learn a set of skills that are state-covering rather than simply diverse. Tessler, Givony, Zahavy, Mankowitz, and Mannor (2017) instead proposed a hierarchical, multi-skill distillation network explicitly allowing *knowledge retention* and *selective transfer* of skills. The Minecraft domain considered in their work 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, Kumar, Tejani, and Isbell (2017); Barreto, Borsa, Hou, Comanici, Aygün, Hamel, Toyama, Mourad, Silver, Precup, et al. (2019); Lu, Grover, Abbeel, and Mordatch (2021) 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, Grover, Abbeel, and Mordatch (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, Arumugam, Lehnert, and Littman (2017); Abel, Umbanhowar, Khetarpal, Arumugam, Precup, and Littman (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 a variety of ways. One popular formulation that is of particular interest in the context of continual RL, is to design algorithms that generalize over goals rather than some other notion of tasks.

An ambitious approach then is to discover general purpose goals without any reward signal as in unsupervised RL (Jin, Krishnamurthy, Simchowitz, and Yu, 2020; Wang, Du, Yang, and Salakhutdinov, 2020a; Zhang, Zhou, and Gu, 2021b). Eysenbach et al. (2018); Gregor, Rezende, and Wierstra (2016); Hausman, Springenberg, Wang, Heess, and Riedmiller (2018); Touati and Ollivier (2021) 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, Oh, and Singh (2018) propose a deep RL adaptation of the work of Kaelbling (1993). Here a goal is defined as a desired raw-pixel observation,

demonstrating empirical benefits in unsupervised learning without a reward signal for improved agent performance on downstream tasks for which goals can be treated as auxiliary tasks. Andreas, Klein, and Levine (2017); Oh, Singh, Lee, and Kohli (2017) assume known information about high-level structural relationships among tasks and similarity between different subtasks, respectively, to enable fast learning across tasks. Meanwhile, Shu, Xiong, and Socher (2017); Geishauser, van Niekerk, Lubis, Heck, Lin, Feng, and Gašić (2022) employ weak supervision from humans to define what skills should be learned in the form of language instructions. Florensa, Held, Geng, and Abbeel (2017); Chen, Yan, Guo, Yang, Su, and Chen (2019) extract 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, Horgan, Gregor, and Silver (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, Žídek, Barreto, Horgan, Hessel, Quan, Oh, van Hasselt, Silver, and Schaul, 2018). Moreover, Zhu, Chang, Zeng, and Tan (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 did not match the original goal of the policy (Andrychowicz, Wolski, Ray, Schneider, Fong, Welinder, McGrew, Tobin, Abbeel, and Zaremba, 2017; Rauber, Ummadisingu, Mutz, and Schmidhuber, 2017; Li, Pinto, and Abbeel, 2020; Kim, Lee, Kim, Ryu, Lee, and Zhang, 2021b; Moro, Likmeta, Prati, Restelli, et al., 2022). 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, Konidaris, Platt, and Saenko, 2017) and modular (Colas, Fournier, Chetouani, Sigaud, and Oudeyer, 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, Li, Gao, He, Chen, Deng, and He, 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, Pascanu, Viola, Soyer, Ballard, Banino, Denil, Goroshin, Sifre, Kavukcuoglu, et al., 2016), pixel control (Jaderberg, Mnih, Czarnecki, Schaul, Leibo, Silver, and Kavukcuoglu, 2016; Hessel, Soyer, Espeholt, Czarnecki, Schmitt, and van Hasselt, 2019), reward prediction (Jaderberg et al., 2016), inverse dynamics prediction (Shelhamer, Mahmoudieh, Argus, and Darrell, 2016), and latent observation prediction (Guo, Pires, Piot, Grill, Altché, Munos, and Azar, 2020).

A natural research direction that follows is to understand how agents can discover useful and general purpose auxiliary tasks that may ease our agent's ability to learn tasks of interest.

Recent work (Veeriah, Hessel, Xu, Rajendran, Lewis, Oh, van Hasselt, Silver, and Singh, 2019) addresses discovering useful auxiliary tasks in the form of the *question functions* of a General Value Function (GVF) (Sutton, Modayil, Degris, Pilarski, and White, 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:

$$v(h_t) = E[c(H_{t+1}) + \gamma(H_{t+1})v(H_{t+1})|H_t = h_t, A_{t+1} \sim \pi(.|h_t)] \tag{15}$$

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 continual non-stationary environments. Moreover, there might be implications of misalignment between auxiliary tasks and the main task at hand (Bellemare, Dabney, Dadashi, Taiga, Castro, Le Roux, Schuurmans, Lattimore, and Lyle, 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 do not 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*, which 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 observations. For a real-world example, we can consider 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 7 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 7, 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., 2014a). For example, the recent work of (Xie et al., 2021) considers a formulation with Markovian transitions between tasks that 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 $p(z'|y, a, z)$.

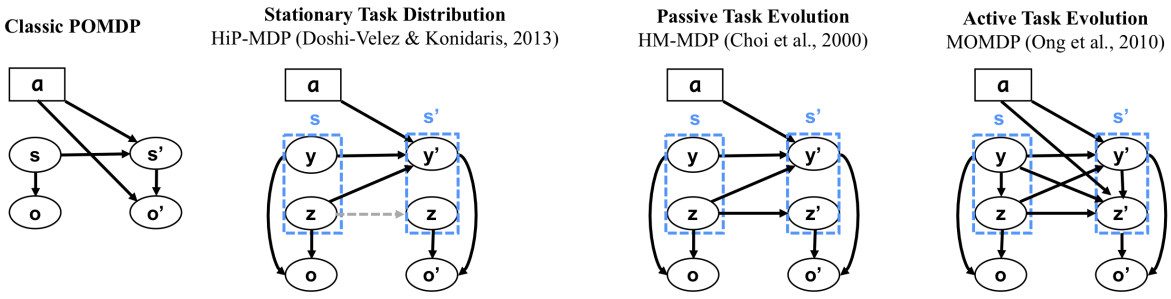

Figure 7: **Useful Classes of Assumptions for POMDPs**: We include from left to right a classic POMDP, a HiP-MDP with stationary *task state* z, a HM-MDP with passive Markovian *task state* evolution, and a MOMDP with an active Markovian *task state* evolution.

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, Shiarlis, Kurin, Hofmann, and Whiteson, 2018; Rakelly, Zhou, Quillen, Finn, and Levine, 2019; Zintgraf, Shiarlis, Igl, Schulze, Gal, Hofmann, and Whiteson, 2019; Humplik, Galashov, Hasenclever, Ortega, Teh, and Heess, 2019; Fakoor, Chaudhari, Soatto, and Smola, 2019; Ortega, Wang, Rowland, Genewein, Kurth-Nelson, Pascanu, Heess, Veness, Pritzel, Sprechmann, et al., 2019; Perez, Such, and Karaletsos, 2020; Dorfman, Shenfeld, and Tamar, 2021; Rigter, Lacerda, and Hawes, 2021; Sodhani, Zhang, and Pineau, 2021; Fu, Tang, Hao, Chen, Feng, Li, and Liu, 2021). However, this framework has also has been readily applicable to more challenging multi-agent learning settings (Da Silva, Basso, Bazzan, and Engel, 2006; Amato, Oliehoek, and Shyu, 2013; Marinescu, Dusparic, and Clarke, 2017; Vezhnevets, Wu,

Leblond, and Leibo, 2019). These approaches have even recently been extended to learn encodings for policies themselves based on limited environment interaction (Harb, Schaul, Precup, and Bacon, 2020; Raileanu, Goldstein, Szlam, and Fergus, 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 (Padakandla et al., 2019; Da Silva et al., 2006; Rosman and Ramamoorthy, 2012; Hadoux, Beynier, and Weng, 2014b; Li, Gu, Zhu, and Zhang, 2019; Kessler, Parker-Holder, Ball, Zohren, and Roberts, 2022; Luo, Jiang, Yu, Zhang, and Zhang, 2022). 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 sample efficiency.

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, Zhan, Taylor, Sucar, and de Cote, 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*. Thus in Bayesian RL, we 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, Zhao, and Schraudolph, 1998) provide a general framework for meta-learning in continuing RL environments. One early promising example was the Success Story Algorithm (SSA) (Schmidhuber, Zhao, and Wiering, 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 on a notion of what is called *meta-training* and *meta-testing*. We highlight some of the key details of this framework in Figure 8. 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., 1992; Bengio, Bengio, and Cloutier, 1990; Schmidhuber, 1992). The *inner loop* is responsible for fast time-scale learning and the *outer loop* is responsible for

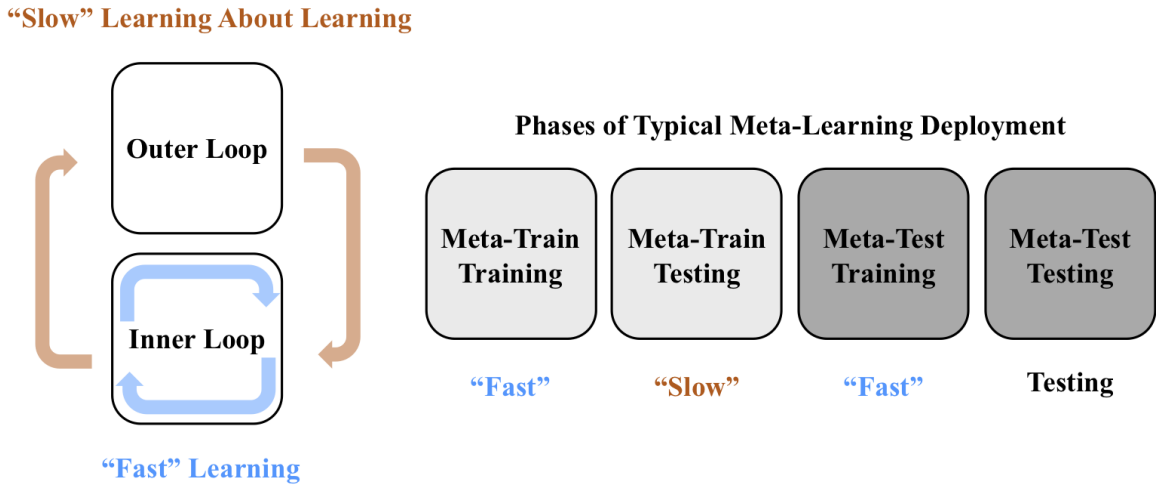

Figure 8: **Phases of Meta-Training & Meta-Testing**. Left: we depict the *inner loop* learning process and the *outer loop* learning process that aims to improve it. Right: we detail the phases of typical meta-learning, including which learning process is used in each phase.

