Article

Not peer-reviewed version

A Systematic Review of Reinforcement Learning for Dynamic Risk Assessment

Houssem Hosni and Clive Asuai *

Posted Date: 28 January 2026

doi: 10.20944/preprints202601.2192.v1

Keywords: reinforcement learning; risk assessment; risk-sensitive RL; safe reinforcement learning; sequential decision-making; AI safety

A Systematic Review of Reinforcement Learning for Dynamic Risk Assessment

Houssem Hosni[1], Clive Asuai [2]*

[1] Department of Computer Engineering, Université de La Rochelle, 17000 La Rochelle, France

[2] Department of Cyber Security, Delta State Polytechnic, Otefe-Oghara, 333106, Nigeria

* Correspondence: clive.asuai@ogharapoly.edu.ng; Tel.: +2347033574980

Abstract

Traditional risk assessment méthodologies are often inadequate in dynamic environments due to their reliance on static historical data. A viable substitute for adaptive, sequential decision-making in the face of uncertainty is Reinforcement Learning (RL). However, the research landscape is fragmented, lacking a unified framework to guide the selection of RL paradigms, such as risk-sensitive, safe, and robust RL, for specific risk categories. To close this gap, this review examines RL's use in risk assessment in a methodical manner. We define a conceptual framework for classifying risk-aware reinforcement learning techniques, compare and contrast their advantages and disadvantages, and pinpoint the main obstacles to dependable implementation. The substantial promise of RL is confirmed by our study, which is based on a systematic evaluation of recent literature (2018–2024) in the fields of finance, autonomous systems, and healthcare. Major issues still exist, nevertheless, such as sample inefficiency, performance-safety trade-offs, and a discrepancy between theoretical assurances and actual dependability. We come to the conclusion that creating hybrid models, setting strict standards, and giving safety and robustness top priority will be necessary for future development in order to facilitate deployment in high-stakes, real-world situations.

Keywords: reinforcement learning; risk assessment; risk-sensitive RL; safe reinforcement learning; sequential decision-making; AI safety

1. Introduction

Modern decision-making in a wide range of crucial fields, such as engineering, cybersecurity, healthcare, and finance, is based on risk assessment. The complexity of modern systems is posing a growing challenge to traditional approaches, which frequently rely on statistical models and historical data analysis. They usually have trouble capturing high-dimensional interactions between risk factors, adapting to dynamic, non-stationary situations, and making the best sequential judgments in the face of extreme uncertainty. Recent failures to foresee systemic financial stress and model the spread of pandemics demonstrate how this inherent inflexibility can result in substantial blind spots and delayed responses to new threats.

These drawbacks have led to an increasing interest in Reinforcement Learning (RL) for risk assessment. In contrast to traditional techniques, RL offers a framework for creating adaptive systems that, with the help of a reward signal, interact with their surroundings to learn the best decision-making policies. This trial-and-error process gives RL special characteristics, such as the capacity to adjust to shifting circumstances, consider long-term effects, and come up with new tactics without direct human guidance. RL presents the possibility of developing dynamic and extremely complex risk assessment systems that may traverse constant stages of uncertainty in fields ranging from algorithmic trading to customized medicine by presenting risk management as a sequential decision-making process

Bottom of Form

 doi:10.20944/preprints202601.2192.v1

However, the burgeoning research in this area is highly fragmented, with applications and methodologies spread across disparate fields. This leads to a significant problem: a lack of a unified understanding of how different RL paradigms, from value-based methods to policy gradient architectures, are being tailored to assess and mitigate risk. Without a consolidated view, it is difficult to identify best practices, common pitfalls, and generalizable principles, thereby hindering the systematic advancement of RL in this critical application area.

This review lays forth a precise scope and goal to fill this gap. In addition to identifying the main obstacles to its broad adoption, it seeks to critically analyze the use of RL as a novel paradigm for risk assessment in order to synthesize its substantial potential for improving decision-making and predictive accuracy. The study conducts a systematic literature review of recent research applying RL to various risk domains, following the PRISMA guidelines where applicable. The analysis focuses on the architectures of RL agents, the formulation of risk-related reward functions, and the environments used for training and validation, drawing insights from studies published between 2018 and 2024.

The primary contribution of this work is a novel conceptual framework and taxonomy for risk-aware RL. This is operationalized through (1) comparative tables that juxtapose different RL approaches against key risk dimensions and (2) a conceptual map that illustrates the relationships between risk types, suitable RL algorithms, and reward function designs. This framework directly addresses the fragmentation in the field by providing researchers and practitioners with a structured guide for selecting and developing appropriate RL models for specific risk contexts.

