Ergodic Risk Measures: Towards a Risk-Aware Foundation for Continual Reinforcement Learning

Juan Sebastian Rojas
University of Toronto
juan.rojas@mail.utoronto.ca

Chi-Guhn Lee
University of Toronto
cglee@mie.utoronto.ca

Abstract

Continual reinforcement learning (continual RL) seeks to formalize the notions of lifelong learning and endless adaptation in RL. In particular, the aim of continual RL is to develop RL agents that can maintain a careful balance between retaining useful information and adapting to new situations. To date, continual RL has been explored almost exclusively through the lens of risk-neutral decision-making, in which the agent aims to optimize the expected long-run performance. In this work, we present the first formal theoretical treatment of continual RL through the lens of *risk-aware* decision-making, in which the behaviour of the agent is directed towards optimizing a measure of long-run performance beyond the mean. In particular, we show that the classical theory of risk measures, widely used as a theoretical foundation in non-continual risk-aware RL, is, in its current form, incompatible with continual learning. Then, building on this insight, we extend risk measure theory into the continual setting by introducing a new class of *ergodic risk measures*, and showing that it is compatible with continual learning. Finally, we provide a case study of continual risk-aware learning, along with empirical results, which show the intuitive appeal of ergodic risk measures in continual settings.

1 Introduction

Reinforcement learning (RL) (Sutton and Barto, 2018) has enjoyed success over the years when tackling certain problems of interest in various domains, ranging from video games to robotics. However, the RL agents behind these successes are typically trained in static environments, and are evaluated on a single task for a finite period of time. By contrast, RL agents deployed in the real world may be required to operate indefinitely in scenarios where the environment and/or task changes over time. This discrepancy has motivated the study of continual reinforcement learning (continual RL) (Ring, 1994; Khetarpal et al., 2022; Abel et al., 2023; Kumar et al., 2025), which formalizes the challenges of lifelong learning and endless adaptation in RL. At the heart of continual RL is the *stability-plasticity dilemma*, through which an agent must learn to preserve sufficient prior knowledge, while still remaining sufficiently flexible to adapt to new streams of experience.

To date, the notion of continual RL has been studied almost exclusively under a *risk-neutral* lens, such that the behaviour of the agent is directed towards optimizing some expectation-based measure of long-run performance (e.g. the expected sum of discounted rewards or the average per-step reward). In this work, we present the first formal theoretical exploration of continual RL through the lens of *risk-aware* decision-making, in which the behaviour of the agent is directed towards optimizing a measure of long-run performance beyond the mean. Through this exploration, we seek to provide a theoretically grounded formalism through which we can frame the notion of risk-awareness in a continual setting. Namely, the first half of our exploration is guided by the following questions:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

▷ *What does it mean for an agent to be risk-aware in a continual setting?*

▷ *What conditions are needed to enable an agent to be risk-aware in a continual setting?*

▷ *What are the implications of risk-awareness as it relates to the stability-plasticity dilemma?*

To answer these questions, we first propose a risk-aware generalization of the continual RL framework proposed in Abel et al. (2025), where we formalize continual risk-aware RL as an ongoing exchange of actions and observations between an agent and an environment, such that the agent receives observations from the environment, processes those observations into *risk-aware observations* based on its *risk attitude* (or *risk tolerance*), then emits actions back to the environment based on the risk-aware observations. Then, we leverage this risk-aware framework to propose a set of axioms that establish the conditions needed to enable an agent to be risk-aware in continual settings.

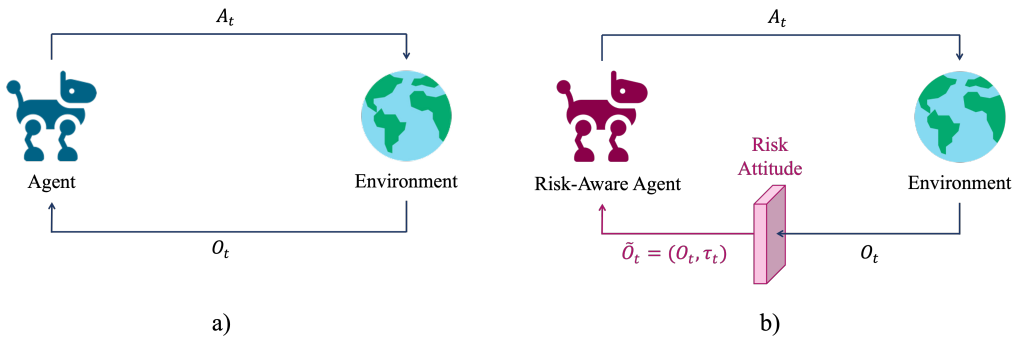

Figure 1: A comparison of: **a)** the continual RL framework proposed in Abel et al. (2025), and **b)** the risk-aware generalization of the framework proposed in this work.

The second half of our exploration is then aimed at examining existing frameworks and methodologies used for non-continual risk-aware RL to see how compatible they are with these axioms. To this end, we examine the classical theory of *risk measures* (e.g. see Chapter 6 of Shapiro et al. (2009)), which has served as a theoretical foundation for risk-aware decision-making in non-continual RL, and find that, in its current form, it is incompatible with the axioms, and hence, continual learning. Then, building on this insight, we extend risk measure theory into the continual setting by proposing a new class of *ergodic risk measures*, and showing that it is compatible with continual learning.

Finally, using the well-known average-reward Markov decision process (MDP) formulation (Puterman, 1994) as a basis, we provide a case study, along with numerical results, which show the intuitive appeal of ergodic risk measures in continual settings. Altogether, these contributions provide, to the best of our knowledge, the first formal theoretical treatment of risk-aware decision-making in a continual (i.e., lifelong) learning setting.

2 Related Work

2.1 Continual Reinforcement Learning

The notions of lifelong learning and endless adaptation in the context of RL have long been studied under different names and perspectives. In recent years, several works (e.g. Khetarpal et al. (2022); Abel et al. (2023); Kumar et al. (2025)) have attempted to unify and frame these diverse sets of works as instances of continual RL. Some of the more common types of RL-related works that can be interpreted as being instances of continual RL include the study of the loss of plasticity in deep RL agents (e.g. Abbas et al. (2023); Dohare et al. (2024)), transfer learning (e.g. Abel et al. (2018); Gimelfarb et al. (2021)), and decision-making in non-stationary environments (e.g. Dick et al. (2014); Luketina et al. (2022)). In this work, we place a great emphasis on the notion of *plasticity*, which has been studied extensively in prior works such as Raghavan and Balaprakash (2021), Chen et al. (2023), and Abel et al. (2025). In our proposed framework, we build on the definition of plasticity (and subsequent continual RL framework) proposed in Abel et al. (2025).

2.2 Risk-Aware Reinforcement Learning

The notion of risk-aware learning and decision-making in the context of RL has been studied under various theoretical frameworks, from the well-established expected utility framework (Howard and Matheson, 1972), to the more contemporary framework of risk measures (e.g. Chapter 6 of Shapiro et al. (2009)). In this work, we focus on the latter framework, which originated in the finance literature (e.g. Rockafellar and Uryasev (2000)), but has since been widely integrated into RL-based works (e.g. Bäuerle and Ott (2011)). Traditionally, non-continual risk-aware RL works have aimed to optimize either a *static* risk measure (e.g. Mead et al. (2025)), or a *nested (dynamic)* risk measure (e.g. Wang and Delage (2025)). The case study presented in this work primarily focuses on the conditional value-at-risk (CVaR) risk measure (Rockafellar and Uryasev, 2000), which has been studied extensively in the discounted setting (e.g. Bäuerle and Ott (2011); Muni et al. (2026)), and, to a lesser extent, in the average-reward setting (e.g. Xia et al. (2023); Rojas and Lee (2025)).

3 Preliminaries

3.1 Continual Reinforcement Learning

Continual reinforcement learning (continual RL) can be viewed as an agent-environment interaction occurring over an infinite time horizon. In particular, following the framework proposed in Abel et al. (2025), this interaction can be modeled as an ongoing exchange of signals between an agent and an environment, where, at each discrete time step, t, the agent emits an action from an action-space, $A_t \in \mathcal{A}$, and the environment emits an observation from an observation-space, $O_t \in \mathcal{O}$ (see Figure 1a)). Accordingly, for a given agent and environment, we can define their interaction as a pair of sequences of random variables, $O_{a:b} \doteq (O_a, O_{a+1}, \dots, O_{b-1}, O_b)$ and $A_{c:d} \doteq (A_c, A_{c+1}, \dots, A_{d-1}, A_d)$, where a, b, c, and d denote discrete time indices.

We next discuss the notion of *plasticity*. In particular, one of the defining aspects of continual RL is the stability-plasticity dilemma, through which an agent must carefully balance the degree to which newly acquired information affects its behaviour relative to previously learned knowledge. More formally, the plasticity, $\mathfrak{P}$, of a given agent, λ, relative to a set of environments, $\mathcal{E}$, with respect to the time intervals $[a:b] \doteq [a, a+1, \dots, b-1, b]$, $1 \leq a \leq b$, and $[c:d] \doteq [c, c+1, \dots, d-1, d]$, $1 \leq c \leq d$, can be defined as follows (Abel et al., 2025):

$$\underset{c:d}{\mathfrak{P}_{a:b}}(\lambda, \mathcal{E}) \doteq \max_{e \in \mathcal{E}} \mathbb{I}(O_{a:b} \to A_{c:d}), \tag{1}$$

where $\mathbb{I}(X_{a:b} \to Y_{c:d}) \doteq \sum_{i=\max(a,c)}^{d} \mathbb{I}(X_{a:\min(b,i)}; Y_i \mid X_{1:a-1}, Y_{1:i-1})$ denotes the *generalized directed information* (Abel et al., 2025) between two sequences of random variables, $X_{a:b}$ and $Y_{c:d}$, and $\mathbb{I}(X;Y)$ denotes the mutual information between random variables X and Y. Here, we utilize the same formal definitions for the agent and environment as Abel et al. (2025). These definitions are also provided in Appendix E. In words, Equation (1) interprets plasticity as a measure of how much a sequence of observations from a set of environments can influence a sequence of the agent's actions. Accordingly, under this framework, the stability-plasticity dilemma can be interpreted as the dilemma associated with determining the right amount of plasticity, $\mathfrak{P}$.

Importantly, the definition for plasticity in Equation (1) applies to both continual and non-continual learning settings. Accordingly, given our focus on continual learning, we will restrict our attention to a set of environments that give rise to meaningful continual learning scenarios. In particular, as discussed in Section 5.2 of Abel et al. (2025), in the context of continual learning, one can think of the set of environments, $\mathcal{E}$, as being comprised of a base environment, e, along with a set of *interventions*, $\mathcal{K}$, such that each intervention, $k \in \mathcal{K}$, creates a new environment, $e' = k(e)$, that produces a (potentially) different sequence of observations than the base environment. Accordingly, assuming a meaningful set of interventions (that yield meaningful continual learning scenarios; see Section 5.2 of Abel et al. (2025)), we can further interpret the agent's plasticity, in the context of continual learning, as its *capacity* to adapt to changes (i.e., interventions) in the environment. For the remainder of this work, we will implicitly assume that our agent is operating with respect to such a set of environments. This is formalized as Assumption 3.1 below:

Assumption 3.1. *The agent operates with respect to a set of environments, $\mathcal{E}$, consisting of a base environment, e, and a set of intervention-induced environments, such that $\mathcal{E} \doteq \{e\} \cup \{k(e) \mid k \in \mathcal{K}\}$, where $\mathcal{K}$ denotes a set of interventions that induce meaningful continual learning scenarios.*

3.2 Risk Measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space, and let $\mathcal{X}$ denote a space of random variables of the form $X : \Omega \to \mathbb{R}$. A *risk measure* (e.g. see Chapter 6 of Shapiro et al. (2009)) is a functional, $\rho : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$, that assigns to each random variable, $X \in \mathcal{X}$, a real value representing the degree of risk associated with X with respect to some *risk attitude* or *risk tolerance*, $\tau \in \mathcal{T}$, where $\mathcal{T}$ denotes a set that captures the agent's risk attitude. In other words, a risk measure is a mapping that quantifies the risk associated with a random variable based on a given risk attitude.