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 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 in *meta-training* by providing 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 be adopted for continual learning deployment at *meta-test* time by training the *inner loop* over a sequence of tasks (Kim et al., 2021a; Javed and White, 2019; Spigler, 2019; Beaulieu, Frati, Miconi, Lehman, Stanley, Clune, and Cheney, 2020; Caccia, Rodriguez, Ostapenko, Normandin, Lin, Caccia, Laradji, Rish, Lacoste, Vazquez, et al., 2020; Co-Reyes, Feng, Berseth, Qui, and Levine, 2021a).

There have been a number of recent improvements on this topic of meta-optimization in multi-task RL (Nichol and Schulman, 2018; Kim, Yoon, Dia, Kim, Bengio, and Ahn, 2018; Rothfuss, Lee, Clavera, Asfour, and Abbeel, 2018; Flennerhag, Moreno, Lawrence, and Damianou, 2018; Nagabandi, Finn, and Levine, 2018; Mendonca, Gupta, Kralev, Abbeel, Levine, and Finn, 2019; Finn, Rajeswaran, Kakade, and Levine, 2019; Lin, Thomas, Yang, and Ma, 2020; Berseth, Zhang, Zhang, Finn, and Levine, 2021; Co-Reyes, Miao, Peng,

Real, Levine, Le, Lee, and Faust, 2021b; Kirsch, Flennerhag, van Hasselt, Friesen, Oh, and Chen, 2022; Wan, Peng, and Gangwani, 2022; Melo, 2022; Nam, Sun, Pertsch, Hwang, and Lim, 2022) and multi-agent RL (Foerster et al., 2018a,1; Kim et al., 2021a; Al-Shedivat, Bansal, Burda, Sutskever, Mordatch, and Abbeel, 2017). 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 (Harb et al., 2020; Sung, Zhang, Xiang, Hospedales, and Yang, 2017; Xu, Cao, and Chen, 2019). 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, Li, Yang, Wang, and Hospedales, 2020; Flennerhag, Schroecker, Zahavy, van Hasselt, Silver, and Singh, 2021) in a true continual RL setting closer to the spirit of self-modifying policies. For example, an approach by Chandak, Theocharous, Shankar, Mahadevan, White, and Thomas (2020a) 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, van Hasselt, and Silver, 2018a; Zahavy, Xu, Veeriah, Hessel, Oh, van Hasselt, Silver, and Singh, 2020; Xu, van Hasselt, Hessel, Oh, Singh, and Silver, 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, Oh, and Singh, 2018; Stadie, Yang, Houthooft, Chen, Duan, Wu, Abbeel, and Sutskever, 2018; Xu, Liu, Zhao, and Peng, 2018b; Houthooft, Chen, Isola, Stadie, Wolski, Ho, and Abbeel, 2018; Yang, Caluwaerts, Iscen, Tan, and Finn, 2019; Zou, Ren, Yan, Su, and Zhu, 2019; Zheng, Oh, Hessel, Xu, Kroiss, van Hasselt, Silver, and Singh, 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; Achiam and Sastry, 2017; Schmidhuber, 1990; Klyubin, Polani, and Nehaniv, 2005; Kaplan and Oudeyer, 2006; Schmidhuber, 2008; Frank, Leitner, Stollenga, Förster, and Schmidhuber, 2014; Mohamed and Rezende, 2015; Bellemare, Srinivasan, Ostrovski, Schaul, Saxton, and Munos, 2016; Houthooft, Chen, Duan, Schulman, De Turck, and Abbeel, 2016; Pathak, Agrawal, Efros, and Darrell, 2017; Karl, Soelch, Becker-Ehmck, Benbouzid, van der Smagt, and Bayer, 2017; Burda, Edwards, Pathak, Storkey, Darrell, and Efros, 2018; Shyam, Jaśkowski, and Gomez, 2019; Liu, Trott, Socher, and Xiong, 2019; Ecoffet, Huizinga, Lehman, Stanley, and Clune, 2019; Sekar, Rybkin, Daniilidis, Abbeel, Hafner, and Pathak, 2020; Steinparz, Schmied, Paischer, Dinu, Patil, Bitto-Nemling, Eghbal-zadeh, and Hochreiter, 2022; Berseth, Geng, Devin, Rhinehart, Finn, Jayaraman, and Levine, 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 in cases where 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 the next task to learn on in deep multi-task RL (Sharma, Jha, Hegde, and Ravindran, 2017), and on learning to elicit valuable information to learn from in non-stationary multi-agent settings (Shu, Xiong, Wu, and Zhu, 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 an open research question. While there has been tremendous progress in the field chasing state-of-the-art results on widely acknowledged benchmarks (Bellemare, Naddaf, Veness, and Bowling, 2013; Brockman, Cheung, Pettersson, Schneider, Schulman, Tang, and Zaremba, 2016) and achieving super human performance (Silver, Hubert, Schrittwieser, Antonoglou, Lai, Guez, Lanctot, Sifre, Kumaran, Graepel, et al., 2018), it is not immediately clear if the aforementioned benchmarks have the sufficient characteristics of desired environments for continual RL. Schaul, van Hasselt, Modayil, White, White, Bacon, Harb, Mourad, Bellemare, and Precup (2018) discuss a long list of open problems in continual RL. We expand on their categorization of *benchmarks* and *metrics*, presenting a discussion on the evaluation of continual RL agents.

6.1 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 (e.g. Mountain Car, Cartpole, and 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 RL over the years.

Specific environments have only allowed us to study and measure agents with respect to specific dimensions. For instance, Taxi (Dietterich, 2000) and four rooms (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, Leibo, Teplyashin, Ward, Wainwright, Küttler, Lefrancq, Green, Valdés, Sadik, et al., 2016) for rich 3D navigation tasks. For more complex agents that can play long-horizon, strategy games accompanied with rich visual observations, several algorithms have also explored Minecraft (Duncan, 2011) and VizDoom (Kempka, Wydmuch, Runc, Toczek, and Jaśkowski, 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 (Machado, Bellemare, Talvitie, Veness, Hausknecht, and Bowling, 2018).

Generating non-stationarity in environments in a principled way is vital for understanding the changes in learning over time. Henderson, Chang, Shkurti, Hansen, Meger, and Dudek (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, Klimov, Hesse, Kim, and Schulman, 2019a), which is part of the larger Procgen (Cobbe, Hesse, Hilton, and Schulman, 2019b) suite of games, leverage procedural generation to create a large set of train and test environments with subtle differences. Additionally, Osband, Doron, Hessel, Aslanides, Sezener, Saraiva, McKinney, Lattimore, Szepezvari, Singh, 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. Furthermore, Yu, Quillen, He, Julian, Hausman, Finn, and Levine (2020c) 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 configuration parameters.

The recent advancement of RL algorithms is in large part due to the emergence of these environments with an increased focus on testing generalization. However, most environments require a clear distinction between train-test boundaries and are dependent on a well defined notion of tasks, 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. 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 formalisms (see Sec. 4). Additionally, desirable characteristics of such benchmarks would include the ability to: 1) train in a progressive and 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) learn causal relationships including affordances associated with objects. Ahmed, Träuble, Goyal, Neitz, Wüthrich, Bengio, Schölkopf, and Bauer (2020) proposed CausalWorld, 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 through embodiment (Kiefer and Trumpp, 2012; Kiela, Bulat, Vero, and Clark, 2016; Lake, Ullman, Tenenbaum, and Gershman, 2017; Bisk, Holtzman, Thomason, Andreas, Bengio, Chai, Lapata, Lazaridou, May, Nisnevich, et al., 2020). Embodied experience grounds learning in intuitive physics and causal reasoning (Lake et al., 2017). Khetarpal, Sodhani, Chandar, and Precup (2018a) discusses similar recommendations for moving towards environments for lifelong agents and also highlights the benefits of embodied cognition.

Jelly Bean World (Platanios, Saparov, and Mitchell, 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.[1]

More recent advances targeted at evaluating continual learning capabilities have addressed some of the aforementioned desiderata of CRL benchmarks. Nekoei, Badrinaaraayanan, Courville, and Chandar (2021) introduced a Hanabi based multi-agent lifelong learning testbed and evaluated many multi-agent RL approaches. Powers, Xing, Kolve, Mottaghi, and Gupta (2021) introduced CORA, a unification of benchmarks, metrics, and baselines incorporating ALE, NetHack, Procgen and CHORES. Johnson, Nguyen, Schreurs, Ewulum, Ashcraft, Fendley, Baker, New, and Vallabha (2022) presented a highly configurable, Unity-based environment for testing continual and lifelong learning systems. Goel, Tatiya, Scheutz, and Sinapov (2021) on the other hand, is a benchmark focused on exploring continual sources of novelty.

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, Jordan, Theocharous, White, and Thomas, 2020b) as case studies for when the nature of RL problems have inbuilt non-stationarity to account for.

6.2 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 reward 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 that 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. Furthermore, for a continual learner, it is vital to measure core metrics, for example, forward and backward transfer (or interference), skill reusability, and skill composition.

A promising approach is to incorporate the idea of *auxiliary metrics* (Schaul et al., 2018) to measure an agent's intelligence with *probe questions* that function in a similar way 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. Specifically, the core desired capabilities of a continual learner can be tested with probe questions.

1. 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 within our taxonomy of formalisms, where the nature of the non-stationarity allows for these lines to be blurry.

6.3 Towards Broader Evaluation Criteria for Continual RL

At the intersection of metrics (what to measure) and domains (how to measure)[2] 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 9 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 continual RL agents. Ideally, proper empirical analysis would result in a measure of the behavior along different dimensions of probe-metrics as shown in the Fig. 9.