2. Conceptual Foundations and Analytical Dimensions

Incorporating reinforcement learning (RL) in risk verification tools requires a formal understanding of how the notion of "risk" is perceived and implemented from the RL perspective. At one level, it is important to make a clear distinction between risk and uncertainty. Although they are frequently interchanged in common (non-specialist) language, their technical meanings are quite different from each other. In the usual usage, risk comes where the probability distributions over outcomes are known or can be estimated [1,2]. In contrast, uncertainty, especially epistemic uncertainty, pertains to when the underlying distributions themselves are unknown or only partially observable [3]. Under the broad spectrum of uncertainty, there are generally two kinds of uncertainty differentiated: aleatoric uncertainty, inherent randomness that is not controllable by us; and epistemic uncertainty, a reflection of our incompleteness in knowledge of a system or environment [4]. Risk sensitivity in RL needs to consider both, especially if the agents need to perform under dynamic and data-limited scenarios.

In reinforcement learning (RL), the formalism is often grounded in Markov Decision Processes (MDPs), where an agent takes actions to maximize cumulative rewards by interacting with a stochastic environment. It's a standard RL objective aiming to maximize the expected return and thus is a kind of risk-neutral [5]. This implies that all outcomes are handled naively on average basis, no matter how slanting they vary. In practice it is, however, unlikely that this is enough for many applications that are restricted by safety constraints, operational limits or very expensive failure events. Therefore, the objective of risk-aware RL is to change the traditional optimization targets to reflect both the expectation and its distributional properties, such as variance, tail risk or worst case [6,7]. Risk can therefore be incorporated at different stages of the RL pipeline: modified reward functions, risk-adjusted value functions, or uncertainty in transition dynamics (e.g., robust RL formulations) etc.

In order to label an RL algorithm as risk-aware, we need to characterize how risks are taken into account by the method. There are a multiple ways by which these approaches can be distinguished. Firstly, there is the risk metric, that is defined in the taper of the learning objective, as for instance through variance, Value-at-Risk (VaR) or Conditional Value-at-Risk (CVaR) [8,9]. A third line of works is the role of constraints at execution, where Constrained MDPs (CMDPs) are used to impose safety or performance bounds during learning [10]. These algorithms try to adversarial optimize

 doi:10.20944/preprints202601.2192.v1

over policies where the environment is treated as uncertain or even hostile, and compute policies that are robust to worst-case environments. Safe RL methods further seek to not violate given safety constraints in learning and/or deployment (Lyapunov functions, safety layers, action shielding) [11,12].

To perform a critical and comparative analysis, a method should be introduced based on a number of dimensions that can make the assessment systematically of the approaches currently in use. The first level is related to the nature of the risk addressed. Quantitative risk is defined as a numerical variance from an expected return which can be modelled using statistical measures [13]. It may be the case that qualitative risk consists of making category judgments on factors or of considering binary safety states, which are naturally modelled based on logic or rule [14]. A third class, structural risk, consists of risks that are rooted on the specific structure of the model, such as approximations errors, unmodeled dynamics or non-stationarities, that are typically tackled through robust or adaptive learning schemes [15,16].

A second dimension relates to the optimization target of the RL agent. Unlike standard RL where the objective is to maximize expected cumulative reward, risk-aware RL could aim at minimizing other risk measures including CVaR, value-at-risk, probabilistic constraint satisfaction and guaranteeing minimum performance [17,18]. These goals lead to very different learning behaviours. For example, CVaR-based optimization in BEH would concentrate on lower tail of the return distribution to enhance outcomes in bad markets pushing the agent conservative in high-risk environments. In contrast, robust optimization often returns policies trading the average-case performance for stability under model perturbations [20].

The third dimension of analysis concerns the assumptions of the environment and model. In standard RL a world transition model is assumed to be known and stationary and full observability is required. In contrast, risk-sensitive settings can be characterized by model uncertainty, partial observability or distributional shifts, and many call for algorithms that can instead adapt or generalize beyond their training distributions [21,22]. The robust RL assumes that an uncertainty set on the transition model is provided, and it searches for the policies that are good for all possible dynamics in that set [23]. Besides, Bayesian and Distributional RL are also different alternatives for capturing epistemic and aleatoric uncertainty using either probabilistic reasoning or by modelling return distributions [24].

A fourth essential dimension concerns the compromise risk-aware RL needs to deal with. Reward maximization and safety are examples of such a trade-off. Adding hard constraints on safety can cause conservative behaviour and low exploration, which is why the achievable reward may be lower [25]. However, maximizing expected return too aggressively may lead the agent to catastrophic failures. Another trade-off is the one between performance and robustness: with robust methods that generalize better in uncertain cases but can become too pessimistic when scenarios are nominal [26,27]. There is also this inherent \"exploration-smart\" trade off especially in situations with unsafe exploration that could lead to irreparable damage. This is alleviated by the safe exploration techniques by limiting the policy space, which can lead to lower convergence speed and flexibility [28].