The precise interpretation of what "risk" means depends on the specific risk measure used; different risk measures capture different risk-related aspects of a random variable. In essence, one can think of a risk measure as any functional that captures distributional characteristics of a random variable, typically beyond just its mean. However, an emphasis is usually placed on deriving risk measures that satisfy certain mathematical properties that can be meaningful in risk-based decision-making contexts. In Appendix A, we provide formal definitions of the various classes of risk measures used in the context of RL, where each (non-mutually exclusive) class of risk measures can be thought of as satisfying a specific set of mathematical properties. We note that it is commonplace to write $\rho(X)$ or $\rho_\tau(X)$ as shorthand for $\rho(X, \tau)$.

3.3 Non-Continual Risk-Aware Reinforcement Learning

Traditionally, non-continual risk-aware RL works have aimed to optimize either a *static* risk measure (e.g. Mead et al. (2025)), or a *nested* risk measure (e.g. Wang and Delage (2025)). We now provide informal definitions of these two classes of risk measures, as well as an informal definition pertaining to the notion of *time consistency* as it relates to risk measures (see Appendix A for formal definitions):

Static Risk Measures (Definition A.2): A *static* risk measure, ρ_N, evaluates risk at a fixed point in time, N. In RL-based contexts, static risk measures can be useful for quantifying the risk of the return at the end of an episode. The primary appeal of static risk measures is their interpretability.

Nested Risk Measures (Definition A.6): A *nested* risk measure is a type of *dynamic* risk measure (i.e., a sequence of one-step risk measures, $\{\rho_t\}_{t=0}^N$; see Definition A.4), that is constructed recursively from the one-step risk measures, such that the risk at time n, ρ_n, can be computed as $\rho_n(X) \doteq \rho_n\big(\rho_{n+1}(\cdots \rho_{N-1}(X)\cdots)\big)$. Nested risk measures are useful because they ensure *time consistency* (described below). In RL-based contexts, nested risk measures are also useful because they can induce Bellman-like recursions with appealing dynamic programming-like properties. However, nested risk measures are typically hard to interpret. We note that typically in the RL literature, the terms "dynamic risk measure" and "nested risk measure" are used interchangeably; however, formally speaking, dynamic risk measures need not be time-consistent, nor have the nested structure.

Time-Consistent Risk Measures (Definition A.5): A dynamic risk measure, $\{\rho_t(X)\}_{t=0}^N$, is said to be *time-consistent* if, for all $X, X' \in \mathcal{X}$ and all $t < N$, we have that $\rho_{t+1}(X) \leq \rho_{t+1}(X') \implies \rho_t(X) \leq \rho_t(X')$. Time-consistent risk measures can be appealing because they ensure that if one future outcome is deemed less risky than another at some time step, t, then that same outcome is not deemed more risky than the other at any other time step. One way to think about time consistency in RL-based contexts is to ask the question: *can the agent change its mind about how relatively risky something is based on new information?* If the answer is *yes*, then there exists a lack of time consistency.

4 Continual Risk-Aware Reinforcement Learning

What does it mean for a lifelong agent to be risk-aware? In this section, we seek to answer this question by formalizing the notion of risk-awareness in the continual setting. In particular, we motivate and propose a risk-aware generalization of the continual RL framework proposed in Abel et al. (2025), such that we formalize continual risk-aware RL as an ongoing exchange of *risk-aware observations* and actions between an agent and an environment (see Figure 1). To this end, we begin by formalizing the notion of a risk-aware observation:

Definition 4.1 (Risk-Aware Observation)**.** *A* risk-aware observation *at time t, $\tilde{O}_t \in \tilde{\mathcal{O}}$, is the tuple (O_t, τ_t), where $O_t \in \mathcal{O}$ denotes an observation from the environment, $\tau_t \in \mathcal{T}$ denotes the agent's risk attitude, and $\tilde{\mathcal{O}}$ denotes the space of risk-aware observations.*

In other words, a risk-aware observation can be interpreted as an observation from the environment that has been "augmented" based on the agent's risk attitude. The primary appeal of such a formulation is that it admits a natural connection to risk measures (see Section 3.2). In particular, we can define an agent's *risk assessment* over a sequence of observations, $O_{a:b}$, as follows:

Definition 4.2 (Risk Assessment)**.** *An agent's* risk assessment *over a sequence of observations collected during the interval* $[a:b]$*,* $U_{a:b} \doteq f(\tilde{O}_{a:b}) \doteq f(O_a, O_{a+1}, \ldots, O_{b-1}, O_b, \tau_a, \tau_{a+1}, \ldots, \tau_{b-1}, \tau_b)$*, is the output of the mapping* $f : \tilde{\mathcal{O}}^* \to \mathcal{U}$*, where* $\tilde{\mathcal{O}}^*$ *denotes the space of risk-aware observation sequences and* $\mathcal{U}$ *denotes the space of risk assessments.*

In essence, a risk assessment allows the agent to capture the risk associated with a sequence of observations collected during a given time interval, such that the risk-aware observations act as a "bridge" between the observations emitted from the environment and the output of the risk assessment. Importantly, if the output of the risk assessment, $U_{a:b}$, is a scalar (i.e., $\mathcal{U} = \mathbb{R}$), then, by definition, the output of the risk assessment corresponds to a risk measure over $O_{a:b}$. For example, $U_{a:b}$ could correspond to a risk measure over a sum of observations collected during the time interval, or an average of risk measures over (one-step) observations collected during the time interval.

We can also use Definition 4.1 to generalize the definition of plasticity from Abel et al. (2025) (i.e., Equation (1)) for a risk-aware agent as follows:

$$\mathfrak{R}_{\substack{a:b \\ c:d}}(\lambda, \tilde{\mathcal{E}}) \doteq \max_{\tilde{e} \in \tilde{\mathcal{E}}} \mathbb{I}(\tilde{O}_{a:b} \to A_{c:d}), \tag{2}$$

where $\tilde{O}_{a:b}$ denotes a sequence of risk-aware observations over the interval $[a:b]$, and $\mathfrak{R}$ denotes the *risk-aware plasticity*. We use $\tilde{\mathcal{E}}$ to denote the set of pairs $\tilde{e} \doteq (e, \tau)$ of an environment, $e \in \mathcal{E}$, and a risk attitude, $\tau \in \mathcal{T}$, which together determine the sequence of risk-aware observations, $\tilde{O}_{a:b}$. As in Section 3.1, we implicitly assume that Assumption 3.1 applies to the set $\tilde{\mathcal{E}}$. In words, the risk-aware plasticity (2) captures the degree to which the agent's risk assessments and subsequent behaviour can be influenced by changes in its risk attitude as well as by new observations from the environment.

Altogether, Definitions 4.1, 4.2, and Equation (2) constitute a risk-aware generalization of the continual RL framework proposed in Abel et al. (2025), such that continual risk-aware RL can be interpreted as an ongoing exchange of risk-aware observations and actions between an agent and an environment that occurs over an infinite time horizon, in which the agent's behaviour is informed and guided by risk assessments performed over sequences of observations from the environment.

4.1 Axioms for Risk-Awareness in the Continual Setting

Having formalized the notion of risk-awareness in the context of continual RL, we now seek to establish the minimum requirements needed for an agent to be risk-aware in a continual learning setting. To this end, we argue that any limitations related to the agent's risk assessments should be direct consequences of the limitations of the agent itself. In particular, we focus on two well-argued limitations (or requirements) of continual RL agents (e.g. see Abel et al. (2023), Kumar et al. (2025), Abel et al. (2025)): 1) any continual RL agent should be able to operate under finite memory constraints, and 2) any continual RL agent should have a non-zero level of plasticity. As such, given these limitations of the agent, we propose the following axioms as the minimum requirements needed for risk-awareness in the continual setting:

Proposition 4.3. *Under Assumption 3.1, the following two axioms are necessary and sufficient conditions for risk-awareness in the continual setting:*

1. ***Feasibility Axiom****: The agent's risk assessments must be accurately computable based on the agent's finite memory. That is, the risk assessments must be computable based on observations collected over a finite time interval* $[a:b]$ *with respect to the agent's risk attitude during that same time interval.*
2. ***Plasticity Axiom****:* $\mathbb{P}(\mathfrak{R} > 0 \text{ infinitely often}) = 1$*. In words, this axiom states that the agent's risk assessments and subsequent behaviour should be influenced by changes in its risk attitude as well as by new observations from the environment infinitely often.*

4.2 Existing Risk Measures and Continual Learning

Having established the minimum requirements needed for risk-awareness in continual learning settings, we now turn our attention towards answering the following question: do the existing classes of risk measures used in non-continual risk-aware RL (i.e., static and nested; see Section 3.3) satisfy the requirements outlined in Proposition 4.3? In other words, if we assume that the output of an agent's risk assessment is a scalar (this is formalized as Assumption 4.4 below), are the existing classes of risk measures compatible with continual learning? As it turns out, static and nested risk measures *do not* satisfy the requirements of Proposition 4.3. This is formally shown in Propositions 4.5 and 4.6 below:

Assumption 4.4. *The output of an agent's risk assessment (see Definition 4.2) is a scalar.*

Proposition 4.5. *Let ρ_N denote a static risk measure (as per Definition A.2) defined for some time step, N. The static risk measure, ρ_N, does not satisfy both axioms of Proposition 4.3.*

Proof. As $N \to \infty$, the Feasibility axiom is violated since the computation of the risk would require the agent to have infinite memory. If $N < \infty$, then any observation that occurs after time step N would not influence the risk assessment, thereby violating the Plasticity axiom. □

Proposition 4.6. *Let $\{\rho_t\}_{t=0}^{\infty}$ denote a nested risk measure and let ρ_n denote the assessed risk at time step n, such that its value can be computed as: $\rho_n(X) \doteq \rho_n\big(\rho_{n+1}(\cdots)\big)$ (see Definition A.6). The nested risk measure, $\{\rho_t\}_{t=0}^{\infty}$, does not satisfy both axioms of Proposition 4.3.*

Proof. Consider the risk assessment at time n, $\rho_n(X) \doteq \rho_n\big(\rho_{n+1}(\cdots)\big)$. By definition, the validity of this risk calculation is contingent on time consistency holding for all $t \geq n$. However, this implies that future observations cannot influence how risky X is relative to other outcomes. Hence, either time consistency is satisfied, which violates the Plasticity axiom, or time consistency is violated, which invalidates the risk calculation at time n, thereby violating the Feasibility axiom. □

4.3 Ergodic Risk Measures for Continual Learning

In the previous section, we showed that static and nested risk measures are not compatible with continual learning. Based on this insight, we now propose a new class of *ergodic risk measures*, and show that it is compatible with Proposition 4.3, and hence, continual learning:

Definition 4.7 (Ergodic Risk Measure)**.** *Let $\{\rho_t\}_{t=0}^{\infty}$ denote a sequence of conditional risk measures (i.e., a dynamic risk measure) defined on a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, with filtration $(\mathcal{F}_t)_{t=0}^{\infty}$, where $\mathcal{F}_t \subseteq \mathcal{F}$ denotes the σ-algebra representing the information available up until time step $t \geq 0$, such that $\rho_t : L^{\infty}(\mathcal{F}_{\infty}) \to L^{\infty}(\mathcal{F}_t)$, where $L^{\infty}(\mathcal{F}_u)$ denotes the space of essentially bounded, $\mathcal{F}_u$-measurable random variables.*

Suppose that for any time step, $t > 0$, there exists a finite time interval, $[n : t]$, such that for all $X \in L^{\infty}(\mathcal{F}_{\infty})$, and some $\hat{\rho}_t(X) \in L^{\infty}(\mathcal{G}_{n:t})$, where $\mathcal{G}_{n:t}$ denotes the σ-algebra generated by the information from time step n up to time step t, we have that:

$$\rho_t(X) - \hat{\rho}_t(X) \approx 0. \tag{3}$$

That is, that the risk assessment at any given time step, t, can be accurately computed based on a preceding, finite subset of the history, such that the risk assessment effectively ceases to depend on the history that occurs prior to the interval $[n : t]$. Furthermore, suppose that $\{\rho_t\}_{t=0}^{\infty}$ is designed in such a way that time consistency, as defined in Definition A.5, is not strictly enforced. Then, we call $\{\rho_t\}_{t=0}^{\infty}$ an ergodic risk measure*.*

In essence, an ergodic risk measure is a non-nested dynamic risk measure, such that its one-step risk assessments, ρ_t, can be *accurately* computed using only observations collected over a rolling finite time interval, $[n : t]$. We note that the size of this interval can vary between different time steps.