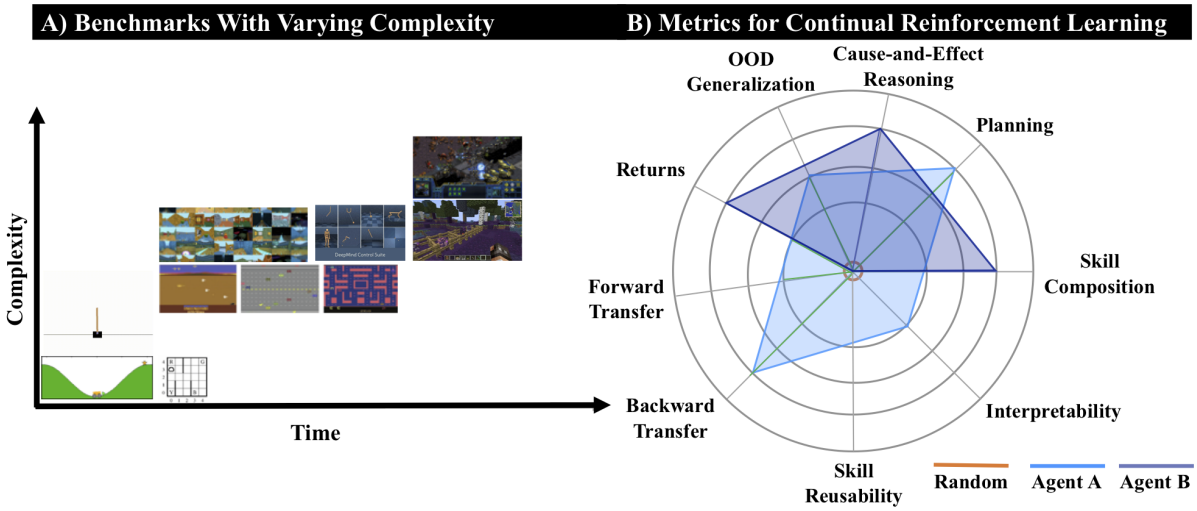

Figure 9: **Evaluating Continual Reinforcement Learning Agents**. A) Depicts the evolution of domains and benchmarks over time commonly used in RL. B) Depicts key metrics for evaluating continual RL agents in the style of bsuite. Such a framework should also offer a knob controlling the degree and nature 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

2. We acknowledge that evaluation of deep RL agents faces several challenges pertaining to reproducibility (see (Henderson, Islam, Bachman, Pineau, Precup, and Meger, 2017b; Khetarpal, Ahmed, Cianflone, Islam, and Pineau, 2018b)). This is even more reason for the community to move towards standardized evaluation benchmarks for continual RL.

current learning greatly interferes with its ability on previous experiences, the agent is experiencing catastrophic forgetting.

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. Much of the existing work uses qualitative analysis 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 behavior 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 for 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*: Designing probes for cause-and-effect analysis will allow a quantitative measure of how well continual RL agents are learning the underlying rules and objects of an environment. 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 first discuss potential connections between continual RL research and findings in the neuroscience community. We 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 human-like continual learners. How humans have the innate ability to perform long-term decision making without a task reward, or rather even delayed sparse rewards, remains largely a mystery. In fact, it is unclear even what utility (Samuelson, 1937) humans optimize for. Nonetheless, there has still been interesting work at the intersection of neuroscience and RL. For example, Hassabis, Kumaran, Summerfield, and Botvinick (2017) provides a survey on neuroscience inspired AI, and Niv (2009) details evidence that RL happens in the human brain. We 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, Lillicrap, Abramson, Wu, Mirza, Carnevale, Ahuja, and Wayne (2019) introduces a temporal value transport algorithm, where the agent uses specific memories to credit past actions.

The neuroscience literature can give us more insight into the nature of task interference that humans experience. For example, Detre, Natarajan, Gershman, and Norman (2013) demonstrate that human memories that are highly activated during continual learning are less likely to be forgotten later. Additionally, Sagiv, Musslick, Niv, and Cohen (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, McNaughton, and O'Reilly, 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, Ajemian, D'Ausilio, Moorman, and Bizzi (2013) demonstrate the role that noisy computations in the human brain may have in avoiding catastrophic forgetting, and Dohare, Mahmood, and Sutton (2021) also demonstrate that noisy computations can improve plasticity.

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, Frank, and Fellous, 2010; Starkweather, Babayan, Uchida, and Gershman, 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, Kurth-Nelson, Kumaran, Tirumala, Soyer, Leibo, Hassabis, and Botvinick, 2018). Moreover, it has been demonstrated that stress can play a big factor in selectively modulating the reward signal for humans (Berghorst, Bogdan, Frank, and Pizzagalli, 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, McDougle, and Collins, 2020).

Leveraging Memory. The neuroscience literature related to the interplay between learning and memory formation is another potential source of insight for building better continual RL agents. 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, Otto, Daw, and Norman, 2018; Vikbladh, Meager, King, Blackmon, Devinsky, Shohamy, Burgess, and Daw, 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, Salakhutdinov, and Levine, 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, 1995,9; Frean and Robins, 1999). Schuck and Niv (2019) also suggest 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, Daw, and Shohamy, 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, Nassar, Dillon, and Frank, 2019), which also mirrors findings in deep RL (Schaul, Quan, Antonoglou, and Silver, 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, Daw, Dayan, and O'Doherty, 2010; Daw, Gershman, Seymour, Dayan, and Dolan, 2011; Doll, Simon, and Daw, 2012). In fact, Langdon, Sharpe, Schoenbaum,

and Niv (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, Otto, Phelps, and Daw, 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, Russek, Cheong, Botvinick, Daw, and Gershman, 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, Miall, and Kawato (1998). Their simulation results corroborate findings associated with fMRI data (Imamizu, 1997; Imamizu, Miyauchi, Tamada, Sasaki, Takino, PuÈtz, Yoshioka, and Kawato, 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, Niv, and Dayan (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 Corroborating 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 learning is either uncertainty-aware or Bayesian in nature (Niv, Duff, and Dayan, 2005; Fleming and Daw, 2017; Babayan, Uchida, and Gershman, 2018; Lowet, Zheng, Matias, Drugowitsch, and Uchida, 2020). There is also significant evidence that substantial hierarchical structure is present in the human learning process (Botvinick, Niv, and Barto, 2009; Gershman, Pesaran, and Daw, 2009; Frank and Badre, 2012; Badre and Frank, 2012; Diuk, Schapiro, Córdova, Ribas-Fernandes, Niv, and Botvinick, 2013; Botvinick and Weinstein, 2014; Solway, Diuk, Córdova, Yee, Barto, Niv, and Botvinick, 2014; Ribas-Fernandes, Shahnazian, Holroyd, and Botvinick, 2019). Moreover, Niv, Daniel, Geana, Gershman, Leong, Radulescu, and Wilson (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. It has been demonstrated that human RL attempts to identify the causal structure in the task at hand (Gershman, 2017; Gershman, Norman, and Niv, 2015).

7.2 Challenges and Open Problems

Due to the lack of a concrete definition and the extremely general nature of CRL, there is still significant unexplored potential to be addressed by future research. 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.[3]

Inductive Biases. There are, without a doubt, many different perspectives within the AI community on how much and to what degree we should embed inductive biases in our agents. 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 such assumptions

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., 2018a; Zahavy et al., 2020), end-to-end skill discovery (Bacon et al., 2017), discovering RL algorithms (Kirsch, van Steenkiste, and Schmidhuber, 2019; Oh, Hessel, Czarnecki, Xu, van Hasselt, Singh, and Silver, 2020), learning what objective to learn (Xu et al., 2020), and other approaches moving more and more towards open-ended learning (Wang, Lehman, Rawal, Zhi, Li, Clune, and Stanley, 2020b).

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. 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, Ahmed, Comanici, Abel, and Precup, 2020b) 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).

3. We refer the reader to (Schaul et al., 2018) for more discussion on open problems in continual RL.

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 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 still has 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, Sutskever, and Hinton, 2012; Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, et al., 2015) was followed by a plethora of research on visualizing (Zeiler and Fergus, 2014) and understanding (Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus, 2013) the nature of 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 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, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, 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 at the problem will solve everything, nor is it 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, McCandlish,

Henighan, Brown, Chess, Child, Gray, Radford, Wu, and Amodei, 2020; Sharma and Kaplan, 2020; Henighan, Kaplan, Katz, Chen, Hesse, Jackson, Jun, Brown, Dhariwal, Gray, et al., 2020). Exploring the relevant *scaling laws* for continual RL will be a potentially fruitful topic to understand more clearly moving forward.

Learning in the Presence of Other Learning Agents. As highlighted in Proposition 3, a key continual RL scenario is when an agent must learn in a stationary environment that contains other agents that are also learning. The non-stationarity experienced in these domains is a key consideration of multi-agent RL (MARL). See Hernandez-Leal, Kartal, and Taylor (2019) for a comprehensive survey on MARL. In these settings, even the correct solution concept can be difficult to define (Kim et al., 2022; Littman, 2001; Wang and Sandholm, 2002; Bowling, 2004; Greenwald and Hall, 2003; Zinkevich, Greenwald, and Littman, 2006; Letcher, Foerster, Balduzzi, Rocktäschel, and Whiteson, 2019). That said, the existence of other agents in the environment could be a blessing if treated appropriately. For example, agents can teach other agents as they learn so that each learns to better perform tasks that are required of them (Torrey and Taylor, 2013; Omidshafiei, Kim, Liu, Tesauro, Riemer, Amato, Campbell, and How, 2019; Kim, Liu, Omidshafiei, Lopez-Cot, Riemer, Habibi, Tesauro, Mourad, Campbell, and How, 2020). Moreover, agents can learn to shape the learning of other agents and thus the nature of the non-stationarity in their environment (Kim et al., 2022; Foerster et al., 2018a,1; Kim et al., 2021a; Letcher et al., 2019; Zhang and Lesser, 2010). MARL is quite far from use in industrial applications and is still largely in its infancy as a research direction. However, it is clear that developing agents capable of acting effectively in large systems with other interacting agents will be an important step towards enabling continual RL agents to integrate with our society and navigate the real world outside of simulation.

8. Conclusion

In summary, we would like to highlight final considerations. While finding the right set of assumptions is an important meta challenge of continual RL, we would like 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 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, Betteridge, Kisiel, Settles, Hruschka, and Mitchell (2010) is one example deployed in a true never-ending continual learning fashion, albeit in a very controlled setting. More recently, Blenderbot (Shuster, Xu, Komeili, Ju, Smith, Roller, Ung, Chen, Arora, Lane, et al., 2022) is also a step in this direction in the context of conversational agents. 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 serves as a useful resource for the continual RL research community that will help us collectively understand and eventually achieve this lofty goal.

Acknowledgements

KK and MR contributed equally to this work. We would like to thank Takuya Ito and Martin Klissarov for providing valuable feedback. We would also like to thank Joelle Pineau for insightful comments on an earlier draft of the paper. We really appreciate the feedback we received from the JAIR reviewers, which helped improve the contributions of this work. Finally, we would like to give a special thank you to Anna Riemer for editing our final manuscript. We acknowledge that this work originated as a class project undertaken in the graduate-level course on Continual Learning: Towards "Broad" AI (IFT-6760B) at Mila, Montreal. IR also acknowledges support from the Canada CIFAR AI Chair Program and the Canada Excellence Research Chairs Program.