These analytical lines provide a structured perspective to reason about how the various RL algorithms perceive and treat the concept of risk. They also serve as a reference for comparative analysis among methods, and to show the different domains of applications, the levels of uncertainty and the operational requirements. Organizing related research along these dimensions enables us to identify methodological tendencies, deficiencies in classical models and potential for combination or novel frameworks to more fully capture the trade-offs in risk-aware decision-making [11,29,31].

 doi:10.20944/preprints202601.2192.v1

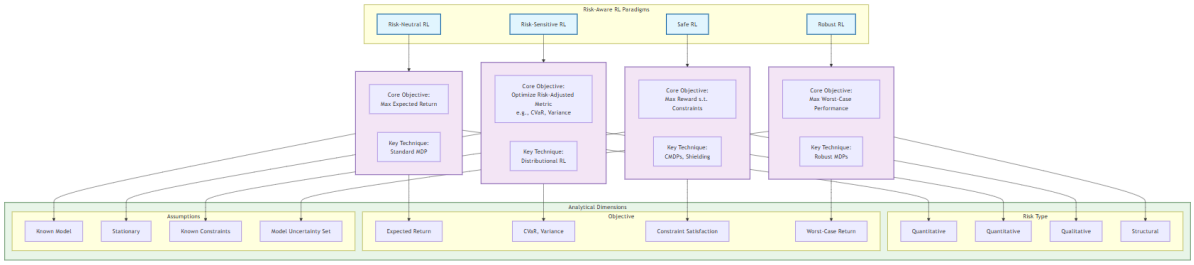


Figure 1. The high level taxonomy map.

3. Classification and Comparative Analysis of Risk-Aware RL Approaches

The classification of RL algorithms on risk assessment is a foundation to organize the various paradigms that have appeared to cope with safety, robustness and uncertainty in decision making. These approaches are disparate in motivation, mathematics and operational premises. There were four main classes in this type: risk-neutral RL, risk-sensitive RL, robust RL, and safe RL. Each type of distribution embodies a different approach to dealing with uncertainty and risk, and their comparison is useful for illuminating their different strengths and weaknesses when used for risk-based applications.

Risk-neutral RL is the classical paradigm of reinforcement learning. In this sense, the agent only cares about obtaining the cumulative reward, without taking into consideration or trying to model the variance or distribution of rewards. The implicit idea is that the expected value is a complete summary statistic for decision-making. This strategy is especially efficient in the presence of low or acceptable variability of performance or in a situation where the cost of not desired events is not disproportionately high. Yet, from a risk assessment standpoint, risk-neutral RL has some key shortcomings. It misses ways to discriminate between policies that have similar expected returns while differing widely in terms of variance or tail risks. Therefore, that paradigm cannot cope with rare but important events that may incur a large cost, and is inadequate in what concerns domains in which risk averse behaviour, reliability, or safety assurances are demanded.

Risk-sensitive RL attempts to overcome these limitations by involving a certain degree of risk-aversion in the objective. Instead of the expected return, these algorithms focus on risk-adjusted performance measures such as the Conditional Value-at-Risk (CVaR), exponential utility functions, or distortion risk measures. For instance, the CVaR optimal criterion is based on the worst-case scenario, i.e., the expected value of the worst returns amongst the specified confidence level, offering therefore a control of the lower tail of the return. Another approach which offers an alternative is exponential utility functions, which penalise variability by rewarding consistently good performance versus extremely good rewards. DRMs alter the weights of various outcomes in the return distribution, and hence constitute a flexible framework to represent diverse risk attitudes. Such methods enable the agent to actively factor in uncertainty in returns and bias the learning process towards safer policies in high stakes environments. Yet the landscape of risk-sensitive methods is generally worse and could be conflicting or taking convergence issues depending on the form of reward distribution and settings of the sensitivity parameter. Furthermore, their performance relies on tail distribution estimation, which is difficult in data poor or non-stationary contexts.

Robust RL provides an additional perspective on risk-awareness by considering model uncertainty and worst-case performance. Instead of relying on a known or accurate model of the environment, robust RL approaches consider a set of plausible models, often referred to as an uncertainty set, and seek a policy that works well under all such models. This is typically cast as a min-max optimization problem, where the agent tries to maximize its return in the assumption that the environment picks the worst of all possible dynamics. The strong RL framework is particularly meaningful to applications in which system parameters are incompletely known or can be changed

 doi:10.20944/preprints202601.2192.v1

by some unknown external circumstances. A well-known formalism in this realm is that of an adversarial MDP, where transition probabilities or reward functions are adversarial corrupted up to some bounded threshold. Although robust RL gives strong guarantees under environmental perturbations, it tends to produce unduly conservative policies at the expense of average performance. In addition, the definition of the uncertainty set has a great impact on the solution effectiveness of the algorithm; too large sets produce "bad" solutions for pessimistic simulations and too small ones might not account for the important dynamic's deviations.