We now show that ergodic risk measures are compatible with both axioms of Proposition 4.3:

Proposition 4.8. *Let $\{\rho_t\}_{t=0}^{\infty}$ denote an ergodic risk measure as defined in Definition 4.7. The ergodic risk measure, $\{\rho_t\}_{t=0}^{\infty}$, is compatible with both axioms of Proposition 4.3.*

Proof. Since an ergodic risk measure is a dynamic risk measure such that time consistency (as defined in Definition A.5) is not strictly enforced, it is compatible with the Plasticity axiom since the risk assessments and subsequent behaviour can be influenced by new observations and any changes in the agent's risk attitude infinitely often. Similarly, by definition, the risk at any time t can be computed using observations collected over a finite time interval with respect to the agent's risk attitude during that same time interval, thereby making it compatible with the Feasibility axiom. □

4.4 Ergodic Risk Measures as Reinforcement Learning Objectives

Having established that ergodic risk measures are compatible with continual learning, we now motivate a corresponding RL objective that can be used to perform risk-aware learning and decision-making in the continual setting. To this end, we will utilize the average-reward MDP formulation (Puterman, 1994) to realize the agent-environment interaction described in Section 3.1. Our choice of the average-reward formulation in this work follows prior work, such as Riemer et al. (2022), Sharma et al. (2022), and Kumar et al. (2025), which argue that the average-reward formulation's emphasis on long-term performance is a natural fit for continual learning settings (also see Appendix C).

More formally, consider a finite average-reward MDP, $\mathcal{M} \doteq \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p \rangle$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathcal{R} \subset \mathbb{R}$ is a bounded set of rewards, and $p : \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S} \to [0, 1]$ is a probabilistic transition function that describes the dynamics of the environment, such that at each discrete time step, $t = 0, 1, 2, \ldots$, an agent chooses an action, $A_t \in \mathcal{A}$, based on its current state, $S_t \in \mathcal{S}$, and receives a reward, $R_{t+1} \in \mathcal{R}$, while transitioning to a (potentially) new state, S_{t+1}, such that $p(s', r \mid s, a) = \mathbb{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$.

Under this framework, the continual RL problem can be viewed as an infinite sequence of average-reward MDPs, $\{\mathcal{M}^k\}_{k=1}^{\infty}$, such that $\mathcal{M}^k \doteq \langle \mathcal{S}_k, \mathcal{A}_k, \mathcal{R}_k, p_k \rangle$, where each $\mathcal{M}^k$ may differ in its state-space, action-space, reward function, and/or transition dynamics based on some indexing function, $\omega : \mathbb{N} \to \mathbb{N}$, such that $\omega(t) = i$ indicates that at time step t, the agent is in MDP $\mathcal{M}^i$.

The agent's goal in this framework is to construct a sequence of stationary policies, $\{\pi^k\}_{k=1}^{\infty}$, that optimizes the long-run (or limiting) average-reward, $\bar{r}$, which is defined as follows for a given stationary policy being followed at time t, $\pi_t \doteq \pi^{\omega(t)}$:

$$\bar{r}^{\pi_t}(s) \doteq \lim_{h \to \infty} \frac{1}{h - n_0} \sum_{n=n_0+1}^{h} \mathbb{E}[R_n \mid S_{n_0} = s, A_{n_0:n-1} \sim \pi^{\omega(n_0)}], \tag{4}$$

where $n_0 = \min\{y : \pi^{\omega(y)} = \pi_t\}$. When working with average-reward MDPs, it is commonplace to invoke *ergodicity-like* assumptions (see below) so that the agent's objective (4) remains well-defined and becomes independent of prior conditions (i.e., $\bar{r}^{\pi_t}(s) = \bar{r}^{\pi_t}$). In Appendix E, we provide a mathematically rigorous discussion on the relationship between ergodicity and continual learning, in which we show that ergodicity does not preclude an agent's ability to be a continual learner.

Assumption 4.9 (Unichain Assumption for Prediction)**.** *For a given MDP, $\mathcal{M}^k$, the Markov chain induced by the policy, π^k, is unichain. That is, the induced Markov chain consists of a single recurrent class and a potentially empty set of transient states.*

Assumption 4.10 (Communicating Assumption for Control)**.** *For a given MDP, $\mathcal{M}^k$, the MDP has a single communicating class. That is, each state in the MDP is accessible from every other state in the MDP under some deterministic stationary policy.*

For clarity, we note that the above assumptions need only apply independently for each MDP, $\mathcal{M}^k$, rather than holistically for the entire sequence of MDPs, $\{\mathcal{M}^k\}_{k=1}^{\infty}$. For example, Assumption 4.10 only requires that each state in a given MDP, $\mathcal{M}^k$, is accessible from every other state in $\mathcal{M}^k$, but not the states of other MDPs, $\mathcal{M}^{i \neq k}$.

Now, to perform risk-aware decision-making in a continual setting, we can consider the risk-aware analogue to the (risk-neutral) average-reward RL objective (4):

$$\bar{\rho}^{\pi_t}(s) \doteq \lim_{h \to \infty} \frac{1}{h - n_0} \sum_{n=n_0+1}^{h} \rho_\tau[R_n \mid S_{n_0} = s, A_{n_0:n-1} \sim \pi^{\omega(n_0)}]. \tag{5}$$

That is, we want to optimize some time-averaged measure of risk pertaining to the *limiting per-step reward distribution* (see Rojas and Lee (2026)) induced when following a given policy, π_t. However, the risk-aware objective presented in Equation (5) is dependent on the initial conditions, which is not tractable in the continual setting. As such, as with the risk-neutral objective (4), we can apply an appropriate ergodicity-like assumption, such as Assumption 4.9 or 4.10, such that the risk-aware objective (5) becomes independent of prior conditions (i.e., $\bar{\rho}^{\pi_t}(s) = \bar{\rho}^{\pi_t}$).

We will also assume that $\rho_\tau[R_n|\cdot]$ can be estimated at every time step (see Remark 4.12):

Assumption 4.11. *Given an appropriate ergodicity-like assumption, such as Assumption 4.9 or 4.10, the* risk functional *in Equation* (5)*, $\rho_\tau[R_n \mid \cdot]$, can be estimated at every time step.*

Remark 4.12. *We note that several common choices of ρ_τ satisfy Assumption 4.11, including $\rho_\tau[R_n|\cdot] = \mathbb{E}[R_n|\cdot]$, $\rho_\tau[R_n|\cdot] = Var[R_n|\cdot]$ (e.g. see Prashanth and Ghavamzadeh (2016)), and $\rho_\tau[R_n|\cdot] = CVaR_\tau[R_n|\cdot]$ (e.g. see Rojas and Lee (2025)).*

Importantly, we can show that, under such assumptions, the risk-aware objective (5) corresponds to an ergodic risk measure, thereby making it compatible with continual learning:

Theorem 4.13. *Given an appropriate ergodicity-like assumption, such as Assumption 4.9 or 4.10, a stationary policy, π_t, and Assumption 4.11, the risk-aware objective* (5) *corresponds to an ergodic risk measure, as defined in Definition 4.7.*

Proof. First, we note that, by construction, the risk-aware objective (5) does not strictly enforce time consistency. Next, we show that for any time step, t, we can accurately compute the risk based on observations collected over a preceding, finite time interval. To this end, consider an MDP, $\mathcal{M}^k$, through which an agent interacts with the environment for the possibly infinite time interval $[a : b]$. Under the stated assumptions, we can evaluate the risk-aware objective (5) at time $a \leq t \leq b$ as $\bar{\rho}_t^{\pi_t} = \frac{1}{(t-a+1)} \sum_{n=a}^{t} \rho_\tau[R_n | A_n \sim \pi^{\omega(a)}]$. If $[a : b]$ is finite, then the desired condition is automatically satisfied. If $[a : b]$ is infinite, then, under the ergodicity-like assumption, we have, by Birkhoff's Ergodic Theorem (Birkhoff, 1931), that $\bar{\rho}_t^{\pi_t}$ converges to a stationary value as $t \to \infty$. This implies that there exists a finite time step, j, such that for all $t \geq j$, $\bar{\rho}_t^{\pi_t}$ converges to a stationary value. We therefore have two cases to check: 1) If $t \leq j$, then, since j is finite, the risk can be computed based on observations collected over the finite time interval $[a : j]$, thereby satisfying the desired condition. 2) If $t > j$, then the convergence of $\bar{\rho}_t^{\pi_t}$ to a stationary value as $t \to \infty$ implies that for any $t > j$, we can find a finite time interval, $[n : t]$ (where $n \geq j$), that can be used to accurately compute the risk assessment, thereby satisfying the desired condition. This completes the proof. □

Remark 4.14. *We note that ergodic risk measures are also compatible with the* ***non-continual*** *average-reward setting. That is, ergodic risk measures provide a formalism for risk-aware objectives of the form* (5)*, which had been studied previously in non-continual settings (e.g. Xia et al. (2023), Rojas and Lee (2025)), but never formalized in the context of risk measure theory (see Appendix D).*

5 Case Study: CVaR as an Ergodic Risk Measure

In this section, we present a case study in which we show how optimizing an ergodic risk measure can induce sensible risk-aware behaviour in continual learning settings. That is, we show empirically how optimizing an ergodic risk measure can enable the agent to successfully adapt in situations where either its risk attitude changes, or the observations from the environment change.

To this end, we focus on optimizing the well-known conditional value-at-risk (CVaR) risk measure (Rockafellar and Uryasev, 2000). More formally, consider a random variable X with a cumulative distribution function, $F(x) = \mathbb{P}(X \leq x)$. The (left-tail) *value-at-risk (VaR)* of X with parameter $\tau \in (0, 1)$ represents the τ-quantile of X, such that $\text{VaR}_\tau(X) = \sup\{x \mid F(x) \leq \tau\}$. If $F(x)$ is continuous at $x = \text{VaR}_\tau(X)$, then $\text{CVaR}_\tau(X) = \mathbb{E}[X \mid X \leq \text{VaR}_\tau(X)]$. Importantly, as per the results in Section 4, and given Assumptions 4.9 or 4.10, the CVaR risk measure can be formulated as the following continual learning objective that corresponds to an ergodic risk measure:

$$\overline{\text{CVaR}}^{\pi_t} \doteq \lim_{h \to \infty} \frac{1}{h - n_0} \sum_{n=n_0+1}^{h} \text{CVaR}_\tau[R_n \mid A_{n_0:n-1} \sim \pi^{\omega(n_0)}]. \tag{6}$$

In this case study, we optimize the CVaR objective (6) via the *RED CVaR Q-learning* algorithm proposed in Rojas and Lee (2025) (see Remark 4.14) in two continual learning tasks. The full set of experiment details can be found in Appendix B.