Appendix A. The Relationship Between Continual Reinforcement Learning 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:

$$J_{\text{continuing}}(\theta) = -\mathbb{E}_{p(x,y^*)}\left[\sum_{k=0}^{T} \ell(Y_{t+k}^*, \hat{Y}_{t+k} = f(X_{t+k}; \theta))\right] \tag{16}$$

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:

$$J_{\text{stationary}}(\theta) = -\sum_{k=0}^{T} \mathbb{E}_{p(x,y^*)}\left[\ell(Y_{t+k}^*, \hat{Y}_{t+k} = f(X_{t+k}; \theta))\right] \tag{17}$$

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:

$$J_{\text{SGD}}(\theta) = -\mathbb{E}_{p(x,y^*)}\left[\ell(Y_t^*, \hat{Y}_t = f(X_t; \theta))\right] \tag{18}$$

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:

$$\begin{aligned} J_{\text{non-stationary}}(\theta) &= -\mathbb{E}_{p(x,y^*,t)}\left[\sum_{k=0}^{T} \ell(Y^*_{t+k}, \hat{Y}_{t+k} = f(X_{t+k}; \theta))\right] \\ &= -\mathbb{E}_{p(x,y^*,t)}\left[\ell(Y^*_t, \hat{Y}_t = f(X_t; \theta)) + \sum_{k=1}^{T} \ell(Y^*_{t+k}, \hat{Y}_{t+k} = f(X_{t+k}; \theta))\right] \end{aligned} \tag{19}$$

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.

References

Emmanuel Bengio, Joelle Pineau, and Doina Precup. Interference and generalization in temporal difference learning. *arXiv preprint arXiv:2003.06350*, 2020.

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

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2019.

Martin Mundt, Yong Won Hong, Iuliia Pliushch, and Visvanathan Ramesh. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *arXiv preprint arXiv:2009.01797*, 2020.

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

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

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

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

Annie Xie, James Harrison, and Chelsea Finn. Deep reinforcement learning amidst continual structured non-stationarity. In *International Conference on Machine Learning*, pages 11393–11403. PMLR, 2021.

Matthew Riemer, Sharath Chandra Raparthy, Ignacio Cases, Gopeshh Subbaraj, Maximilian Puelma Touzel, and Irina Rish. Continual learning in environments with polynomial mixing times. *Advances in Neural Information Processing Systems*, 2022.

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

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

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

Haitham Bou Ammar, Eric Eaton, Paul Ruvolo, and Matthew Taylor. Online multi-task learning for policy gradient methods. In *International Conference on Machine Learning*, pages 1206–1214, 2014.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Jesse Farebrother, Marlos C Machado, and Michael Bowling. Generalization and regularization in dqn. *arXiv preprint arXiv:1810.00123*, 2018.

Alex Nichol, Vicki Pfau, Christopher Hesse, Oleg Klimov, and John Schulman. Gotta learn fast: A new benchmark for generalization in rl. *arXiv preprint arXiv:1804.03720*, 2018.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.

Niels Justesen, Ruben Rodriguez Torrado, Philip Bontrager, Ahmed Khalifa, Julian Togelius, and Sebastian Risi. Illuminating generalization in deep reinforcement learning through procedural level generation. *arXiv preprint arXiv:1806.10729*, 2018.

Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018a.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. *arXiv preprint arXiv:1812.02341*, 2018.

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

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

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continuum learning. *Advances in Neural Information Processing Systems*, 2017.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *ICLR*, 2019.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *ICLR*, 2019.

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

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In *International Conference on Machine Learning*, pages 7447–7458. PMLR, 2021.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *ICML*, pages 157–163. Morgan Kaufmann Publishers Inc., 1994. ISBN 1-55860-335-2. URL `http://dl.acm.org/citation.cfm?id=3091574.3091594`.

Dong Ki Kim, Matthew Riemer, Miao Liu, Jakob N Foerster, Michael Everett, Chuangchuang Sun, Gerald Tesauro, and Jonathan P How. Influencing long-term behavior in multiagent reinforcement learning. *Advances in Neural Information Processing Systems*, 2022.

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

Samuel PM Choi, Dit-Yan Yeung, and Nevin L Zhang. Hidden-mode markov decision processes for nonstationary sequential decision making. In *Sequence Learning*, pages 264–287. Springer, 2000.

Sylvie CW Ong, Shao Wei Png, David Hsu, and Wee Sun Lee. Planning under uncertainty for robotic tasks with mixed observability. *The International Journal of Robotics Research*, 29(8):1053–1068, 2010.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Simon S Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient rl with rich observations via latent state decoding. *arXiv preprint arXiv:1901.09018*, 2019b.

Christian Gumbsch, Martin V Butz, and Georg Martius. Sparsely changing latent states for prediction and planning in partially observable domains. *Advances in Neural Information Processing Systems*, 34:17518–17531, 2021.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pages 740–747, 1999.

Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. Stochastic dynamic programming with factored representations. *Artificial Intelligence*, 121(1):49–107, 2000. ISSN 0004-3702. https://doi.org/10.1016/S0004-3702(00)00033-3. URL `https://www.sciencedirect.com/science/article/pii/S0004370200000333`.

Qi Cai, Zhuoran Yang, and Zhaoran Wang. Sample-efficient reinforcement learning for pomdps with linear function approximations. *arXiv preprint arXiv:2204.09787*, 2022.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Weitong Zhang, Jiafan He, Dongruo Zhou, Amy Zhang, and Quanquan Gu. Provably efficient representation learning in low-rank markov decision processes. *arXiv preprint arXiv:2106.11935*, 2021a.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. Adarl: What, where, and how to adapt in transfer reinforcement learning. *arXiv preprint arXiv:2107.02729*, 2021.

Dipendra Misra, Qinghua Liu, Chi Jin, and John Langford. Provable rich observation reinforcement learning with combinatorial latent states. In *International Conference on Learning Representations*, 2021.

Sammie Katt, Frans A Oliehoek, and Christopher Amato. Bayesian reinforcement learning in factored pomdps. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 7–15, 2019.

Shagun Sodhani, Franziska Meier, Joelle Pineau, and Amy Zhang. Block contextual mdps for continual learning. In *Learning for Dynamics and Control Conference*, pages 608–623. PMLR, 2022.

Beining Han, Chongyi Zheng, Harris Chan, Keiran Paster, Michael Zhang, and Jimmy Ba. Learning domain invariant representations in goal-conditioned block mdps. *Advances in Neural Information Processing Systems*, 34:764–776, 2021.

Carlos Guestrin, Daphne Koller, Chris Gearhart, and Neal Kanodia. Generalizing plans to new environments in relational mdps. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pages 1003–1010, 2003.

Craig Boutilier, Raymond Reiter, and Bob Price. Symbolic dynamic programming for first-order mdps. In *IJCAI*, volume 1, pages 690–700, 2001.

Carlos Diuk, Andre Cohen, and Michael L Littman. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 240–247, 2008.

Pashootan Vaezipoor, Andrew C Li, Rodrigo A Toro Icarte, and Sheila A Mcilraith. Ltl2action: Generalizing ltl instructions for multi-task rl. In *International Conference on Machine Learning*, pages 10497–10508. PMLR, 2021.

Yuqian Jiang, Suda Bharadwaj, Bo Wu, Rishi Shah, Ufuk Topcu, and Peter Stone. Temporal-logic-based reward shaping for continuing reinforcement learning tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7995–8003, 2021a.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip S Thomas. Learning action representations for reinforcement learning. *arXiv preprint arXiv:1902.00183*, 2019.

Eric D Langlois and Tom Everitt. How rl agents behave when their actions are modified. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11586–11594, 2021.

Ayush Jain, Norio Kosaka, Kyung-Min Kim, and Joseph J Lim. Know your action set: Learning action relations for reinforcement learning. In *International Conference on Learning Representations*, 2021.

Brandon Trabucco, Mariano Phielipp, and Glen Berseth. Anymorph: Learning transferable polices by inferring agent morphology. In *International Conference on Machine Learning*, pages 21677–21691. PMLR, 2022.

Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block mdps. *arXiv preprint arXiv:2003.06016*, 2020.

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

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

Robert 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.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020a.

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

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

Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan 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.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

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

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

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

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

Emmanuel Hadoux, Aurélie Beynier, and Paul Weng. Solving hidden-semi-markov-mode markov decision problems. In *International Conference on Scalable Uncertainty Management*, pages 176–189. Springer, 2014a.

Erwan Lecarpentier and Emmanuel 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.

Wenhao Li, Xiangfeng Wang, Bo Jin, Junjie Sheng, and Hongyuan Zha. Dealing with non-stationarity in marl via trust-region decomposition. In *International Conference on Learning Representations*, 2021.

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

Wang Chi Cheung, David Simchi-Levi, and Ruihao 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.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. A kernel-based approach to non-stationary reinforcement learning in metric spaces. *arXiv preprint arXiv:2007.05078*, 2020.

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

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

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

Satinder Singh, Richard L Lewis, and Andrew 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.

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

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

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

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

Anirudh Srinivasan, Dzmitry Bahdanau, Maxime Chevalier-Boisvert, and Yoshua Bengio. Automated curriculum generation for policy gradients from demonstrations. *arXiv preprint arXiv:1912.00444*, 2019.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic curriculum learning for deep rl: A short survey. *arXiv preprint arXiv:2003.04664*, 2020.

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

Roberta Raileanu, Maxwell Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. Automatic data augmentation for generalization in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:5402–5415, 2021.

Jiayu Chen, Yuanxin Zhang, Yuanfan Xu, Huimin Ma, Huazhong Yang, Jiaming Song, Yu Wang, and Yi Wu. Variational automatic curriculum learning for sparse-reward cooperative multi-agent problems. *Advances in Neural Information Processing Systems*, 34:9681–9693, 2021.

Suyoung Lee and Sae-Young Chung. Improving generalization in meta-rl with imaginary tasks from latent dynamics mixture. *Advances in Neural Information Processing Systems*, 34:27222–27235, 2021.

Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pages 1515–1528, 2018.