Safe reinforcement learning has an emphasis on the need of an agent to satisfy certain safety criteria during learning and deployment. In this regime, one seeks to prevent the system from entering failure states or violating formal safety constraints, regardless of any potential reward. One popular formalism for safe RL is Constrained Markov Decision Pro- cesses (CMDPs): the safe agent aims to maximize rewards under the constraints of some cost function(s). Other techniques include barrier functions that introduce safety margins to the optimization and shielding methods that veto unsafe actions in favour of actions that are verifiably safe. Such issues are particularly crucial in real-world systems where exploration can threaten physical assets or human users – for example, in the case of autonomous vehicles, medical decision-making systems, or industrial automation. Though safe RL offers a principled framework for safety guarantees, its use of constraint satisfaction and conservative exploration can significantly slow down learning and restrict the policy's flexibility. Constructing such constraints which are both practical and computable is an important challenge, in particular in high-dimensional or partially observed settings.

For a systematic comparison of these approaches, it is helpful to analyse them along the analytical lines that have been introduced before: kind of risk treated, objective of optimization, environmental assumptions, trade-offs. Risk-neutral RL: Risk neutral RL, meanwhile, does not consider any particular form of risk and works under the assumptions of known and stationary environment maximizing expected returns with no risk constrained. Risk-sensitive RL addresses quantitative risk by adapting the optimization criterion based on statistical measures, and remaining under the assumption of a stationary environment. Robust RL mitigates structural and epistemic risk through sufficient relaxation of environment certainty to optimize for the worst-case without fixing the model parameters, even accepting their uncertainty. Safe RL considers qualitative and quantitative safety specifications, where external safety constraints are incorporated directly into learning, and, generally, more stringent assumptions on admissible behaviour during exploration are considered.

Each method has unique strengths and drawbacks that will make it more applicable than the other in specific applications. Risk neutral RL is suitable for low risk, high volume environments, where average performance outweighs considerations of deviation. Risk-sensitive RL is better for applications in which the stochasticity is quantifiable and the tail risks are measurable, like the financial portfolio management. Strong RL is generally preferred for mission-critical problems in the presence of adversarial perturbations or unknown dynamics, for example in autonomous vehicle navigation in unstructured environments. Safe RL becomes crucial in safety critical systems when hard constraints shall not be violated, not even in learning.

The classification of risk-aware RL methods shows that a single paradigm is not the best one in all risk dimensions or application domains. Each of this face the trade-off between safety, performance, robustness, and generality. Their relative strengths and weakness demonstrate the significance of choosing or creating RL algorithms that cater to particular objectives, risk preferences and uncertainty environments in the target domain. An overarching synthesis of these paradigms is that hybrid strategies that can simultaneously leverage all the 3 sources of optimality by MCF-like policies are probably more balanced, though they bring complexity to computation, convergence and interpretability.

 doi:10.20944/preprints202601.2192.v1

Table 1. Comparative Analysis of Risk-Aware RL Paradigms.

Paradigm	Core Objective	Key Techniques	Strengths	Weaknesses	Ideal Application Domain
Risk-Neutral RL	Maximize expected return.	Q-Learning, Policy Gradients.	Simple, well-understood, efficient in low-risk settings.	Ignores variance and tail-risk; unsuitable for safety-critical tasks.	Game playing, recommendation systems.
Risk-Sensitive RL	Optimize a risk-adjusted metric (e.g., CVaR, variance).	Distributional RL, Exponential utility.	Directly controls for downside risk; quantifiable risk preferences.	Can be less stable; sensitive to parameter tuning; may underestimate model uncertainty.	Financial portfolio optimization.
Safe RL	Maximize reward while satisfying safety constraints.	Constrained MDPs (CMDPs), Lagrangian methods, Shielding.	Provides hard safety guarantees during learning and deployment.	Can be overly conservative; limits exploration; challenging constraint design.	Autonomous driving, medical dosing.
Robust RL	Perform well under worst-case model perturbations.	Robust MDPs, Adversarial training.	Resilient to model misspecification and adversarial conditions.	Often pessimistic; sacrifices average-case	

4. METHODOLOGIES AND ENABLING TECHNIQUES: CRITICAL REVIEW

The application of RL, when deployed in risk-sensitive domains, has called naturally for a wide increment of methods and enabling techniques, which extend standard RL models capturing risk-constraints, uncertainty, and safety requirements. These techniques cut across multiple layers of algorithmic development, including how risk is formulated inside decision procedures, which

 doi:10.20944/preprints202601.2192.v1

optimization frameworks are selected, how uncertainty is handled, and the techniques used for safe exploration. Such components need to be reviewed in a critical way, in order to evaluate their theoretical foundations, the possibility of practical application, and the fit to the needs of actual risk assessment systems.