In the first task, we consider a continual variation of the *red-pill blue-pill (RPBP)* task (Rojas and Lee, 2025). More specifically, in the regular (non-continual) RPBP task, an agent, at each time step, can take either a *red pill*, which takes them to the *red world* state, or a *blue pill*, which takes them to the *blue world* state. Each state has its own per-step reward distribution, such that for a sufficiently low CVaR parameter, τ, the red world state has a reward distribution with a lower (worse) mean but a higher (better) CVaR compared to the blue world state. That is, in the regular RPBP task, for a sufficiently low CVaR parameter, τ, we would expect a risk-neutral agent to learn a policy that prefers to stay in the blue world, and a risk-averse agent to learn a policy that prefers to stay in the red world.

In the continual variation of RPBP considered in this task, the risk attitude of the agent, which is governed by the CVaR parameter, τ, changes over time from risk-neutral ($\tau \approx 1$) to risk-averse ($\tau \approx 0$). Here, we would expect the agent to first learn to stay in the blue world state, but change its preference to the red world state as its risk attitude changes from risk-neutral to risk-averse. More formally, this task can be viewed as a continual learning task with a changing reward function. We refer to this task as the τ-RPBP task.

In the second task, we consider another continual variation of RPBP in which the per-step reward distributions of the states change over time, such that the agent is required to continually adapt and find the state with the better CVaR (given a fixed risk attitude, τ). More formally, this task can be viewed as a continual learning task with a changing state-space (such that a given state is effectively replaced with a state with a different reward distribution). We refer to this task as the $\mathcal{S}$-RPBP task.

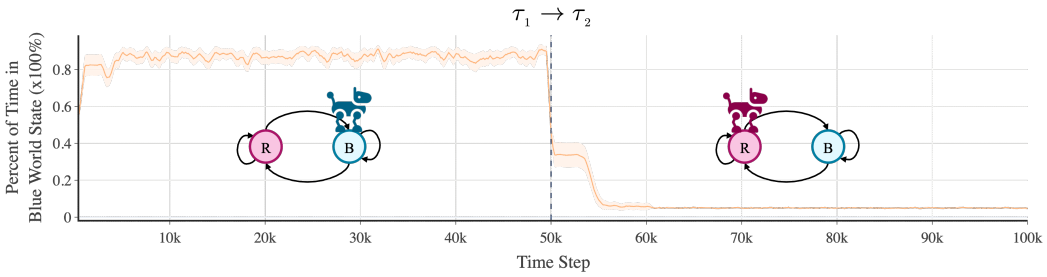

Figure 2: Rolling percent of time that the agent stays in the blue world state as learning progresses in the τ-RPBP task. A solid line denotes the mean percent of time spent in the blue world state, and the shaded region denotes a 95% confidence interval over 50 runs. As shown in the figure, the agent correctly learns to stay in the blue world state in the beginning, and then correctly changes its preference to the red world state once its risk attitude changes from risk-neutral to risk-averse.

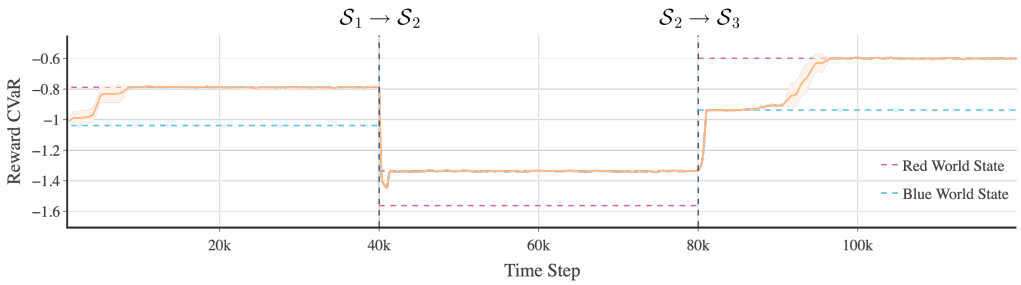

Figure 3: Rolling reward CVaR as learning progresses in the $\mathcal{S}$-RPBP task. A solid line denotes the mean CVaR, and the shaded region denotes a 95% confidence interval over 10 runs. The blue and red dashed lines denote the reward CVaR of the blue and red world states, respectively. As shown in the figure, the agent is able to continually adapt and find the state with the better CVaR.

In terms of empirical results, Figures 2 and 3 show the resulting agent behaviour as learning progresses in both tasks. In particular, Figure 2 shows that in the τ-RPBP task, the agent correctly learns to stay in the blue world state in the beginning, and then correctly changes its preference to the red world state once its risk attitude changes from risk-neutral to risk-averse. Similarly, Figure 3 shows that in the $\mathcal{S}$-RPBP task, the agent is able to continually adapt and find the state with the better CVaR.

6 Discussion

In this work, we took the first steps towards developing a risk-aware foundation for continual RL. In particular, we formalized the notion of risk-awareness in continual settings, and used the resulting framework to establish the conditions needed to enable an agent to be risk-aware in continual settings. We then examined the classical theory of risk measures, and showed that, in its current form, it is incompatible with continual learning. Then, building on this insight, we extended risk measure theory into the continual setting by introducing a new class of *ergodic risk measures*, and showing that it is compatible with continual learning. Finally, we provided a CVaR-based case study, along with numerical results, which showed the intuitive appeal of ergodic risk measures in continual settings.

More broadly, in *non-continual* settings, we note that ergodic risk measures offer a new and appealing alternative for risk-aware learning and decision-making. In particular, when utilizing the average-reward formulation, ergodic risk measures offer several advantages in comparison to the static and nested risk measures that are typically used in non-continual RL settings. A mathematically rigorous discussion of these advantages is provided in Appendix D. In summary, we show that ergodic risk measures can retain some notion of time consistency while remaining highly interpretable, thereby capturing the appeal of both static and nested risk measures. Moreover, we show that ergodic risk measures are able to circumvent some of the challenges and non-trivialities that have been historically associated with risk-aware RL objectives in non-continual settings.

We also note that this (theoretically motivated) move away from static and nested risk measures towards ergodic risk measures mirrors recent trends in the finance literature. In particular, the branch of *ergodicity economics* (Peters, 2019), which has gained popularity in recent years, argues that economic models should aim to optimize objectives (or *observables*) that (provably) satisfy ergodicity-like properties. As such, the use of ergodicity-like assumptions in our approach coincides with recent trends in other (non-RL) branches of the decision-making literature. Finally, we note that although some prior works in the RL literature have questioned the compatibility between ergodicity and continual learning (e.g. Elelimy et al. (2025)), we show in Appendix E, through a mathematically rigorous discussion on the relationship between ergodicity and continual learning, that ergodicity does not preclude an agent's ability to be a continual learner.

All in all, this work represents the first formal theoretical exploration of risk-aware decision-making in a continual learning setting. Moving forward, we believe that the theoretical foundation that has been established, including the introduction of a theoretically sound risk-aware objective that is stable-yet-adaptable, will enable further progress in the development of risk-aware lifelong agents.

Acknowledgments

We gratefully acknowledge funding from NSERC Discovery Grant # RGPIN-2026-04777, as well as computing resources provided by the Digital Research Alliance of Canada. We thank anonymous ICLR reviewer 6n6g for their valuable feedback and suggestions during an earlier review process. Finally, we thank the anonymous NeurIPS reviewers and area chair for their useful feedback and commentary during the review process.

References

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C Machado. Loss of plasticity in continual deep reinforcement learning. In *Proceedings of the 2nd Conference on Lifelong Learning Agents (CoLLAs 2023)*, March 2023.

David Abel, Yuu Jinnai, Yue (sophie) Guo, G Konidaris, and M Littman. Policy and value transfer in lifelong reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

David Abel, Andre Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

David Abel, Michael Bowling, André Barreto, Will Dabney, Shi Dong, Steven Hansen, Anna Harutyunyan, Khimya Khetarpal, Clare Lyle, Razvan Pascanu, Georgios Piliouras, Doina Precup, Jonathan Richens, Mark Rowland, Tom Schaul, and Satinder Singh. Plasticity as the mirror of empowerment. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Math. Finance*, 9(3):203–228, July 1999.

Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. The MIT Press, 2023.

George D Birkhoff. Proof of the ergodic theorem. *Proc. Natl. Acad. Sci. U. S. A.*, 17(12):656–660, December 1931.

Kang Boda and Jerzy A Filar. Time consistent dynamic risk measures. *Math. Methods Oper. Res. (Heidelb.)*, 63(1):169–186, February 2006.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Math. Methods Oper. Res.*, 74(3):361–379, December 2011.

Qi Chen, Changjian Shui, Ligong Han, and Mario Marchand. On the stability-plasticity dilemma in continual meta-learning: Theory and algorithm. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Travis Dick, Andras Gyorgy, and Csaba Szepesvari. Online learning in markov decision processes with changing cost sequences. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, August 2024.

Esraa Elelimy, David Szepesvari, Martha White, and Michael Bowling. Rethinking the foundations for continual reinforcement learning. *Reinf. Learn. J.*, 6:2174–2194, 2025.

Michael Gimelfarb, André Barreto, Scott Sanner, and Chi-Guhn Lee. Risk-aware transfer in reinforcement learning using successor features. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Manage. Sci.*, 18(7):356–369, March 1972.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *J. Artif. Intell. Res.*, 75:1401–1476, December 2022.

Saurabh Kumar, Henrik Marklund, Ashish Rao, Yifan Zhu, Hong Jun Jeon, Yueyang Liu, and Benjamin Van Roy. Continual learning as computationally constrained reinforcement learning. *Found. Trends® Mach. Learn.*, 18(5):913–1053, 2025.

Jelena Luketina, Sebastian Flennerhag, Yannick Schroecker, David Abel, Tom Zahavy, and Satinder Singh. Meta-gradients in non-stationary environments. In *Proceedings of The 1st Conference on Lifelong Learning Agents (CoLLAs 2022)*, 2022.

Harry Mead, Clarissa Costen, Bruno Lacerda, and Nick Hawes. Return capping: Sample-efficient CVaR policy gradient optimisation. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Aneri Muni, Vincent Taboga, Esther Derman, Pierre-Luc Bacon, and Erick Delage. Reward redistribution for CVaR MDPs using a bellman operator on L-infinity. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Ole Peters. The ergodicity problem in economics. *Nat. Phys.*, 15(12):1216–1221, December 2019.

L A Prashanth and Mohammad Ghavamzadeh. Variance-constrained actor-critic algorithms for discounted and average reward MDPs. *Mach. Learn.*, 105(3):367–417, December 2016.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

Krishnan Raghavan and Prasanna Balaprakash. Formalizing the generalization-forgetting trade-off in continual learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

Matthew Riemer, Sharath Chandra Raparthy, Ignacio Cases, Gopeshh Subbaraj, Maximilian Puelma Touzel, and Irina Rish. Continual learning in environments with polynomial mixing times. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

Mark Ring. *Continual Learning In Reinforcement Environments*. PhD thesis, University of Texas at Austin, 1994.

R Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000.

Juan Sebastian Rojas and Chi-Guhn Lee. Burning RED: Unlocking subtask-driven reinforcement learning and risk-awareness in average-reward markov decision processes. *Reinforcement Learning Journal*, 6:431–477, 2025.

Juan Sebastian Rojas and Chi-Guhn Lee. A differential perspective on distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Math. Program.*, 125(2):235–261, October 2010.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming : Modeling and theory*. MPS-SIAM, Philadelphia, PA, USA, 2009.

Archit Sharma, Kelvin Xu, Nikhil Sardana, Abhishek Gupta, Karol Hausman, Sergey Levine, and Chelsea Finn. Autonomous reinforcement learning: Formalism and benchmarking. In *International Conference on Learning Representations (ICLR)*, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction, 2nd edition*. MIT Press, November 2018.

Weikai Wang and Erick Delage. Planning and learning in average risk-aware MDPs. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Li Xia, Luyao Zhang, and Peter W Glynn. Risk-sensitive markov decision processes with long-run CVaR criterion. *Prod. Oper. Manag.*, 32(12):4049–4067, December 2023.