Sébastien Racaniere, Andrew K Lampinen, Adam Santoro, David P Reichert, Vlad Firoiu, and Timothy P Lillicrap. Automated curricula through setter-solver interactions. *arXiv preprint arXiv:1909.12892*, 2019.

Archit Sharma, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Autonomous reinforcement learning via subgoal curricula. *Advances in Neural Information Processing Systems*, 34:18474–18486, 2021.

Pascal Klink, Carlo D'Eramo, Jan Peters, and Joni Pajarinen. Boosted curriculum reinforcement learning. In *International Conference on Learning Representations*, 2021.

Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Replay-guided adversarial environment design. *Advances in Neural Information Processing Systems*, 34:1884–1897, 2021b.

Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. Intrinsic motivation and automatic curricula via asymmetric self-play. *arXiv preprint arXiv:1703.05407*, 2017.

Stephane Doncieux, Nicolas Bredeche, Léni Le Goff, Benoît Girard, Alexandre Coninx, Olivier Sigaud, Mehdi Khamassi, Natalia Díaz-Rodríguez, David Filliat, Timothy Hospedales, et al. Dream architecture: a developmental approach to open-ended learning in robotics. *arXiv preprint arXiv:2005.06223*, 2020.

Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor 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.

Jakob Foerster, Gregory Farquhar, Maruan Al-Shedivat, Tim Rocktäschel, Eric P Xing, and Shimon Whiteson. Dice: The infinitely differentiable monte-carlo estimator. *arXiv preprint arXiv:1802.05098*, 2018b.

Dong Ki Kim, Miao Liu, Matthew D Riemer, Chuangchuang Sun, Marwa Abdulhai, Golnaz Habibi, Sebastian Lopez-Cot, Gerald Tesauro, and Jonathan How. A policy gradient algorithm for learning to learn in multiagent reinforcement learning. In *International Conference on Machine Learning*, pages 5541–5550. PMLR, 2021a.

Diana Borsa, Thore Graepel, and John Shawe-Taylor. Learning shared representations in multi-task reinforcement learning. *arXiv preprint arXiv:1603.02041*, 2016.

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

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The Benefit of Multitask Representation Learning. *JMLR*, page 32, 2016.

Carlo D'Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, and Jan Peters. Sharing knowledge in multi-task deep reinforcement learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rkgpv2VFvr`.

Guanya Shi, Kamyar Azizzadenesheli, Michael O'Connell, Soon-Jo Chung, and Yisong Yue. Meta-adaptive nonlinear control: Theory and algorithms. *Advances in Neural Information Processing Systems*, 34:10013–10025, 2021.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020b.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021a.

Brian Cheung, Alexander Terekhov, Yubei Chen, Pulkit Agrawal, and Bruno Olshausen. Superposition of many models into one. In *Advances in Neural Information Processing Systems*, pages 10868–10877, 2019.

Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. Supermasks in superposition. *arXiv preprint arXiv:2006.14769*, 2020.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541. ACM, 2006.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

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

Matthew Riemer, Elham Khabiri, and Richard Goodwin. Representation stability as a regularizer for improved text analytics transfer learning. *arXiv preprint arXiv:1704.03617*,

2016.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

Jonathan Schwarz, Jelena Luketina, Wojciech M Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. *arXiv preprint arXiv:1805.06370*, 2018.

Glen Berseth, Cheng Xie, Paul Cernek, and Michiel Van de Panne. Progressive reinforcement learning with distillation for multi-skilled motion control. *arXiv preprint arXiv:1802.04765*, 2018.

Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Policy consolidation for continual reinforcement learning. *arXiv preprint arXiv:1902.00255*, 2019.

René Traoré, Hugo Caselles-Dupré, Timothée Lesort, Te Sun, Guanghang Cai, Natalia Díaz-Rodríguez, and David Filliat. Discorl: Continual reinforcement learning via policy distillation. *arXiv preprint arXiv:1907.05855*, 2019.

Dhruva Tirumala, Hyeonwoo Noh, Alexandre Galashov, Leonard Hasenclever, Arun Ahuja, Greg Wayne, Razvan Pascanu, Yee Whye Teh, and Nicolas Heess. Exploiting hierarchy for learning and transfer in kl-regularized rl. *arXiv preprint arXiv:1903.07438*, 2019.

Maximilian Igl, Gregory Farquhar, Jelena Luketina, JW Böhmer, and Shimon Whiteson. Transient non-stationarity and generalisation in deep reinforcement learning. In *9th International Conference on Learning Representations*, 2021.

Qingfeng Lan, Yangchen Pan, Jun Luo, and A Rupam Mahmood. Memory-efficient reinforcement learning with knowledge consolidation. *arXiv preprint arXiv:2205.10868*, 2022.

Tiantian Zhang, Xueqian Wang, Bin Liang, and Bo Yuan. Catastrophic interference in reinforcement learning: A solution based on context division and knowledge distillation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

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

Yangchen Pan, Muhammad Zaheer, Adam White, Andrew Patterson, and Martha White. Organizing experience: a deeper look at replay mechanisms for sample-based planning in continuous state domains. *arXiv preprint arXiv:1806.04624*, 2018.

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

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

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, pages 348–358, 2019.

Youngmin Oh, Jinwoo Shin, Eunho Yang, and Sung Ju Hwang. Model-augmented prioritized experience replay. In *International Conference on Learning Representations*, 2021.

Christian Henning, Maria Cervera, Francesco D'Angelo, Johannes Von Oswald, Regina Traber, Benjamin Ehret, Seijin Kobayashi, Benjamin F Grewe, and João Sacramento. Posterior meta-replay for continual learning. *Advances in Neural Information Processing*

Systems, 34:14135–14149, 2021.

Xu-Hui Liu, Zhenghai Xue, Jingcheng Pang, Shengyi Jiang, Feng Xu, and Yang Yu. Regret minimization experience replay in off-policy reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17604–17615, 2021b.

James Queeney, Yannis Paschalidis, and Christos G Cassandras. Generalized proximal policy optimization with sample reuse. *Advances in Neural Information Processing Systems*, 34:11909–11919, 2021.

Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34:27475–27490, 2021.

Andrew Lampinen, Stephanie Chan, Andrea Banino, and Felix Hill. Towards mental time travel: a hierarchical memory for reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34:28182–28195, 2021.

David Venuto, Elaine Lau, Doina Precup, and Ofir Nachum. Policy gradients incorporating the future. *arXiv preprint arXiv:2108.02096*, 2021.

Pierre Liotet, Francesco Vidaich, Alberto Maria Metelli, and Marcello Restelli. Lifelong hyper-policy optimization with multiple importance sampling regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7525–7533, 2022.

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

Craig Atkinson, Brendan McCane, Lech Szymanski, and Anthony Robins. Pseudo-rehearsal: Achieving deep reinforcement learning without catastrophic forgetting. *arXiv preprint arXiv:1812.02464*, 2018.

Zachary Daniels, Aswin Raghavan, Jesse Hostetler, Abrar Rahman, Indranil Sur, Michael Piacentino, and Ajay Divakaran. Model-free generative replay for lifelong reinforcement learning: Application to starcraft-2. *arXiv preprint arXiv:2208.05056*, 2022.

Matthew Riemer, Tim Klinger, Djallel Bouneffouf, and Michele Franceschini. Scalable recollections for continual lifelong learning. *arXiv preprint arXiv:1711.06761*, 2017.

Lucas Caccia, Eugene Belilovsky, Massimo Caccia, and Joelle Pineau. Online learned continual compression with stacked quantization module. *arXiv preprint arXiv:1911.08019*, 2019.

Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Continual reinforcement learning with complex synapses. *arXiv preprint arXiv:1802.07239*, 2018.

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

Thomas L Griffiths, Frederick Callaway, Michael B Chang, Erin Grant, Paul M Krueger, and Falk Lieder. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29:24–30, 2019.

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

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

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

Kenji Doya, Kazuyuki Samejima, Ken-ichi Katagiri, and Mitsuo Kawato. Multiple model-based reinforcement learning. *Neural computation*, 14(6):1347–1369, 2002.

Coline Devin, Abhishek Gupta, Trevor Darrell, Pieter Abbeel, and Sergey 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.

Kevin Frans, Jonathan Ho, Xi Chen, Pieter Abbeel, and John Schulman. Meta learning shared hierarchies. *arXiv preprint arXiv:1710.09767*, 2017.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

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

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

Louis Kirsch, Julius Kunze, and David Barber. Modular networks: learning to decompose neural computation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2414–2423, 2018.

Prajit Ramachandran and Quoc V Le. Diversity and depth in per-example routing models. *ICLR*, 2018.

Michael B Chang, Abhishek Gupta, Sergey Levine, and Thomas L Griffiths. Automatically composing representation transformations as a means for generalization. *arXiv preprint arXiv:1807.04640*, 2018.

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

Ferran Alet, Tomás Lozano-Pérez, and Leslie P Kaelbling. Modular meta-learning. *arXiv preprint arXiv:1806.10166*, 2018.

Ignacio Cases, Clemens Rosenbaum, Matthew Riemer, Atticus Geiger, Tim Klinger, Alex Tamkin, Olivia Li, Sandhini Agarwal, Joshua D Greene, Dan 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.

Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-task reinforcement learning with soft modularization. *arXiv preprint arXiv:2003.13661*, 2020.

Seungwon Lee, Sima Behpour, and Eric Eaton. Sharing less is more: Lifelong learning in deep networks with selective layer transfer. In *International Conference on Machine Learning*, pages 6065–6075. PMLR, 2021.

Wei-Cheng Tseng, Jin-Siang Lin, Yao-Min Feng, and Min Sun. Toward robust long range policy transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9958–9966, 2021.

Jorge A Mendez and Eric Eaton. How to reuse and compose knowledge for a lifetime of tasks: A survey on continual learning and functional composition. *arXiv preprint*

arXiv:2207.07730, 2022.

Jorge A Mendez, Harm van Seijen, and Eric Eaton. Modular lifelong reinforcement learning via neural composition. *arXiv preprint arXiv:2207.00429*, 2022.

Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, and Roberta Raileanu. Building a subspace of policies for scalable continual learning. In *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*, 2022.

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

Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Smash: one-shot model architecture search through hypernetworks. *arXiv preprint arXiv:1708.05344*, 2017.

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

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

Clemens Rosenbaum, Ignacio Cases, Matthew Riemer, and Tim Klinger. Routing networks and the challenges of modular and compositional computation. *arXiv preprint arXiv:1904.12774*, 2019.

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

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

Amy Zhang, Harsh Satija, and Joelle Pineau. Decoupling dynamics and reward for transfer learning. *arXiv preprint arXiv:1804.10689*, 2018b.