Most RL formulations are based on the Markov Decision Process (MDP) formalism, which offers a mathematical framework for representing processes of decision-making under uncertainty. The classical MDPs are, however, risk-neutral; they are designed to maximize the expected sum of rewards coming back over a horizon. To deal with risk in the decision model, MDPs have been extended into risk-aware models such as Constrained Markov Decision Processes (CMDPs). In CMDPs the agent tries to maximize a primary objective, usually the expected sum of rewards, subject, however, to one or more cumulative cost functions that have to stay below given levels. This extension enables the representation of safety constraints and resource constraints, and CMDPs become the choice in safety-critical systems. Theoretically, CMDPs are non-convex and present the issues of optimal convergence, and feasibility. For example, it is not straightforward how to show that the set of feasible policies has a convex interface, when constraints or dynamics are non-linear. Second, it is not always decidable a priori whether there is an admissible policy that is optimal for both objectives; for example, in high-dimensional or non-stationary environments, it may not exist. These conceptual challenges manifest themselves in the tractability and scalability of risk-aware RL models that are rooted in CMDPs.

To solve risk-aware MDPs and risk-aware CMDPs, sophisticated optimization toolboxes and solvers for multi-objective optimization and constrained optimization are needed. One of the most popular approaches are policy gradient methods, which estimate gradients of expected return with respect to policy parameters and modify them accordingly. More advanced versions such as actor-critic methods integrate value and policy updates, which forms the trade-off between bias and variance of the gradient estimates. When there are constraints, primal–dual and Lagrangian approaches are commonly used. These strategies involve using dual variables (or Lagrange multipliers) to produce saddle-point formulations of implies in order that these can be solved by the use of gradient-based methods. Although these approaches offer a theoretically sound way to deal with constraints, they are inherently numerically unstable and sensitive to tuning, and often need plenty of iterations to converge in practice. Furthermore, the stochastic approximation used for the gradient estimation can have an impact on both convergence rates and robustness, particularly in the case of problems with sparse rewards or delayed feedback. These solvers, however, have theoretical beauty but could face a problem concerning verifiable performance guarantees in real time or safety-critical situations, where failure to converge has real consequences.

In addition to maximizing expected value, recent work in RL has been focusing on distributional and Bayesian methods to more accurately model the uncertainty inherent in sequential decision-making. This differs from the classical settings by estimating the entire probability distribution over returns instead of just one expected return. This more informative representation allows the agent to measure risk metrics, e.g., variance, quantiles and tail risk (e.g., CVaR) directly on the learned return distribution. Algorithms such as C51, Quantile Regression DQN, and Implicit Quantile Networks are examples of this shift, providing new opportunities for incorporating risk preferences into learning. In contrast, Bayesian RL can account for epistemic uncertainty by tracking a posterior distribution over the model's parameters or value functions such that the agent can reason in terms of uncertainty in its predictions. Bayesian techniques are especially powerful in data-poor or partially observed problems where uncertainties regarding dynamics or rewards can easily have a large impact on the robustness of policies. Nonetheless, both distributional and Bayesian approaches are not scalable to a large number of wiki concept pairs. Distributional approaches will only add more storage and computation to preserve complete distributions, and the resulting Bayesian inference is usually infeasible in high dimensions even if reduced to an approximation that loses fidelity.

One of the primary operational challenges in risk-sensitive RL is that of safe exploration, i.e., an agent’s ability to effectively learn while remaining within safety bounds during the learning process.

 doi:10.20944/preprints202601.2192.v1

Classical exploration methods, like ε-greedy or Boltzmann exploration, promotes random or stochastic action selection, in order to explore the state-action space. But in safety-critical settings, such experimentation may result in disastrous failures. In response to this, a number of solutions have been proposed in the literature aiming at providing safety during the operation, such as reactive safety guarantees (i.e., safety affordance) or shielding, where responses to unsafe actions are provided on line, through the use of methods based on control theory, e.g., Lyapunov-based methods, which enforce stability constraints. Other approaches are predictive safety filters that simulate the consequences of candidate actions and discard those that result in constraint violations. Although these methods can ensure of safety in training time, they are also prone to some trade-offs between learning efficiency and exploration efficiency. Safety during learning naturally results in a trade-off between the agent's ability to explore potentially high-reward but initially uncertain actions, possibly impeding convergence or leading to suboptimal policies.