A Risk Measures

In this appendix, we provide formal definitions of the various classes of risk measures used in the context of RL, where each (non-mutually exclusive) class of risk measures can be thought of as satisfying a specific set of mathematical properties:

Definition A.1 (Coherent Risk Measure; adapted from Artzner et al. (1999))**.** *A risk measure, ρ, is called coherent if it satisfies the following four axioms for all random variables $X, X' \in \mathcal{X}$:*

1. *Monotonicity: If $X \leq X'$ almost surely, then $\rho(X) \leq \rho(X')$.*
2. *Translation Invariance: For all $c \in \mathbb{R}$, $\rho(X + c) = \rho(X) + c$.*
3. *Positive Homogeneity: For all $c \geq 0$, $\rho(cX) = c\rho(X)$.*
4. *Subadditivity: $\rho(X + X') \leq \rho(X) + \rho(X')$.*

Coherent risk measures are useful because they enforce a form of self-consistency in how risk is quantified and compared. In particular, monotonicity ensures that if a random variable, X, always yields outcomes that are no worse than outcomes induced by another random variable, X', then X should be considered less risky than X'. Translation invariance requires that adding a constant amount to X simply shifts its risk by that same amount. Positive homogeneity enforces scale-consistency, such that doubling the size of X also doubles its risk. Finally, subadditivity formalizes the idea of diversification, requiring that the risk of two combined random variables cannot exceed the sum of their individual risks. We note that the positive homogeneity and subadditivity properties also ensure that coherent risk measures are convex.

Definition A.2 (Static Risk Measure; adapted from Artzner et al. (1999))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space, and let $\mathcal{F}_N \subseteq \mathcal{F}$ denote the σ-algebra representing the information available at time N. Denote by $L^\infty(\mathcal{F}_N)$ the space of essentially bounded, $\mathcal{F}_N$-measurable random variables. A static risk measure is a mapping, $\rho : L^\infty(\mathcal{F}_N) \to \mathbb{R}$, that assigns to each random variable, $X \in L^\infty(\mathcal{F}_N)$, a single real value.*

In essence, a static risk measure evaluates risk at a fixed point in time. In RL-based contexts, static risk measures can be useful for quantifying the risk associated with the return at the end of an episode. One of the primary appeals of static risk measures is their interpretability.

Definition A.3 (Conditional Risk Measure; adapted from Ruszczyński (2010))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space with filtration $(\mathcal{F}_t)_{t=0}^N$, where $\mathcal{F}_t \subseteq \mathcal{F}$ represents the information available up to time t. Denote by $L^\infty(\mathcal{F}_N)$ the space of essentially bounded, $\mathcal{F}_N$-measurable random variables, and by $L^\infty(\mathcal{F}_t)$ the space of essentially bounded, $\mathcal{F}_t$-measurable random variables. A conditional risk measure at time t is a mapping, $\rho_t : L^\infty(\mathcal{F}_N) \to L^\infty(\mathcal{F}_t)$, that assigns to each random variable, $X \in L^\infty(\mathcal{F}_N)$, a conditional risk evaluation, $\rho_t(X)$, that is $\mathcal{F}_t$-measurable and satisfies the following monotonicity property: if $X \leq X'$ almost surely, then $\rho_t(X) \leq \rho_t(X')$.*

In essence, a conditional risk measure at time t is a mapping that evaluates the risk of future outcomes (e.g. at time $N > t$) based on the information available up to and including time t. The monotonicity property ensures that if a future outcome, X, always yields less loss (or more reward) than another outcome, X', then X is never assigned a higher risk than X'.

Definition A.4 (Dynamic Risk Measure; adapted from Ruszczyński (2010))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space with filtration $(\mathcal{F}_t)_{t=0}^N$. A dynamic risk measure is a* sequence *of conditional risk measures, $\{\rho_t\}_{t=0}^N$, where each $\rho_t : L^\infty(\mathcal{F}_N) \to L^\infty(\mathcal{F}_t)$ assigns to every random variable $X \in L^\infty(\mathcal{F}_N)$ a conditional risk evaluation, $\rho_t(X)$, that is $\mathcal{F}_t$-measurable.*

In essence, a dynamic risk measure is a time-indexed sequence of conditional risk measures. In RL-based contexts, dynamic risk measures can be useful for capturing the sequential nature of decision-making (i.e., that actions taken at each time step can potentially influence future outcomes, and hence, future risk evaluations).

Definition A.5 (Time-Consistent Risk Measure; adapted from Boda and Filar (2006))**.** *Let $\{\rho_t\}_{t=0}^N$ be a dynamic risk measure defined on a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, with filtration $(\mathcal{F}_t)_{t=0}^N$. The dynamic risk measure is said to be time-consistent if, for all random variables $X, X' \in L^\infty(\mathcal{F}_N)$ and all $t < N$,*

$$\rho_{t+1}(X) \leq \rho_{t+1}(X') \implies \rho_t(X) \leq \rho_t(X'). \tag{A.1}$$

Time-consistent risk measures can be appealing because they ensure that if one future outcome is deemed less risky than another at some time step, t, then that same outcome is not deemed more risky than the other at any other time step. One way to think about time consistency in RL-based contexts is to ask the question: *can the agent change its mind about how relatively risky something is based on new information?* If the answer is *yes*, then there exists a lack of time consistency.

Definition A.6 (Nested Risk Measure; adapted from Ruszczyński (2010))**.** *A nested risk measure is a dynamic risk measure, $\{\rho_t\}_{t=0}^N$, that is constructed recursively from one-step conditional risk measures. More formally, given one-step conditional risk measures, $\rho_t : L^\infty(\mathcal{F}_{t+1}) \to L^\infty(\mathcal{F}_t)$, the risk at time n, ρ_n, can be evaluated as:*

$$\rho_n(X) \doteq \rho_n\big(\rho_{n+1}(\cdots \rho_{N-1}(X)\cdots)\big). \tag{A.2}$$

Nested risk measures are useful because they ensure time consistency, such that the risk evaluation at earlier times is consistent with future evaluations. In RL-based contexts, nested risk measures are also useful because they can induce Bellman-like recursions with appealing dynamic programming-like properties. One of the drawbacks of nested risk measures is that they are typically hard to interpret. We note that typically in the RL literature, the terms "dynamic risk measure" and "nested risk measure" are used interchangeably; however, formally speaking, dynamic risk measures need not be time-consistent, nor have the nested structure.

Definition A.7 (Markov Risk Measure; adapted from Ruszczyński (2010))**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space, and let $\{S_t\}_{t=0}^N$ denote a Markov process where each state, $S_t : \Omega \to \mathcal{S}$, takes values in a measurable state-space, $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, such that $\mathcal{F}_t \doteq \sigma(S_0, \ldots, S_t)$. Here, $\mathcal{B}(\mathcal{S})$ denotes the Borel σ-algebra on $\mathcal{S}$. A one-step conditional risk measure, $\rho_t : L^\infty(\mathcal{F}_{t+1}) \to L^\infty(\mathcal{F}_t)$, is called a Markov risk measure if, for every $X \in L^\infty(\mathcal{F}_{t+1})$, the risk assessment satisfies $\rho_t(X) \in L^\infty(\sigma(S_t))$, where $\sigma(S_t) \subseteq \sigma(S_0, \ldots, S_t) = \mathcal{F}_t$. That is, the risk assessment $\rho_t(X)$ is $\sigma(S_t)$-measurable, such that it only depends on the current state, S_t.*

Markov risk measures are useful because they enforce a one-step time dependence structure that makes them compatible with MDP-based RL solution methods. From a risk perspective, this means that the assessment of risk at each time step only depends on the information available at that time step, rather than the entire history of past information. Note that in an MDP setting (as opposed to the simpler Markov process described in Definition A.7), the "state" can be characterized as a state-action pair. That is, $\rho_t(X) \in L^\infty(\sigma(S_t, A_t))$ for some A_t in a measurable action-space, $\mathcal{A}$.

B Numerical Experiments

This appendix contains details regarding the numerical experiments performed as part of this work. The overall aim of the experiments was to show how optimizing an ergodic risk measure can induce sensible risk-aware behaviour in continual learning settings. That is, we aimed to show empirically how optimizing an ergodic risk measure can enable the agent to successfully adapt in situations where either its risk attitude changes, or the observations from the environment change.

To this end, we focused on optimizing the well-known conditional value-at-risk (CVaR) risk measure (Rockafellar and Uryasev, 2000). More formally, consider a random variable, X, with a finite mean on a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, and with a cumulative distribution function, $F(x) = \mathbb{P}(X \leq x)$. The (left-tail) *value-at-risk (VaR)* of X with parameter $\tau \in (0, 1)$ represents the τ-quantile of X, such that $\text{VaR}_\tau(X) = \sup\{x \mid F(x) \leq \tau\}$. When $F(x)$ is continuous at $x = \text{VaR}_\tau(X)$, $\text{CVaR}_\tau(X)$ can be interpreted as the expected value of X conditioned on X being less than or equal to $\text{VaR}_\tau(X)$, such that $\text{CVaR}_\tau(X) = \mathbb{E}[X \mid X \leq \text{VaR}_\tau(X)]$.

As per Section 5, the CVaR risk measure can be formulated as the continual learning objective (6), which is displayed below as Equation (B.1) for convenience:

$$\overline{\text{CVaR}}^{\pi_t} \doteq \lim_{h\to\infty} \frac{1}{h - n_0} \sum_{n=n_0+1}^{h} \text{CVaR}_\tau[R_n \mid A_{n_0:n-1} \sim \pi^{\omega(n_0)}]. \tag{B.1}$$

In other words, we aimed to optimize the (left-tail) conditional value-at-risk associated with the limiting per-step reward distribution induced when following stationary policy π_t. More specifically, our aim was to optimize the CVaR objective (B.1) in two continual learning tasks via the ***RED CVaR Q-learning*** algorithm proposed in Rojas and Lee (2025). Importantly, the RED CVaR Q-learning algorithm was designed to optimize the CVaR associated with the long-run per-step reward distribution of an average-reward MDP, which precisely corresponds to the continual learning objective (B.1). The RED CVaR Q-learning algorithm (Algorithm 1) is shown below:

Algorithm 1 RED CVaR Q-Learning (Tabular) (Rojas and Lee, 2025)

Input: the policy π to be used (e.g., ε-greedy)
Algorithm parameters: step size parameters α, α_{CVaR}, α_{VaR}; CVaR parameter τ
Initialize $Q(s, a)\ \forall s, a$ (e.g. to zero)
Initialize CVaR arbitrarily (e.g. to zero)
Initialize VaR arbitrarily (e.g. to zero)
Obtain initial S
while still time to train **do**
  $A \leftarrow$ action given by π for S
  Take action A, observe R, S'
  $\tilde{R} = \text{VaR} - \frac{1}{\tau}\max\{\text{VaR} - R, 0\}$
  $\delta = \tilde{R} - \text{CVaR} + \max_a Q(S', a) - Q(S, A)$
  $Q(S, A) = Q(S, A) + \alpha\delta$
  $\text{CVaR} = \text{CVaR} + \alpha_{\text{CVaR}}\delta$
  if $R \geq \text{VaR}$ **then**
    $\text{VaR} = \text{VaR} + \alpha_{\text{VaR}}(\delta + \text{CVaR} - \text{VaR})$
  else
    $\text{VaR} = \text{VaR} + \alpha_{\text{VaR}}\left(\left(\frac{\tau}{\tau - 1}\right)\delta + \text{CVaR} - \text{VaR}\right)$
  end if
  $S = S'$
end while
return Q

In terms of the two continual learning tasks considered in this work, we considered two continual variations of the *red-pill blue-pill (RPBP)* task (Rojas and Lee, 2025). More specifically, in the regular (non-continual) RPBP task, an agent, at each time step, can take either a *red pill*, which takes them to the *red world* state, or a *blue pill*, which takes them to the *blue world* state. Each state has its own characteristic per-step reward distribution, such that for a sufficiently low CVaR parameter, τ,

the red world state has a reward distribution with a lower (worse) mean but a higher (better) CVaR compared to the blue world state. That is, in the regular RPBP task, for a sufficiently low CVaR parameter, τ, we would expect a risk-neutral agent to learn a policy that prefers to stay in the blue world, and a risk-averse agent to learn a policy that prefers to stay in the red world.