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

Amy Zhang, Zachary C Lipton, Luis Pineda, Kamyar Azizzadenesheli, Anima Anandkumar, Laurent Itti, Joelle Pineau, and Tommaso Furlanello. Learning causal state representations of partially observable environments. *arXiv preprint arXiv:1906.10437*, 2019.

Philippe Hansen-Estruch, Amy Zhang, Ashvin Nair, Patrick Yin, and Sergey Levine. Bisimulation makes analogies in goal-conditioned reinforcement learning. *arXiv preprint arXiv:2204.13060*, 2022.

C Chace Ashcraft, Benjamin Stoler, Chigozie Ewulum, and Susama Agarwala. Structural similarity for improved transfer in reinforcement learning. *arXiv preprint arXiv:2207.13813*, 2022.

Cameron Allen, Neev Parikh, Omer Gottesman, and George Konidaris. Learning markov state abstractions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:8229–8241, 2021.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:5541–5552, 2020.

Samuel Sokota, Caleb Y Ho, Zaheen Ahmad, and J Zico Kolter. Monte carlo tree search with iteratively refining state abstractions. *Advances in Neural Information Processing*

Systems, 34:18698–18709, 2021.

Brandon Cui, Yinlam Chow, and Mohammad Ghavamzadeh. Control-aware representations for model-based reinforcement learning. In *International Conference on Learning Representations*, 2021.

Aidan Curtis, Tom Silver, Joshua B Tenenbaum, Tomás Lozano-Pérez, and Leslie Kaelbling. Discovering state and action abstractions for generalized task and motion planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5377–5384, 2022.

Marwa Abdulhai, Dong Ki Kim, Matthew Riemer, Miao Liu, Gerald Tesauro, and Jonathan P How. Context-specific representation abstraction for deep option learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5959–5967, 2022.

Milos Hauskrecht, Nicolas Meuleau, Leslie Pack Kaelbling, Thomas Dean, and Craig 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.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Marios C Machado, Marc G Bellemare, and Michael 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.

Marlos C Machado, Clemens Rosenbaum, Xiaoxiao Guo, Miao Liu, Gerald Tesauro, and Murray Campbell. Eigenoption discovery through the deep successor representation. *arXiv preprint arXiv:1710.11089*, 2017b.

Jean Harb, Pierre-Luc Bacon, Martin Klissarov, and Doina Precup. When waiting is not an option: Learning options with a deliberation cost. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Khimya Khetarpal, Martin Klissarov, Maxime Chevalier-Boisvert, Pierre-Luc Bacon, and Doina Precup. Options of interest: Temporal abstraction with interest functions. *arXiv preprint arXiv:2001.00271*, 2020a.

Matthew Riemer, Ignacio Cases, Clemens Rosenbaum, Miao Liu, and Gerald 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.

Martin Klissarov and Doina Precup. Flexible option learning. *Advances in Neural Information Processing Systems*, 34:4632–4646, 2021.

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

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*, 2018.

Matthew Riemer, Miao Liu, and Gerald Tesauro. Learning abstract options. *NIPS*, 2018.

Maximilian Igl, Andrew Gambardella, Jinke He, Nantas Nardelli, N Siddharth, Wendelin Böhmer, and Shimon Whiteson. Multitask soft option learning. *arXiv preprint arXiv:1904.01033*, 2019.

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

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giro-i Nieto, and Jordi Torres. Explore, discover and learn: Unsupervised discovery of state-covering skills. *arXiv preprint arXiv:2002.03647*, 2020.

Chen Tessler, Shahar Givony, Tom Zahavy, Daniel J Mankowitz, and Shie Mannor. A deep hierarchical approach to lifelong learning in minecraft. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Himanshu Sahni, Saurabh Kumar, Farhan Tejani, and Charles Isbell. Learning to compose skills. *arXiv preprint arXiv:1711.11289*, 2017.

André Barreto, Diana Borsa, Shaobo Hou, Gheorghe Comanici, Eser Aygün, Philippe Hamel, Daniel Toyama, Shibl Mourad, David Silver, Doina Precup, et al. The option keyboard: Combining skills in reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 13031–13041, 2019.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Reset-free lifelong learning with skill-space planning. In *International Conference on Learning Representations*, 2021.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Reset-free lifelong learning with skill-space planning. *NeurIPS Workshop on Deep Reinforcement Learning*, 2020.

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael L Littman. Toward good abstractions for lifelong learning. In *NIPS Workshop on Hierarchical Reinforcement Learning*, 2017.

David Abel, Nate Umbanhowar, Khimya Khetarpal, Dilip Arumugam, Doina Precup, and Michael Littman. Value preserving state-action abstractions. In *International Conference on Artificial Intelligence and Statistics*, pages 1639–1650. PMLR, 2020.

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

Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826, 2020a.

Weitong Zhang, Dongruo Zhou, and Quanquan Gu. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34:1582–1593, 2021b.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

Karol Hausman, Jost Tobias Springenberg, Ziyu Wang, Nicolas Heess, and Martin Riedmiller. Learning an embedding space for transferable robot skills. *ICLR*, 2018.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.

Vivek Veeriah, Junhyuk Oh, and Satinder Singh. Many-goals reinforcement learning. *arXiv preprint arXiv:1806.09605*, 2018.

Jacob Andreas, Dan Klein, and Sergey 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.

Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet 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.

Tianmin Shu, Caiming Xiong, and Richard Socher. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. *arXiv preprint arXiv:1712.07294*, 2017.

Christian Geishauser, Carel van Niekerk, Nurul Lubis, Michael Heck, Hsien-Chin Lin, Shutong Feng, and Milica Gašić. Dynamic dialogue policy transformer for continual reinforcement learning. *arXiv preprint arXiv:2204.05928*, 2022.

Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. *arXiv preprint arXiv:1705.06366*, 2017.

Dagui Chen, Qi Yan, Shangqi Guo, Zhile Yang, Xin Su, and Feng Chen. Learning effective subgoals with multi-task hierarchical reinforcement learning. *Scaling-Up Reinforcement Learning (SURL) Workshop. URl: http://surl. tirl. info/proceedings/SURL-2019_paper_10. pdf*, 2019.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pages 1312–1320, 2015a.

Daniel J Mankowitz, Augustin Žídek, André Barreto, Dan Horgan, Matteo Hessel, John Quan, Junhyuk Oh, Hado van Hasselt, David Silver, and Tom Schaul. Unicorn: Continual learning with a universal, off-policy agent. *arXiv preprint arXiv:1802.08294*, 2018.

Fengda Zhu, Xiaojun Chang, Runhao Zeng, and Mingkui Tan. Continual reinforcement learning with diversity exploration and adversarial self-correction. *arXiv preprint arXiv:1906.09205*, 2019.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, pages 5048–5058, 2017.

Paulo Rauber, Avinash Ummadisingu, Filipe Mutz, and Juergen Schmidhuber. Hindsight policy gradients. *arXiv preprint arXiv:1711.06006*, 2017.

Alexander C Li, Lerrel Pinto, and Pieter Abbeel. Generalized hindsight for reinforcement learning. *arXiv preprint arXiv:2002.11708*, 2020.

Kibeom Kim, Min Whoo Lee, Yoonsung Kim, JeHwan Ryu, Minsu Lee, and Byoung-Tak Zhang. Goal-aware cross-entropy for multi-target reinforcement learning. *Advances in Neural Information Processing Systems*, 34:2783–2795, 2021b.

Lorenzo Moro, Amarildo Likmeta, Enrico Prati, Marcello Restelli, et al. Goal-directed planning via hindsight experience replay. In *International Conference on Learning Representations*, pages 1–16, 2022.

Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. Learning multi-level hierarchies with hindsight. *arXiv preprint arXiv:1712.00948*, 2017.

Cédric Colas, Pierre Fournier, Mohamed Chetouani, Olivier Sigaud, and Pierre-Yves Oudeyer. Curious: intrinsically motivated modular multi-goal reinforcement learning. In *International conference on machine learning*, pages 1331–1340, 2019.

Xiujun Li, Lihong Li, Jianfeng Gao, Xiaodong He, Jianshu Chen, Li Deng, and Ji He. Recurrent reinforcement learning: a hybrid approach. *arXiv preprint arXiv:1509.03044*, 2015.

Guillaume Lample and Devendra Singh Chaplot. Playing fps games with deep reinforcement learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3796–3803, 2019.

Evan Shelhamer, Parsa Mahmoudieh, Max Argus, and Trevor Darrell. Loss is its own reward: Self-supervision for reinforcement learning. *arXiv preprint arXiv:1612.07307*, 2016.

Daniel Guo, Bernardo Avila Pires, Bilal Piot, Jean-bastien Grill, Florent Altché, Rémi Munos, and Mohammad Gheshlaghi Azar. Bootstrap latent-predictive representations for multitask reinforcement learning. *arXiv preprint arXiv:2004.14646*, 2020.

Vivek Veeriah, Matteo Hessel, Zhongwen Xu, Janarthanan Rajendran, Richard L Lewis, Junhyuk Oh, Hado P van Hasselt, David Silver, and Satinder Singh. Discovery of useful questions as auxiliary tasks. In *Advances in Neural Information Processing Systems*, pages 9306–9317, 2019.

Richard S Sutton, Joseph Modayil, Michael Delp Thomas Degris, Patrick M Pilarski, and Adam White. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 2017.

Marc Bellemare, Will Dabney, Robert Dadashi, Adrien Ali Taiga, Pablo Samuel Castro, Nicolas Le Roux, Dale Schuurmans, Tor Lattimore, and Clare Lyle. A geometric perspective on optimal representations for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4360–4371, 2019.

Luisa M Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. *arXiv preprint arXiv:1810.03642*, 2018.

Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. *arXiv preprint arXiv:1903.08254*, 2019.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.

Jan Humplik, Alexandre Galashov, Leonard Hasenclever, Pedro A Ortega, Yee Whye Teh, and Nicolas Heess. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*, 2019.

Rasool Fakoor, Pratik Chaudhari, Stefano Soatto, and Alexander J Smola. Meta-q-learning. *arXiv preprint arXiv:1910.00125*, 2019.

Pedro A Ortega, Jane X Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Heess, Joel Veness, Alex Pritzel, Pablo Sprechmann, et al. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

Christian F Perez, Felipe Petroski Such, and Theofanis Karaletsos. Generalized hidden parameter mdps transferable model-based rl in a handful of trials. *arXiv preprint arXiv:2002.03072*, 2020.