From a critical perspective, the theory promise of these methods does not seem to match their real-world applications. Most algorithms make assumptions, based on perfect observability, stationarity, or the knowledge of the constraint functions, that are not satisfied in practice. In addition, safety claims established using formal methods often do not cover the complete spread of the environmental variability witnessed in operational scenarios. For instance, while the policy convergence in CMDPs can be mathematically proven under certain conditions, this theoretical guarantee can be broken in practice by noise, approximation errors, and computational limitation. On the other hand, in the case of robust optimization, worst-case performance can only be guaranteed and realized if the true environment is in the modeled uncertainty set, and if it is, strictly better performance can be achieved. Such shortcomings indicate a gap between formality correctness and practical reliability for risk-aware RL, and challenges the transportability and generalization of risk-aware RL algorithms in more realistic domains.

The enabling techniques of risk-aware reinforcement learning are of impressive theoretical depth and algorithmic novelty. Methods such as constrained optimization, distributional modelling, and safe exploration, offer powerful resources to deal with the intricate problems of the risk-sensitive aspect in sequential decision making. Nevertheless, their successful use in practical applications is currently limited by issues of scalability, model assumptions, and computational feasibility. It is an open challenge to attempt to bridge this gap between theory and empirical reliability, which calls for the proposal of more practical and robust algorithms, benchmarking sets to evaluate the performance in a standardised way, and the incorporation, to an extent not yet achieved, of domain knowledge into action-learning systems in order to obtain robust, interpretable and safe learning under uncertainty.

5. Application Domains: Cross-Sectoral Comparison

Reinforcement Learning (RL) shows a remarkable diversity in terms of the realization and performance as well as the applicability when being utilized in risk assessment. Each domain features distinct attributes regarding environmental uncertainty, operational limitations, data availability, and safety mandates, shaping how RL methods can be chosen, configured, and executed. Across these sectors, looking at finance, energy and infrastructure, autonomous systems and cyber security we see the strengths and weaknesses of RL in tackling risk within real-world domains. This section investigates how risk-aware RL is used in these areas (i.e., how characters that are maximin safe are already considered), what kind of methodological changes need to be made to account for the domain specifics, and to which extend methods for one domain can be directly applied to the other.

In finance, the role of risk management functions has long been part of decision support systems. RL is commonly used for portfolio optimization, asset allocation and trading strategy under market uncertainty in this field. As part of model-based and model-free learning techniques in RL, risk-aware RL techniques, namely those developed based on risk-sensitive formulations, can be utilized to capture/learn and optimize risk-return trade-offs. In these formulations, they employ

 doi:10.20944/preprints202601.2192.v1

variance, VaR, and CVaR as risk measures to measure and manage downside risks. Distributional RL and Bayesian inference through the modelling of market volatility have also appeared in agents that are able to account for more complicated stochastic behaviour at the markets beyond estimating mean returns. However, these learning dynamics do not have a good performance in the non-stationary financial environment, featuring high-dimensionality with the existence of adversarial participators. Second, practical constraints such as regulations and the need for transparency and explainability further prevent people to deploy black-box models in real-life applications. Importantly, despite the fact that finance has been a promising domain for risk-sensitive RL applications, deployed real-world applications are generally restricted to environment heavy simulated environments (i.e. back testing) and are not full-fledged autonomous systems trading live in markets.

In energy systems and infrastructure, RL methods are becoming popular to assist decision making in safety-critical operations like power grid control, demand-response management, and resource allocation in smart grids. These systems are usually subject to strict safety and reliability requirements, and such systems, if incorrectly controlled, may cause severe physical or economical degradation. The unpredictability of the demand, renewable generation, and failures of devices call for both safe and robust RL. For example, CMDPs, restrictive MDP formulations, have been used for stability as well as performance to maintain while robust RL formulations can be used to deal with partial observability or unexpected environmental disturbances. Moreover, the energy industry takes advantage of model-based policy learning, multiplying the power of going hybrid and integrating domain-specific models with learning policies, which at this point turn hybrid, i.e., mix physics-based simulations with RL. But these hybrid techniques generally need significant computational cost and are not easily scalable for wide-area networking. In addition, the time scale and real-time requirements of control applications also limit the employment of algorithms that converge slowly or have high computational requirements, which affects the selection of RL methods as well.