We now discuss the two continual variations of the RPBP task considered in this work:

B.1 τ-RPBP Task

In the first task, we considered a continual variation of RPBP, such that the *risk attitude* of the agent, which is governed by the CVaR parameter, τ, changes over time from risk-neutral ($\tau = 0.9$) to risk-averse ($\tau = 0.1$). Here, we would expect the agent to first learn to stay in the blue world state, but change its preference to the red world state as its risk attitude changes from risk-neutral to risk-averse. More formally, this task can be viewed as a continual learning task, $\{\mathcal{M}^k\}_{k=1}^2$, with a changing reward function, such that $\mathcal{M}^k \doteq \langle \mathcal{S}, \mathcal{A}, \mathcal{R}_k, p \rangle$, where

$$\tilde{R}_{t,k} = \text{VaR}_t - \frac{1}{\tau_k}(\text{VaR}_t - R_t)^+ \text{ (see Algorithm 1)}, \tag{B.2}$$

with $\tau_1 = 0.9$ and $\tau_2 = 0.1$. The indexing function, ω, was defined such that $\omega(t) = 1$ for $t < 50{,}000$, and $\omega(t) = 2$ otherwise. That is, the agent's risk attitude changes from risk-neutral to risk-averse at $t = 50{,}000$.

In terms of the hyperparameters used with the RED CVaR Q-learning algorithm, we used the tuned hyperparameters from Rojas and Lee (2025). That is, $\alpha =$ 2e-2, $\alpha_{\text{CVaR}} \doteq \eta_{\text{CVaR}}\alpha$, where $\eta_{\text{CVaR}} =$ 1e-1, and $\alpha_{\text{VaR}} \doteq \eta_{\text{VaR}}\alpha$, where $\eta_{\text{VaR}} =$ 1e-1. We used an ε-greedy policy with a fixed epsilon of 0.1, and set all initial guesses to zero. The reward distribution parameters used are the same as the parameters used in the $k = 1$ portion of the $\mathcal{S}$-RPBP task discussed in Section B.2 below. The results for this *τ-RPBP* task are shown in Figure 2.

B.2 $\mathcal{S}$-RPBP Task

In the second task, we considered another continual variation of RPBP. In this variation, the characteristic per-step reward distributions of the states change over time, such that the agent is required to continually adapt and find the state with the better CVaR (given a fixed risk attitude, τ). More formally, this task can be viewed as a continual learning task with a changing state-space, such that a given state is effectively replaced with a state with a different per-step reward distribution. In other words, we have a continual learning task, $\{\mathcal{M}^k\}$, such that $\mathcal{M}^k \doteq \langle \mathcal{S}_k, \mathcal{A}, \mathcal{R}, p \rangle$. For a given $\mathcal{S}_k$, the red world state reward distribution is characterized as a Gaussian distribution with mean μ_{red} and standard deviation σ_{red}. Conversely, the blue world state reward distribution is characterized as a mixture of two Gaussian distributions with means $\mu_{\text{blue-a}}$ and $\mu_{\text{blue-b}}$, standard deviations, $\sigma_{\text{blue-a}}$ and $\sigma_{\text{blue-b}}$, and a mixing coefficient of 0.5.

In the experiment performed, we set $k \in \{1, 2, 3\}$. For all k and all states, we set the standard deviation to 0.05. For $k = 1$, we set $\mu_{\text{red}} = -0.7$, $\mu_{\text{blue-a}} = -1.0$, and $\mu_{\text{blue-b}} = -0.2$. For $k = 2$, we set $\mu_{\text{red}} = -1.5$, $\mu_{\text{blue-a}} = -1.25$, and $\mu_{\text{blue-b}} = -1.0$. For $k = 3$, we set $\mu_{\text{red}} = -0.5$, $\mu_{\text{blue-a}} = -0.9$, and $\mu_{\text{blue-b}} = -0.5$. The indexing function, ω, was defined such that $\omega(t) = 1$ for $t < 40{,}000$, $\omega(t) = 2$ for $40{,}000 \leq t < 80{,}000$, and $\omega(t) = 3$ otherwise.

In terms of the hyperparameters used with the RED CVaR Q-learning algorithm, we used the tuned hyperparameters from Rojas and Lee (2025). That is, $\alpha =$ 2e-2, $\alpha_{\text{CVaR}} \doteq \eta_{\text{CVaR}}\alpha$, where $\eta_{\text{CVaR}} =$ 1e-1, and $\alpha_{\text{VaR}} \doteq \eta_{\text{VaR}}\alpha$, where $\eta_{\text{VaR}} =$ 1e-1. We used a fixed CVaR parameter, τ, of 0.25, an ε-greedy policy with a fixed epsilon of 0.1, and set all initial guesses to zero. The results for this *$\mathcal{S}$-RPBP* task are shown in Figure 3.

B.3 Compute Time and Resources

The experiments were conducted on an Apple M1 SoC (8-core CPU). For the τ-RPBP experiment, each run took roughly 6 seconds to compute on a single CPU, for an approximate total of 5 CPU minutes across the 50 runs. For the $\mathcal{S}$-RPBP experiment, each run took roughly 4 seconds to compute on a single CPU, for an approximate total of 40 CPU seconds across the 10 runs.

C Discounted MDPs and Plasticity

In this appendix, we provide a brief discussion that explores some of the challenges that discounted MDPs (Puterman, 1994) face when trying to maintain a non-zero level of risk-aware plasticity (as defined in Equation (2)) in continual learning settings. More formally, a finite discounted MDP is the tuple $\mathcal{M} \doteq \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p, \gamma \rangle$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathcal{R} \subset \mathbb{R}$ is a bounded set of rewards, $p : \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S} \to [0, 1]$ is a probabilistic transition function that describes the dynamics of the environment, and $\gamma \in [0, 1)$ is the discount factor. At each discrete time step, $t = 0, 1, 2, \ldots$, an agent chooses an action, $A_t \in \mathcal{A}$, based on its current state, $S_t \in \mathcal{S}$, and receives a reward, $R_{t+1} \in \mathcal{R}$, while transitioning to a (potentially) new state, S_{t+1}, such that $p(s', r \mid s, a) = \mathbb{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$. In a discounted MDP, an agent aims to find a policy, $\pi : \mathcal{S} \to \mathcal{A}$, that optimizes the expected discounted sum of rewards. More formally, following the notation established in Section 4.4, for a given (discounted) MDP, $\mathcal{M}^k \in \{\mathcal{M}^k\}_{k=1}^{\infty}$, the performance of a stationary policy, π_t, can be characterized at discrete time step N (where $N > n_0 \geq 0$) via the state-action value function, $Q_N^{\pi_t}$, for all $s \in \mathcal{S}_k$ and $a \in \mathcal{A}_k$ as follows:

$$Q_N^{\pi_t}(s, a) \doteq \mathbb{E}\left[\sum_{n=n_0}^{N-1} \gamma^{n-n_0} R_{n+1} \mid S_{n_0} = s, A_{n_0} = a, A_{n_0+1:N-1} \sim \pi^{\omega(n_0)}\right]. \tag{C.1}$$

Accordingly, as with the average-reward objective (4), we can consider the risk-aware analogue to the (risk-neutral) discounted objective (C.1):

$$\rho_N^{\pi_t}(s, a) \doteq \rho_\tau\left[\sum_{n=n_0}^{N-1} \gamma^{n-n_0} R_{n+1} \mid S_{n_0} = s, A_{n_0} = a, A_{n_0+1:N-1} \sim \pi^{\omega(n_0)}\right]. \tag{C.2}$$

That is, the risk-aware objective (C.2) aims to optimize some measure of risk pertaining to the distribution of discounted sum of rewards induced when following policy π_t from time step n_0 up until time step N. Here, $\rho_N^{\pi_t}$ can be interpreted as a risk assessment (see Definition 4.2) performed at time step N via a static risk measure (see Definition A.2). In Proposition 4.5, we showed that static risk measures cannot maintain a non-zero level of risk-aware plasticity after time step N. Building on this result, we will now show how utilizing the discounted MDP formulation can induce challenges when trying to maintain a non-zero level of risk-aware plasticity even before time step N.

In particular, consider Figure C.1, which evaluates the $\gamma^\ell \doteq \gamma^{n-n_0}$ term in the risk-aware objective (C.2) for $\ell \geq 0$ across various discount factors, $\gamma \in \{0.5, 0.75, 0.9, 0.99, 0.999\}$. As shown in the figure, the γ^ℓ term evaluates to ≈ 0 for all the discount factors considered by $\ell = 6000$. Accordingly, rewards (i.e., observations) observed after $\ell = 6000$ effectively cease to affect the risk assessment. This implies that the agent is unable to maintain a non-zero level of risk-aware plasticity for all discount factors considered for any $\ell \geq 6000$. As such, if the risk assessment is performed beyond this time step, then this implies that utilizing the discounted MDP formulation can induce plasticity challenges even before the (static) risk assessment is performed. Crucially, the average-reward MDP formulation does not utilize the γ^ℓ term, and hence does not face such plasticity challenges.

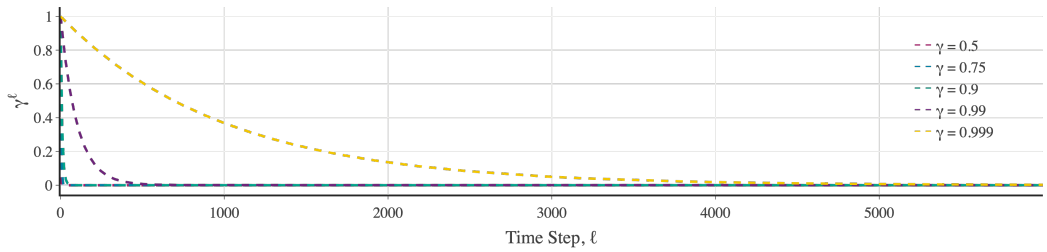

Figure C.1: Evaluation of the $\gamma^\ell \doteq \gamma^{n-n_0}$ term in the risk-aware objective (C.2) for $\ell \geq 0$ across various discount factors, $\gamma \in \{0.5, 0.75, 0.9, 0.99, 0.999\}$. As shown in the figure, the γ^ℓ term evaluates to ≈ 0 for all the discount factors considered by $\ell = 6000$, thereby eliminating the influence of rewards observed after $\ell = 6000$ on the risk assessment.

D Ergodic Risk Measures in Non-Continual Settings

In this appendix, we provide a theoretical analysis that shows that in *non-continual* settings, ergodic risk measures offer a new and appealing alternative for risk-aware learning and decision-making. First, we show that the risk-aware RL objective (5) (which corresponds to an ergodic risk measure) can retain some notion of time consistency while remaining highly interpretable, thereby capturing the appeal of both static and nested risk measures. Then, we show that the risk-aware RL objective (5) is able to circumvent some of the challenges and non-trivialities that have been historically associated with risk-aware RL objectives in non-continual settings.