Ron Dorfman, Idan Shenfeld, and Aviv Tamar. Offline meta reinforcement learning–identifiability challenges and effective data collection strategies. *Advances in Neural Information Processing Systems*, 34:4607–4618, 2021.

Marc Rigter, Bruno Lacerda, and Nick Hawes. Risk-averse bayes-adaptive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1142–1154, 2021.

Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In *International Conference on Machine Learning*, pages 9767–9779. PMLR, 2021.

Haotian Fu, Hongyao Tang, Jianye Hao, Chen Chen, Xidong Feng, Dong Li, and Wulong Liu. Towards effective context for meta-reinforcement learning: an approach based on contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7457–7465, 2021.

Bruno C Da Silva, Eduardo W Basso, Ana LC Bazzan, and Paulo M Engel. Dealing with non-stationary environments using context detection. In *Proceedings of the 23rd international conference on Machine learning*, pages 217–224, 2006.

Christopher Amato, Frans A Oliehoek, and Eric Shyu. Scalable bayesian reinforcement learning for multiagent pomdps. In *Citeseer*. Citeseer, 2013.

Andrei Marinescu, Ivana Dusparic, and Siobhán 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.

Alexander Sasha Vezhnevets, Yuhuai Wu, Remi Leblond, and Joel Leibo. Options as responses: Grounding behavioural hierarchies in multi-agent rl. *arXiv preprint arXiv:1906.01470*, 2019.

Jean Harb, Tom Schaul, Doina Precup, and Pierre-Luc Bacon. Policy evaluation networks. *arXiv preprint arXiv:2002.11833*, 2020.

Roberta Raileanu, Max Goldstein, Arthur Szlam, and Rob Fergus. Fast adaptation via policy-dynamics value functions. *arXiv preprint arXiv:2007.02879*, 2020.

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

Emmanuel Hadoux, Aurélie Beynier, and Paul Weng. Sequential decision-making under non-stationary environments via sequential change-point detection. In *Learning over multiple contexts (LMCE)*, 2014b.

Siyuan Li, Fangda Gu, Guangxiang Zhu, and Chongjie 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.

Samuel Kessler, Jack Parker-Holder, Philip Ball, Stefan Zohren, and Stephen J Roberts. Same state, different task: Continual reinforcement learning without interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages

7143–7151, 2022.

Fan-Ming Luo, Shengyi Jiang, Yang Yu, Zongzhang Zhang, and Yi-Feng Zhang. Adapt to environment sudden changes by learning a context sensitive policy. In *Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event*, 2022.

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

Sultan Javed Majeed and Marcus Hutter. On q-learning convergence for non-markov decision processes. In *IJCAI*, pages 2546–2552, 2018.

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

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

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

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

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

Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. *Learning a synaptic learning rule*. Citeseer, 1990.

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

Khurram Javed and Martha White. Meta-learning representations for continual learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, pages 1818–1828. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/8458-meta-learning-representations-for-continual-learning.pdf`.

Giacomo 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`.

Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. Learning to continually learn. *arXiv preprint arXiv:2002.09571*, 2020. URL `https://arxiv.org/abs/2002.09571`.

Massimo Caccia, Pau Rodriguez, Oleksiy Ostapenko, Fabrice Normandin, Min Lin, Lucas Caccia, Issam Laradji, Irina Rish, Alexande Lacoste, David Vazquez, et al. Online fast adaptation and knowledge accumulation: a new approach to continual learning. *Advances in Neural Information Processing Systems*, 2020.

John D Co-Reyes, Sarah Feng, Glen Berseth, Jie Qui, and Sergey Levine. Accelerating online reinforcement learning via model-based meta-learning. In *Learning to Learn-Workshop at ICLR 2021*, 2021a.

Taesup Kim, Jaesik Yoon, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *arXiv preprint arXiv:1806.03836*, 2018.

Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, and Pieter Abbeel. Promp: Proximal meta-policy search. *arXiv preprint arXiv:1810.06784*, 2018.

Sebastian Flennerhag, Pablo G Moreno, Neil D Lawrence, and Andreas Damianou. Transferring knowledge across learning processes. *arXiv preprint arXiv:1812.01054*, 2018.

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

Russell Mendonca, Abhishek Gupta, Rosen Kralev, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Guided meta-policy search. In *Advances in Neural Information Processing Systems*, pages 9653–9664, 2019.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. *arXiv preprint arXiv:1902.08438*, 2019.

Zichuan Lin, Garrett Thomas, Guangwen Yang, and Tengyu Ma. Model-based adversarial meta-reinforcement learning. *arXiv preprint arXiv:2006.08875*, 2020.

Glen Berseth, Zhiwei Zhang, Grace Zhang, Chelsea Finn, and Sergey Levine. Comps: Continual meta policy search. *arXiv preprint arXiv:2112.04467*, 2021.

John D Co-Reyes, Yingjie Miao, Daiyi Peng, Esteban Real, Sergey Levine, Quoc V Le, Honglak Lee, and Aleksandra Faust. Evolving reinforcement learning algorithms. *arXiv preprint arXiv:2101.03958*, 2021b.

Louis Kirsch, Sebastian Flennerhag, Hado van Hasselt, Abram Friesen, Junhyuk Oh, and Yutian Chen. Introducing symmetries to black box meta reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7202–7210, 2022.

Michael Wan, Jian Peng, and Tanmay Gangwani. Hindsight foresight relabeling for meta-reinforcement learning. In *International Conference on Learning Representations*, 2022.

Luckeciano C Melo. Transformers are meta-reinforcement learners. In *International Conference on Machine Learning*, pages 15340–15359. PMLR, 2022.

Taewook Nam, Shao-Hua Sun, Karl Pertsch, Sung Ju Hwang, and Joseph J Lim. Skill-based meta-reinforcement learning. In *International Conference on Learning Representations*, 2022.

Maruan Al-Shedivat, Trapit Bansal, Yuri Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. *arXiv preprint arXiv:1710.03641*, 2017.

Flood Sung, Li Zhang, Tao Xiang, Timothy Hospedales, and Yongxin Yang. Learning to learn: Meta-critic networks for sample efficient learning. *arXiv preprint arXiv:1706.09529*, 2017.

Zhixiong Xu, Lei Cao, and Xiliang Chen. Learning to learn: Hierarchical meta-critic networks. *IEEE Access*, 7:57069–57077, 2019.

Wei Zhou, Yiying Li, Yongxin Yang, Huaimin Wang, and Timothy M Hospedales. Online meta-critic learning for off-policy actor-critic methods. *arXiv preprint arXiv:2003.05334*, 2020.

Sebastian Flennerhag, Yannick Schroecker, Tom Zahavy, Hado van Hasselt, David Silver, and Satinder Singh. Bootstrapped meta-learning. *arXiv preprint arXiv:2109.04504*, 2021.

Yash Chandak, Georgios Theocharous, Shiv Shankar, Sridhar Mahadevan, Martha White, and Philip S Thomas. Optimizing for the future in non-stationary mdps. *arXiv preprint arXiv:2005.08158*, 2020a.

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

Zhongwen Xu, Hado P van Hasselt, and David Silver. Meta-gradient reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2402–2413, 2018a.

Tom Zahavy, Zhongwen Xu, Vivek Veeriah, Matteo Hessel, Junhyuk Oh, Hado van Hasselt, David Silver, and Satinder Singh. Self-tuning deep reinforcement learning. *arXiv preprint arXiv:2002.12928*, 2020.

Zhongwen Xu, Hado P van Hasselt, Matteo Hessel, Junhyuk Oh, Satinder Singh, and David Silver. Meta-gradient reinforcement learning with an objective discovered online. *Advances in Neural Information Processing Systems*, 33, 2020.

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

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

Bradly C Stadie, Ge Yang, Rein Houthooft, Xi Chen, Yan Duan, Yuhuai Wu, Pieter Abbeel, and Ilya Sutskever. Some considerations on learning to explore via meta-reinforcement learning. *arXiv preprint arXiv:1803.01118*, 2018.

Tianbing Xu, Qiang Liu, Liang Zhao, and Jian Peng. Learning to explore with meta-policy gradient. *arXiv preprint arXiv:1803.05044*, 2018b.

Rein Houthooft, Yuhua Chen, Phillip Isola, Bradly Stadie, Filip Wolski, OpenAI Jonathan Ho, and Pieter Abbeel. Evolved policy gradients. In *Advances in Neural Information Processing Systems*, pages 5400–5409, 2018.

Yuxiang Yang, Ken Caluwaerts, Atil Iscen, Jie Tan, and Chelsea 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.

Haosheng Zou, Tongzheng Ren, Dong Yan, Hang Su, and Jun Zhu. Reward shaping via meta-learning. *arXiv preprint arXiv:1901.09330*, 2019.

Zeyu Zheng, Junhyuk Oh, Matteo Hessel, Zhongwen Xu, Manuel Kroiss, Hado van Hasselt, David Silver, and Satinder Singh. What can learned intrinsic rewards capture? *arXiv preprint arXiv:1912.05500*, 2019.

Jiirgen Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent n eu al networks for dynamic reinforcement learning and planning in non-stationary environments. *Citeseer*, 1990.

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

Frederic Kaplan and Pierre-Yves Oudeyer. Curiosity-driven development. In *Proceedings of the International Workshop on Synergistic Intelligence Dynamics.* Citeseer, 2006.

Jürgen 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.

Mikhail Frank, Jürgen Leitner, Marijn Stollenga, Alexander Förster, and Jürgen Schmidhuber. Curiosity driven reinforcement learning for motion planning on humanoids. *Frontiers in neurorobotics*, 7:25, 2014.

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

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in neural information processing systems*, pages 1471–1479, 2016.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor 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.

Maximilian Karl, Maximilian Soelch, Philip Becker-Ehmck, Djalel Benbouzid, Patrick van der Smagt, and Justin Bayer. Unsupervised real-time control through variational empowerment. *arXiv preprint arXiv:1710.05101*, 2017.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International Conference on Machine Learning*, pages 5779–5788, 2019.

Hao Liu, Alexander Trott, Richard Socher, and Caiming Xiong. Competitive experience replay. *arXiv preprint arXiv:1902.00528*, 2019.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. *arXiv preprint arXiv:2005.05960*, 2020.