The application of RL for autonomous systems (self-driving cars, quadcopters, robotic arms, etc.) presents challenges in real-time decision making, controlled navigation, and model disturbances. Such systems function in a dynamic, uncertain, and frequently unstructured world, in which safety and robustness are pivotal. RL has been used to learn control policies for motion planning, collision avoidance, and adaptive behaviour in response to changes in the environment. Given the high exploration failure rate, learning not to fail in exploration, e.g., with safe RL strategies such as action shielding, reward shaping, and Lyapunov-based methods, is crucial to guarantee constraint satisfaction at both training and test time. Furthermore, mismatch in the model between simulation and the real world (sim-to-real gap) presents a great risk to policy transfer. In order to counteract this problem, strong RL and domain adaptation strategies are used to reduce the gap and retain the same performance after deployment. However, the adoption of RL in embodied autonomous agents is hampered by data efficiency, safety validation, and interpretability, which are necessary conditions for regulatory approval and public acceptance.

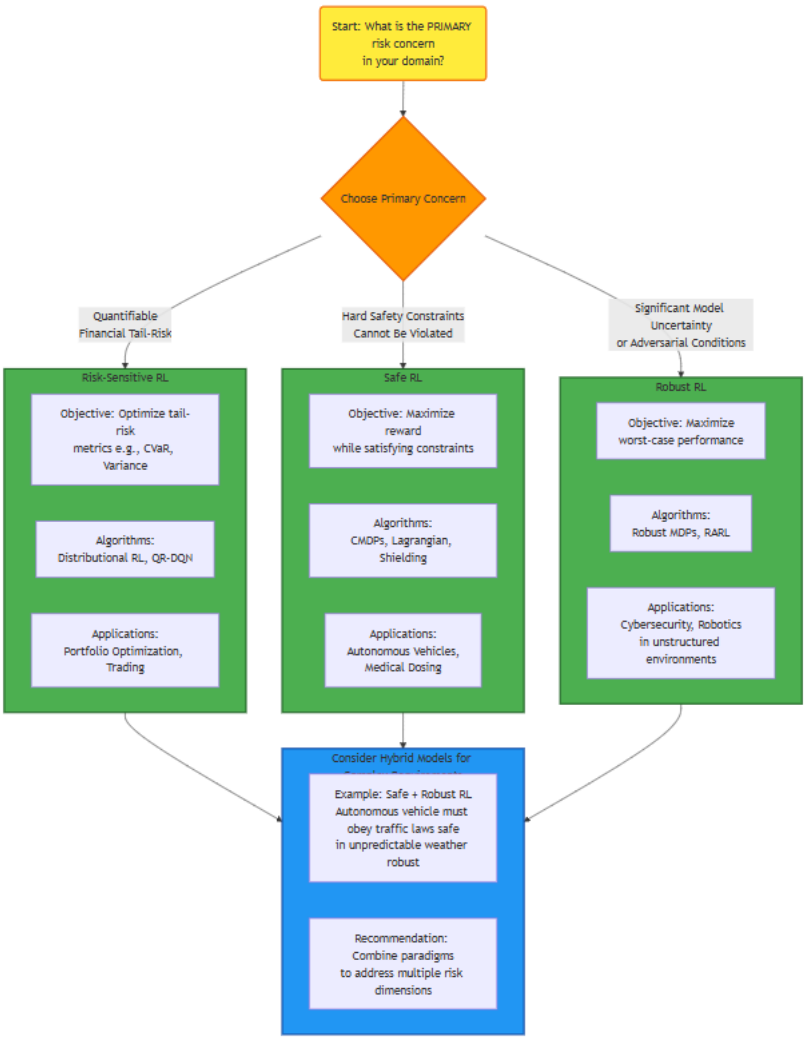
In the field of cybersecurity, while reinforcement learning applied to cybersecurity is used to deal with dynamic threats situations in a complex and adversarial environment. Applications range from intrusion detection to dynamic response to cyberattacks, and learning of network defence strategies. In particular, cybersecurity atmospheres are featured with partial observability, adversarial intervention and dynamic change of attack strategies, which render the situation highly non-stationary and difficult to formulate. In such environments, strong RL and adversarial training techniques are particularly noteworthy, because it enables the agent to be armed against and adapt to the worst-case situations. Besides, distributional RL can be employed to measure risk levels related to potential attack paths. The core problem to address in this perspective is the low and slow feedback at learning time (i.e., we may not observe at the moment if an intrusion succeeded or failed, nor be able to attribute it to specific actions). Furthermore, due to the high-stakes nature of cybersecurity, our policies need to be robust and auditable, making deployment of black-box or

 doi:10.20944/preprints202601.2192.v1

heavily-parameterized RL models challenging. Despite these difficulties, due to the ability of RL to adapt to ever-changing threats, it seems to be a promising tool for defensive strategies in the long-term, assuming that the models are sufficiently explainable and robust.