For convenience, we simplify our notation for the risk-aware RL objective (5) in non-continual settings as follows:

$$\bar{\rho}^{\pi}(s) \doteq \lim_{h\to\infty} \frac{1}{h} \sum_{n=1}^{h} \rho[R_n \mid S_0 = s, A_{0:n-1} \sim \pi], \tag{D.1}$$

such that it can be evaluated at time t as:

$$\bar{\rho}^{\pi}_t(s) = \frac{1}{t} \sum_{n=1}^{t} \rho[R_n \mid S_0 = s, A_{0:n-1} \sim \pi], \tag{D.2}$$

where π denotes some policy and $\rho[R_n|\cdot]$ corresponds to a risk functional which captures some distributional characteristic of the per-step rewards (e.g. variance, CVaR, etc.) with respect to some fixed risk attitude. Here, the non-continual risk-aware objective (D.1) can be viewed as being a specific instance of the more general risk-aware objective (5). Accordingly, given that we showed in Theorem 4.13 that the risk-aware objective (5) corresponds to an ergodic risk measure, it directly follows that the non-continual risk-aware objective (D.1) also corresponds to an ergodic risk measure. This is formalized as Corollary D.1 below:

Corollary D.1. *Given an appropriate ergodicity-like assumption, such as Assumption 4.9 or 4.10, and a stationary policy, π, the non-continual risk-aware objective* (D.1) *corresponds to an ergodic risk measure, as defined in Definition 4.7.*

Proof. This follows directly from Theorem 4.13. □

For clarity, we note that as with the continual risk-aware objective (5), when we apply an appropriate ergodicity-like assumption, such as Assumption 4.9 or 4.10, the risk-aware objective (D.1) becomes independent of prior conditions (i.e., $\bar{\rho}^{\pi}(s) = \bar{\rho}^{\pi}$).

D.1 Ergodic Risk Measures Retain Some Notion of Time Consistency

In this section, we show that ergodic risk measures can retain some notion of time consistency while remaining highly interpretable, thereby capturing the appeal of both static and nested risk measures. In particular, we will show below that the risk-aware RL objective (D.1) satisfies a weaker notion of time consistency, which we call *local time consistency*:

Definition D.2 (Local Time Consistency)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space with filtration $(\mathcal{F}_t)_{t=0}^{\infty}$, and let $\{\rho_t\}_{t=0}^{\infty}$ denote a sequence of conditional risk measures, such that $\rho_t : L^{\infty}(\mathcal{F}_{\infty}) \to L^{\infty}(\mathcal{F}_t)$. The sequence $\{\rho_t\}_{t=0}^{\infty}$ is said to satisfy the* local time consistency *property if there exists a time step, $n \geq 0$, and a possibly infinite horizon length, $m \in \mathbb{N} \cup \{\infty\}$, such that, for all $n \leq t < n + m$ and all $X, X' \in L^{\infty}(\mathcal{F}_{\infty})$, we have that: $\rho_{t+1}(X) \leq \rho_{t+1}(X') \implies \rho_t(X) \leq \rho_t(X')$. That is, time consistency holds within some subset of the time-horizon. Note that when $n = 0$ and $m \to \infty$, this reduces to the standard definition of time consistency (Definition A.5).*

In essence, the above definition for local time consistency only requires that time consistency holds for some subset of the history, rather than the entire history. We note that such a notion of time consistency could be useful in a continual learning setting as it could provide some measure of *stability*. That is, while we want the agent to have the flexibility to change its risk preferences over time, it would likely be problematic if the agent changed its risk preferences at every time step.

We will also make use of a relaxed variant of this property, which we refer to as *ϵ-local time consistency*, to account for the fact that in RL we are often dealing with approximated risk evaluations up to some arbitrary error threshold:

Definition D.3 (ϵ-Local Time Consistency)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space with filtration $(\mathcal{F}_t)_{t=0}^{\infty}$, and let $\{\rho_t\}_{t=0}^{\infty}$ denote a sequence of conditional risk measures, such that $\rho_t : L^{\infty}(\mathcal{F}_{\infty}) \to L^{\infty}(\mathcal{F}_t)$. The sequence $\{\rho_t\}_{t=0}^{\infty}$ is said to satisfy the* ϵ-local time consistency *property if, for a given $\epsilon > 0$, there exists a time step, $n \geq 0$, and a possibly infinite horizon length, $m \in \mathbb{N} \cup \{\infty\}$, such that, for all $n \leq t < n + m$ and all $X, X' \in L^{\infty}(\mathcal{F}_{\infty})$, we have that: $\rho_{t+1}(X) \leq \rho_{t+1}(X') \implies \rho_t(X) \leq \rho_t(X') + \epsilon$. That is, time consistency holds within some subset of the time-horizon up to an error threshold, ϵ.*

We will now show that, under ergodicity-like assumptions, the risk-aware RL objective (D.1) satisfies the ϵ-local time consistency property, such that $\epsilon \to 0$ as $t \to \infty$:

Proposition D.4. *Given an appropriate ergodicity-like assumption, such as Assumption 4.9 or 4.10, and a stationary policy, π, the risk-aware objective* (D.1) *satisfies the ϵ-local time consistency property (as defined in Definition D.3), such that $\epsilon \to 0$ as $t \to \infty$.*

Proof. Let $\bar{\rho}_t^{\pi}$ denote the risk evaluated at time t (as per Equation (D.1)) under the stationary policy. Under the ergodicity-like assumption, we have, by Birkhoff's Ergodic Theorem (Birkhoff, 1931), that $\bar{\rho}_t^{\pi}$ converges to a stationary value as $t \to \infty$. This implies that there exists a finite time step, j, such that for all $t \geq j$, $\bar{\rho}_t^{\pi}$ converges to a stationary value. As such, for all $t \geq j$, the evaluated risk at time t becomes arbitrarily indistinguishable from the evaluated risk at time $t + 1$, such that $|\bar{\rho}_{t+1}^{\pi}(X) - \bar{\rho}_t^{\pi}(X)| \leq \tilde{\epsilon}$ with respect to some error threshold, $\tilde{\epsilon}$, where $\tilde{\epsilon} \to 0$ as $t \to \infty$. Accordingly, for any future sequences, X and X', it directly follows that $\bar{\rho}_{t+1}^{\pi}(X) \leq \bar{\rho}_{t+1}^{\pi}(X') \implies \bar{\rho}_t^{\pi}(X) \leq \bar{\rho}_t^{\pi}(X') + 2\tilde{\epsilon}$ for all $t \geq j$, thereby satisfying the ϵ-local time consistency property (where $\epsilon = 2\tilde{\epsilon}$), such that $\tilde{\epsilon} \to 0$ as $t \to \infty$. □

As such, with Proposition D.4, we have shown that our interpretable risk-aware objective (D.1), satisfies some notion of time consistency, thereby capturing the appeal of both static (i.e., interpretability) and nested (i.e., time consistency) risk measures in the non-continual setting.

D.2 Ergodic Risk Measures and Optimal Policies

In this section, we show that the risk-aware RL objective (D.1) is able to circumvent some of the challenges and non-trivialities that have been historically associated with risk-aware RL objectives in non-continual settings. In particular, it is well-known that risk-aware RL objectives often admit non-Markovian optimal policies (Bellemare et al., 2023; Bowling et al., 2023). However, we will show below that the risk-aware objective (D.1), which corresponds to an ergodic risk measure, provably admits a stationary, deterministic optimal policy, given the following assumption:

Assumption D.5. *Consider the risk functional in Equation* (D.1)*, ρ, which captures some distributional characteristic of the limiting per-step reward distribution induced when following a given policy. We assume that ρ is bounded, and that, under an ergodicity-like assumption, the limiting per-step reward distribution induced by the policy contains all the information required to evaluate ρ.*

In essence, the above assumption ensures that the per-step reward distribution contains all the information needed to evaluate the risk functional, ρ, at a given time step. In the context of the average-reward MDP formulation, we find this to be a reasonable assumption to make, given that in the risk-aware setting, we aim to find a policy that induces a limiting per-step reward distribution that is optimal with respect to some measure of risk. For example, if our risk functional corresponds to the variance of the distribution (such that our aim is to find a policy that induces a limiting per-step reward distribution with the best long-run variance), then Assumption D.5 implies that if we know this distribution, then we can compute its variance without needing additional information.

Next, given Assumption D.5, we can make use of Theorem 8.9.3 of Puterman (1994) to prove the following result:

Theorem D.6. *Consider the risk-aware objective* (D.1)*. Suppose that, under an appropriate ergodicity-like assumption, such as Assumption 4.9 or 4.10, this objective exists and is unique. Furthermore, suppose that Assumption D.5 holds, and that there exists some policy that is optimal with respect to the risk-aware objective* (D.1)*. Then, there exists a stationary, deterministic policy that is optimal with respect to the risk-aware objective* (D.1)*.*

Proof. Consider the risk functional from Equation (D.1) under the ergodicity-like assumption: $\rho[R_n \mid A_{0:n-1} \sim \pi]$. Given Assumption D.5, we know that this risk functional can be computed entirely based on information from the per-step rewards, R_t. Consequently, given that the per-step rewards are computed based on the current states and actions (i.e., S_t and A_t), the risk functional can be viewed as the following mapping: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action-spaces of an average-reward MDP. Accordingly, we can interpret ρ as a ("risk-aware") reward function, in the usual MDP sense, such that the risk-aware objective can be written as follows:

$$\bar{\rho}^{\pi} \doteq \lim_{h \to \infty} \frac{1}{h} \sum_{n=1}^{h} \rho[R_n \mid A_{0:n-1} \sim \pi] \tag{D.3a}$$

$$= \sum_{s \in \mathcal{S}} \mu_{\pi}(s) \sum_{a \in \mathcal{A}} \pi(a|s) \hat{r}(s,a), \tag{D.3b}$$

where $\hat{r}(s,a)$ denotes the risk-aware reward function (for some $s \in \mathcal{S}, a \in \mathcal{A}$) that corresponds to the risk functional, ρ, and μ_{π} denotes the limiting distribution of states induced by policy π (we know that μ_{π} exists and is unique given the ergodicity-like assumption).

As such, we can see from Equation (D.3) that optimizing our risk-aware objective, $\bar{\rho}^{\pi}$, amounts to optimizing a standard average-reward MDP objective with a modified (i.e., "risk-aware") reward function that satisfies the requirements of reward functions for average-reward MDPs. Consequently, it follows from Theorem 8.9.3 of Puterman (1994) that if there exists some policy that is optimal with respect to $\bar{\rho}^{\pi}$, then, there exists a stationary, deterministic policy that is optimal with respect to $\bar{\rho}^{\pi}$. We note that Theorem 8.9.3 of Puterman (1994) assumes finite state and action-spaces, as well as bounded rewards. It can be easily shown that these assumptions are satisfied by Assumption D.5 and by the MDP formulation considered in this work (see Section 4.4). This completes the proof. □

In essence, Theorem D.6 states that if an optimal policy exists for the risk-aware objective (D.1), then there exists a stationary, deterministic policy that is also optimal with respect to this objective. Intuitively, this follows from Equation (D.3b), which shows that the objective depends neither on time nor the history, thereby making a stationary, deterministic policy sufficient for optimality.

Finally, we can leverage Theorems 8.4.5 or 9.1.8 of Puterman (1994) to show that an optimal policy indeed exists:

Theorem D.7. *Consider the risk-aware objective* (D.1)*. Suppose that, under an appropriate ergodicity-like assumption, such as Assumption 4.9 or 4.10, this objective exists and is unique. Furthermore, suppose that Assumption D.5 holds. Then, there exists a policy that is optimal with respect to the risk-aware objective* (D.1)*.*

Proof. We showed in the proof for Theorem D.6 how optimizing our risk-aware objective, $\bar{\rho}^{\pi}$, ultimately amounts to optimizing a standard average-reward MDP objective with a modified reward function that satisfies the requirements of reward functions for average-reward MDPs. Consequently, it directly follows from Theorems 8.4.5 or 9.1.8 of Puterman (1994) (depending on which ergodicity-like assumption is used) that there exists some policy that is optimal with respect to $\bar{\rho}^{\pi}$. We note that these theorems from Puterman (1994) assume finite state and action-spaces, as well as bounded rewards. It can be easily shown that these assumptions are satisfied by Assumption D.5 and by the MDP formulation considered in this work (see Section 4.4). This completes the proof. □

As such, with Theorems D.6 and D.7, we have shown that the risk-aware objective (D.1), which corresponds to an ergodic risk measure, provably admits a stationary, deterministic optimal policy, thereby circumventing some of the challenges and non-trivialities that have been historically associated with risk-aware RL objectives in non-continual settings. We conclude this section with a remark related to the interpretation of the risk-aware reward function, $\hat{r}(s,a)$:

Remark D.8. *At first glance, the interpretation of the risk functional, ρ, as a reward function, $\hat{r}(s,a)$, in Equation* (D.3) *may give the impression that we are "averaging away" the risk from our objective. However, this is far from reality. In particular, Equation* (D.3) *can be thought of as computing the "time-averaged risk" associated with the limiting per-step reward distribution (e.g. the average long-run variance or the average long-run CVaR of R_t). Accordingly, the objective* (D.1) *precisely captures some (time-averaged) distributional characteristic associated with the limiting per-step reward distribution beyond the mean, thereby making it a risk-aware objective.*

E Ergodicity and Continual Learning

In this appendix, we provide a mathematically rigorous discussion on the relationship between ergodicity and continual learning, in which we show that ergodicity-like assumptions, such as Assumptions 4.9 or 4.10, do not preclude an agent's ability to be a continual learner. Given that this discussion is risk-agnostic (i.e., it equally applies to risk-neutral and risk-aware settings), we will limit our discussion to risk-neutral agents; however, the same arguments apply to risk-aware agents.