Christian Steinparz, Thomas Schmied, Fabian Paischer, Marius-Constantin Dinu, Vihang Patil, Angela Bitto-Nemling, Hamid Eghbal-zadeh, and Sepp Hochreiter. Reactive exploration to cope with non-stationarity in lifelong reinforcement learning. *arXiv preprint arXiv:2207.05742*, 2022.

Glen Berseth, Daniel Geng, Coline Manon Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. Smirl: Surprise minimizing reinforcement learning in unstable environments. In *International Conference on Learning Representations*, 2020.

Sahil Sharma, Ashutosh Jha, Parikshit Hegde, and Balaraman Ravindran. Learning to multi-task by active sampling. *arXiv preprint arXiv:1702.06053*, 2017.

Tianmin Shu, Caiming Xiong, Ying Nian Wu, and Song-Chun Zhu. Interactive agent modeling by learning to probe. *arXiv preprint arXiv:1810.00510*, 2018.

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

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

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

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

Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016.

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

Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech 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.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61: 523–562, 2018.

Peter Henderson, Wei-Di Chang, Florian Shkurti, Johanna Hansen, David Meger, and Gregory Dudek. Benchmark environments for multitask learning in continuous domains. *arXiv preprint arXiv:1708.04352*, 2017a.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International Conference on Machine Learning*, pages 1282–1289. PMLR, 2019a.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *arXiv preprint arXiv:1912.01588*, 2019b.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepezvari, Satinder Singh, et al. Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*, 2019.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pages 1094–1100, 2020c.

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wüthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

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

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

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

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

Khimya Khetarpal, Shagun Sodhani, Sarath Chandar, and Doina Precup. Environments for lifelong reinforcement learning. *arXiv preprint arXiv:1811.10732*, 2018a.

Emmanouil Antonios Platanios, Abulhair Saparov, and Tom Mitchell. Jelly bean world: A testbed for never-ending learning. *arXiv preprint arXiv:2002.06306*, 2020.

Hadi Nekoei, Akilesh Badrinaaraayanan, Aaron Courville, and Sarath Chandar. Continuous coordination as a realistic scenario for lifelong learning. In *International Conference on Machine Learning*, pages 8016–8024. PMLR, 2021.

Sam Powers, Eliot Xing, Eric Kolve, Roozbeh Mottaghi, and Abhinav Gupta. Cora: Benchmarks, baselines, and metrics as a platform for continual reinforcement learning agents. *arXiv preprint arXiv:2110.10067*, 2021.

Erik C Johnson, Eric Q Nguyen, Blake Schreurs, Chigozie S Ewulum, Chace Ashcraft, Neil M Fendley, Megan M Baker, Alexander New, and Gautam K Vallabha. L2explorer: A lifelong reinforcement learning assessment environment. *arXiv preprint arXiv:2203.07454*, 2022.

Shivam Goel, Gyan Tatiya, Matthias Scheutz, and Jivko Sinapov. Novelgridworlds: A benchmark environment for detecting and adapting to novelties in open worlds. *International Foundation for Autonomous Agents and Multiagent Systems, AAMAS*, 2021.

Yash Chandak, Scott M. Jordan, Georgios Theocharous, Martha White, and Philip S. Thomas. Towards safe policy improvement for non-stationary mdps, 2020b.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *arXiv preprint arXiv:1709.06560*, 2017b.

Khimya Khetarpal, Zafarali Ahmed, Andre Cianflone, Riashat Islam, and Joelle Pineau. Re-evaluate: Reproducibility in evaluating reinforcement learning algorithms. *Reproducibility in Machine Learning Workshop, (ICML)*, 2018b.

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

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245 – 258, 2017.

ISSN 0896-6273. https://doi.org/10.1016/j.neuron.2017.06.011. URL http://www.sciencedirect.com/science/article/pii/S0896627317305093.

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

Chia-Chun Hung, Timothy Lillicrap, Josh Abramson, Yan Wu, Mehdi Mirza, Federico Carnevale, Arun Ahuja, and Greg Wayne. Optimizing agent behavior over long time scales by transporting value. *Nature communications*, 10(1):1–12, 2019.

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

Yotam Sagiv, Sebastian Musslick, Yael Niv, and Jonathan D Cohen. Efficiency of learning vs. processing: Towards a normative theory of multitasking. *arXiv preprint arXiv:2007.03124*, 2020.

James L McClelland, Bruce L McNaughton, and Randall 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.

Robert Ajemian, Alessandro D'Ausilio, Helene Moorman, and Emilio 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.

Shibhansh Dohare, A Rupam Mahmood, and Richard S Sutton. Continual backprop: Stochastic gradient descent with persistent randomness. *arXiv preprint arXiv:2108.06325*, 2021.

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

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

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

Clara Kwon Starkweather, Benedicte M Babayan, Naoshige Uchida, and Samuel J Gershman. Dopamine reward prediction errors reflect hidden-state inference across time. *Nature neuroscience*, 20(4):581–589, 2017.

Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860–868, 2018.

Lisa H Berghorst, Ryan Bogdan, Michael J Frank, and Diego A Pizzagalli. Acute stress selectively reduces reward sensitivity. *Frontiers in human neuroscience*, 7:133, 2013.

Milena Rmus, Samuel McDougle, and Anne Collins. The role of executive function in shaping reinforcement learning. *Current Opinion in Behavioral Sciences*, 2020.

Anne GE Collins and Michael 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.

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

Ida Momennejad, A Ross Otto, Nathaniel D Daw, and Kenneth A Norman. Offline replay supports planning in human reinforcement learning. *Elife*, 7:e32548, 2018.

Oliver M Vikbladh, Michael R Meager, John King, Karen Blackmon, Orrin Devinsky, Daphna Shohamy, Neil Burgess, and Nathaniel D Daw. Hippocampal contributions to model-based planning and spatial memory. *Neuron*, 102(3):683–693, 2019.

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

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

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

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

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

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

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

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

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015b.

Jan Gläscher, Nathaniel Daw, Peter Dayan, and John 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.

Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69 (6):1204–1215, 2011.

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

Angela J Langdon, Melissa J Sharpe, Geoffrey Schoenbaum, and Yael Niv. Model-based predictions for dopamine. *Current Opinion in Neurobiology*, 49:1–7, 2018.

Dylan A Simon and Nathaniel 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.

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

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

Ida Momennejad, Evan M Russek, Jin H Cheong, Matthew M Botvinick, Nathaniel Douglass Daw, and Samuel J Gershman. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680–692, 2017.

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

Hiroshi 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.

Hiroshi Imamizu, Satoru Miyauchi, Tomoe Tamada, Yuka Sasaki, Ryousuke Takino, Benno PuÈtz, Toshinori Yoshioka, and Mitsuo Kawato. Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature*, 403(6766):192–195, 2000.

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

Yael Niv, Michael O Duff, and Peter Dayan. Dopamine, uncertainty and td learning. *Behavioral and brain Functions*, 1(1):6, 2005.

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

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

Adam S Lowet, Qiao Zheng, Sara Matias, Jan Drugowitsch, and Naoshige Uchida. Distributional reinforcement learning in the brain. *Trends in Neurosciences*, 2020.

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

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

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

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

Carlos Diuk, Anna Schapiro, Natalia Córdova, José Ribas-Fernandes, Yael Niv, and Matthew 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.

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

Alec Solway, Carlos Diuk, Natalia Córdova, Debbie Yee, Andrew G Barto, Yael Niv, and Matthew M Botvinick. Optimal behavioral hierarchy. *PLOS Comput Biol*, 10(8): e1003779, 2014.

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

of cognitive neuroscience, 31(1):8–23, 2019.

Yael Niv, Reka Daniel, Andra Geana, Samuel J Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21):8145–8157, 2015.

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

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

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

Louis Kirsch, Sjoerd van Steenkiste, and Juergen Schmidhuber. Improving generalization in meta reinforcement learning using learned objectives. In *International Conference on Learning Representations*, 2019.

Junhyuk Oh, Matteo Hessel, Wojciech M Czarnecki, Zhongwen Xu, Hado P van Hasselt, Satinder Singh, and David Silver. Discovering reinforcement learning algorithms. *Advances in Neural Information Processing Systems*, 33, 2020.

Rui Wang, Joel Lehman, Aditya Rawal, Jiale Zhi, Yulun Li, Jeff Clune, and Kenneth O Stanley. Enhanced poet: Open-ended reinforcement learning through unbounded invention of learning challenges and their solutions. *arXiv preprint arXiv:2003.08536*, 2020b.

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

Anna Harutyunyan. What is an agent? *arxiv*, 2020.

Khimya Khetarpal, Zafarali Ahmed, Gheorghe Comanici, David Abel, and Doina Precup. What can i do here? a theory of affordances in reinforcement learning. *International Conference on Machine Learning*, 2020b.

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

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

Harry 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.

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

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019.

Michael L. Littman. Friend-or-foe q-learning in general-sum games. In *ICML*, page 322–328, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Xiaofeng Wang and Tuomas Sandholm. Reinforcement learning to play an optimal nash equilibrium in team markov games. In *NeurIPS*, page 1603–1610. MIT Press, 2002.

Michael Bowling. Convergence and no-regret in multiagent learning. *Advances in neural information processing systems*, 17, 2004.

Amy Greenwald and Keith Hall. Correlated-Q learning. In *ICML*, page 242–249. AAAI Press, 2003. ISBN 1577351894.

Martin Zinkevich, Amy Greenwald, and Michael Littman. Cyclic equilibria in markov games. In *NeurIPS*, volume 18. MIT Press, 2006. URL `https://proceedings.neurips.cc/paper/2005/file/9752d873fa71c19dc602bf2a0696f9b5-Paper.pdf`.

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. In *ICLR*, 2019. URL `https://openreview.net/forum?id=SyGjjsC5tQ`.

Lisa Torrey and Matthew Taylor. Teaching on a budget: Agents advising agents in reinforcement learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1053–1060, 2013.

Shayegan Omidshafiei, Dong Ki Kim, Miao Liu, Gerald Tesauro, Matthew Riemer, Christopher Amato, Murray Campbell, and Jonathan P How. Learning to teach in cooperative multiagent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6128–6136, 2019.

Dong Ki Kim, Miao Liu, Shayegan Omidshafiei, Sebastian Lopez-Cot, Matthew Riemer, Golnaz Habibi, Gerald Tesauro, Sami Mourad, Murray Campbell, and Jonathan P How. Learning hierarchical teaching policies for cooperative agents. In *AAMAS*, 2020.

Chongjie Zhang and Victor R. Lesser. Multi-agent learning with policy prediction. In *AAAI*, 2010.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022.