To make our arguments, we will leverage the language and formalism introduced in Abel et al. (2023), in which an instance of RL is considered to be an instance of continual RL (CRL) if, informally, the best agent never stops learning. To this end, we now provide key definitions from Abel et al. (2023):

Definition E.1 (Definition 2.1 of Abel et al. (2023))**.** *An* agent-environment interface *is the pair* $(\mathcal{A}, \mathcal{O})$*, such that* $|\mathcal{A}| \geq 2$ *and* $|\mathcal{O}| \geq 1$*, where* $\mathcal{O}$ *and* $\mathcal{A}$ *denote observation and action-spaces.*

Definition E.2 (Definition 2.2 of Abel et al. (2023))**.** *The* histories *with respect to interface* $(\mathcal{A}, \mathcal{O})$ *are the set of sequences of action-observation pairs,*

$$\mathcal{H} = \bigcup_{t=0}^{\infty} (\mathcal{A} \times \mathcal{O})^t. \tag{E.1}$$

Definition E.3 (Definition 2.3 of Abel et al. (2023))**.** *An* environment *with respect to interface* $(\mathcal{A}, \mathcal{O})$ *is a function,* $e : \mathcal{H} \times \mathcal{A} \to \Delta(\mathcal{O})$*, where* $\Delta(\mathcal{O})$ *denotes the probability simplex over the set of observations,* $\mathcal{O}$*.*

Definition E.4 (Definition 2.4 of Abel et al. (2023))**.** *An* agent *with respect to interface* $(\mathcal{A}, \mathcal{O})$ *is a function,* $\lambda : \mathcal{H} \to \Delta(\mathcal{A})$*, where* $\Delta(\mathcal{A})$ *denotes the probability simplex over the set of actions,* $\mathcal{A}$*.*

Definition E.5 (Definition 2.5 of Abel et al. (2023))**.** *The* realizable histories *of a given agent-environment pair,* (λ, e)*, define the set of histories of any length that can occur with non-zero probability from the interaction of* λ *and* e*, such that:*

$$\mathcal{H}^{\lambda,e} = \overline{\mathcal{H}} = \bigcup_{t=0}^{\infty} \left\{ h_t \in \mathcal{H}_t : \prod_{k=0}^{t-1} e(o_{k+1} \mid h_k, a_k) \lambda(a_k \mid h_k) > 0 \right\}. \tag{E.2}$$

Definition E.6 (Definition 2.8 of Abel et al. (2023))**.** *Let* Λ *denote a set of agents and* $\mathcal{E}$ *denote a set of environments. The* performance*,* $\upsilon : \mathcal{H} \times \Lambda \times \mathcal{E} \to [v_{min}, v_{max}]$*, is a bounded function for fixed constants,* $v_{min}, v_{max} \in \mathbb{R}$*.*

In words, $\mathcal{H}$ denotes the set of all possible finite interaction histories (with $\overline{\mathcal{H}}$ denoting the subset of realizable histories), where a history, $h_t \in \mathcal{H}$, corresponds to the sequence of observations and actions up to time t (this is analogous to the sequences $O_{1:t}$ and $A_{1:t}$, as defined in Section 3.1 of this work). An agent is defined as a mapping, $\lambda : \mathcal{H} \to \Delta(\mathcal{A})$, that outputs a probability distribution over actions given a history. Following Abel et al. (2023), we let Λ denote the set of all such agents considered in a given RL problem. The performance function, $\upsilon(\lambda, e)$, expresses some statistic of the future rewards produced by the interaction between an agent, λ, and an environment, e. We note that the definition of performance (i.e., Definition E.6) can be viewed as a risk-neutral instance of the definition of a risk assessment (see Definition 4.2) proposed in this work.

To formalize the learning process, Abel et al. (2023) utilize the concept of an *agent basis*, $\Lambda_B \subset \Lambda$, which represents a core set of fixed, baseline behaviours (e.g. a finite set of stationary policies or fixed neural network weights) that an agent implicitly searches over as it interacts with the environment. To dictate how an agent switches between these base behaviours, Abel et al. (2023) introduce the concept of a learning rule, which gives rise to the *generates* operator, as follows:

Definition E.7 (Definition 3.2 of Abel et al. (2023))**.** *A* learning rule *over an agent basis,* Λ_B*, is a function,* $\sigma : \mathcal{H} \to \Lambda_B$*, that selects a base agent for each history.*

Definition E.8 (Definition 3.4 of Abel et al. (2023))**.** *We say a basis* Λ_B generates Λ *in* e*, denoted as* $\Lambda_B \models \Lambda$*, if and only if there exists a set of learning rules,* Σ*, such that:*

$$\forall_{\lambda \in \Lambda} \exists_{\sigma \in \Sigma} \forall_{h \in \overline{\mathcal{H}}} \lambda(h) = \sigma(h)(h). \tag{E.3}$$

Intuitively, a basis, Λ_B, generates an agent set, Λ, if every agent in Λ can be understood as simply switching between the base agents (i.e., the agents in Λ_B) according to some learning rule.

Given this search process, an agent is said to "stop learning" if it eventually adopts one of these fixed base behaviours permanently. In particular, an agent, λ, is said to *reach* Λ_B in environment e, denoted as $\lambda \rightsquigarrow \Lambda_B$, where $(\rightsquigarrow)$ denotes the *reaches* operator, if there exists some realizable history after which λ becomes equivalent to some base agent, $\lambda_B \in \Lambda_B$, for all future histories. By contrast, an agent that never permanently settles on a single base strategy is said to never reach the basis, such that $\lambda \not\rightsquigarrow \Lambda_B$. Through this formalism, Abel et al. (2023) propose the following definitions for a continual learning agent and continual RL:

Definition E.9 (Definition 4.1 of Abel et al. (2023))**.** *An agent, λ, is a* continual learning agent *in e relative to Λ_B if and only if the basis generates the agent (such that $\Lambda_B \models \{\lambda\}$) and the agent never reaches the basis (such that $\lambda \not\rightsquigarrow \Lambda_B$).*

Definition E.10 (Definition 4.2 of Abel et al. (2023))**.** *Consider an RL problem, (e, υ, Λ). Let $\Lambda_B \subset \Lambda$ be a basis such that $\Lambda_B \models \Lambda$ and let $\Lambda^* = \arg\max_{\lambda \in \Lambda} \upsilon(\lambda, e)$. We say $(e, \upsilon, \Lambda, \Lambda_B)$ defines a* continual RL *problem if:*

$$\forall_{\lambda^* \in \Lambda^*} \lambda^* \not\rightsquigarrow \Lambda_B. \tag{E.4}$$

Informally, Definition E.10 states that an instance of RL is considered to be an instance of continual RL if the best agent never stops learning.

As such, having formalized the notion of continual RL and a continual learning agent, we are now ready to begin our discussion on the role of ergodicity in continual RL. In particular, we will show that ergodicity-like assumptions, such as Assumptions 4.9 or 4.10, do not preclude an agent's ability to be a continual learner. To make this argument, we will seek to answer the following question:

Suppose that there exists some continual RL (CRL) problem, $(e, \upsilon, \Lambda, \Lambda_B)$, such that it satisfies Definition E.10. Can we determine whether the environment component of this CRL problem, e, satisfies an ergodicity-like assumption, based solely on the fact that we know that the overarching problem satisfies the definition of a CRL problem?

In other words, we want to determine whether ergodicity-like assumptions are strictly incompatible with continual RL. That is, if we know that the problem is an instance of CRL, does it automatically invalidate the possibility that the environment is ergodic (and vice-versa)? In Theorem E.11 below, we answer this question, and show that ergodicity neither implies nor precludes continual learning:

Theorem E.11. *Let $(e, \upsilon, \Lambda, \Lambda_B)$ denote a continual RL problem, such that it satisfies Definition E.10. The fact that this continual RL problem satisfies Definition E.10 neither implies nor precludes the condition that the environment, e, satisfies an ergodicity-like assumption.*

Proof. We observe that the environment must either satisfy an ergodicity-like assumption, or not satisfy an ergodicity-like assumption. Accordingly, we proceed with a proof by contradiction for both mutually exclusive cases:

Case 1: Suppose, for the sake of contradiction, that every valid continual RL problem strictly requires a non-ergodic environment. Now consider an environment with a single state and two available actions, a_1 and a_2. It can be easily shown that this environment satisfies standard ergodicity-like assumptions. Now, let the reward function change periodically, such that a_1 yields the optimal reward for the first n steps, a_2 yields the optimal reward for the next m steps, and so forth, with the pattern repeating indefinitely, such that an optimal agent must continually adapt its policy to track the shifting rewards. Because the agent never permanently settles on a single stationary base behaviour (i.e., $\lambda \not\rightsquigarrow \Lambda_B$), this constitutes a valid continual RL problem within an ergodic environment, thereby contradicting the assumption that CRL strictly implies a non-ergodic environment.

Case 2: Suppose, for the sake of contradiction, that every valid continual RL problem strictly requires an ergodic environment. Now consider an environment consisting of two isolated, absorbing states, each with two available actions, such that switching between the two states is structurally impossible. It can be easily shown that this environment violates any standard ergodicity-like assumption. At time $t = 0$, the agent is placed into one of these states at random. Within each state, the rewards associated with each available action continually change over time. Accordingly, because the optimal action continues to shift indefinitely, the agent must continually adapt its policy and never reaches

a fixed basis (such that $\lambda \not\rightsquigarrow \Lambda_B$). As such, this constitutes a valid continual RL problem within a non-ergodic environment, thereby contradicting the assumption that CRL strictly implies an ergodic environment.

Hence, given that both cases result in a contradiction, we can conclude that the satisfaction of Definition E.10 neither implies nor precludes the condition that the environment, e, satisfies an ergodicity-like assumption. This concludes the proof. □

As such, with Theorem E.11, we have shown that ergodicity and continual learning are not strictly incompatible. We note that although the counterexamples used in the proof for Theorem E.11 both make use of non-stationary reward functions, there are many possible counterexamples that could be considered. For instance, one could consider an agent that has a highly constrained function approximator (such as a linear function approximator) operating in a large, stationary environment. Because the agent cannot represent the entire environment at once, it must continually update its representations as it explores different regions of the environment, thereby constituting an instance of continual RL (as per Definition E.10). Importantly, whether or not the environment satisfies ergodicity-like assumptions would not impact whether this example constitutes an instance of continual RL, given that the agent would be required to update its representations both in (large, stationary) ergodic and (large, stationary) non-ergodic environments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction are supported by theoretical and empirical results presented in the main body.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are communicated via the various assumptions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Theoretical results contain complete proofs and specify relevant assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Reproducibility details are included in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Open-source code is available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experiment details are included in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experiment results show 95% confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details regarding the compute resources are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is foundational research and not tied to particular applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work is foundational research and poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The open-source code includes setup and run instructions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.