A cross-domain critical review of relevant works makes clear the way in which domain-specific constraints may shape the choice and the adaptation of RL techniques. Regulatory and ethical limitations in finance and cybersecurity highlight the necessity of model interpretability and transparency. In energy and autonomous systems, real time operation and safe operation are paramount and they call for the use of constrained and robust RL methodologies. The degree to which methods developed in a particular context can be exported to another will depend on the degree to which the risk structures, observability conditions and systemic dynamics align. Such as strong RL methods developed for robotics that can be used for cyber-scenarios as well, in terms of adversarial resilience. In comparison financial domain-specific methods often need significant re-engineering to be used in infrastructure / control systems because of a different risk formalisation and requirements of real-time behaviour.

The application of risk-aware RL in diverse domains illustrates its flexibility and limitations. Although the fundamental mathematical approach is universal, the practical application of this model is highly sensitive to the nature of risk relationship, system latency, data feedback and the regulatory constraints. The knowledge of these contexts is critical for properly translating theoretical progress in RL into robust and practical risk assessment solutions. In addition, this comparative analysis also highlights the necessity for more standard studies and transferability researches, which could close methodological gaps between fields and promote cross-domain development of risk-aware reinforcement learning.



 doi:10.20944/preprints202601.2192.v1

Figure 2. The decision framework for practitioners.

6. Evaluation Metrics and Benchmarking Standards

A comprehensive evaluation of RL algorithms in risk-sensitive settings would have to define and use metrics beyond performance that also capture risk related properties like safety, robustness, and uncertainty. Compared with traditional RL context, in which the performance objective is commonly to maximise the expectation of accumulative reward, in risk-aware RL, evaluation has additional dimensions related to the standard spirit of evaluation. In this section, we provide a critique of existing practices for evaluating risk-aware RL, indicate their deficiencies in providing a fair comparison between risk-aware algorithms and benchmark domain spaces for risk and describe standard metrics and evaluation frameworks adapted to risk assessment that are necessary.

Current evaluation methods in risk-sensitive RL are typically based on measures that have been defined or used in financial engineering and operations research, where the modelling of risk is more developed. Among the approaches used, the Conditional Value-at-Risk (CVaR) is popular, which measures the expected loss in the poorer tail of the outcomes and allows agents to be penalized for bad tail events. This makes the CVaR-based assessments especially important in the settings in which having extreme (negative) costs, though very unlikely, has a large impact (e.g., finance, energy, or autonomous systems). Another common measure is the variance of the returns, which indicates how the outcomes spread around and, therefore, how stable the learned policy is. Nevertheless, such a measure of variation might not provide complete description of the extreme value behaviour, such as asymmetric and heavy-tailed risk distributions, and might need additional measures. Moreover, it is common to evaluate an agent's capability under assumed uncertain or adversarial scenarios using robustness metrics such as worst-case return under environment perturbations, adversarial success rates, or performance under an adversarial or uncertainty set. Interest in these metrics is heightened in robust RL, where algorithms seek to perform well across a class of plausible models rather than optimizing over nominal counterparts.

However, much remains to be addressed for the evaluation landscape of today. The first issue is that there are no general, domain-independent benchmarks which can be used by all risk-aware algorithms. The majority of the existing benchmarks in RL, such as introduced by OpenAI Gym or DeepMind Control Suite, emphasize the optimization of the expected reward and do not naturally support risk-sensitive objectives or constraints related to system safety. While attempts have been made, for example by OpenAI (Safety Gym) to incorporate constraint-based evaluation in RL environments, such adoption is limited and the risk formulations considered are narrow. Furthermore, experiments tend to rely on custom-built benchmarks and ad hoc risk metrics, which hinders the reproducibility and comparability of findings across studies. This situation is worsened by the absence of any standard reporting procedures, where important assessment details, such as how constraints are defined, what constitutes an acceptable risk, or what levels of uncertainty are allowed, are not reported and therefore unclear in any reporting of results, rendering findings ambiguous or non-repeatable.

Another important drawback is the low external validity of many evaluation methods. Simulation environments, although beneficial at development and testing, seldom achieve the full complexity, variation, stochastic of operational domains, including power networks, medical systems, or financial markets. Consequently, while a model may perform extremely well in a simplified simulation, it may fail in real-world conditions where the data is noisy, the feedback is delayed, or the system dynamics can change unexpectedly. Also, most risk-aware RL models are tested on synthetic distributions that might not correspond to real-world risk structures, rendering the assumption of e.g. non-Gaussian returns or safety violations taking a correlated shape infeasible. These problems create a gap between theoretical risk modelling and real-world deployment, which is a critical obstacle to the wide spread use of RL in safety critical or regulation sensitive domains.

Considering these limitations, it is desirable to make a step towards standardised evaluation dimensions tailored to risk-aware RL. One such direction can be to formalize the risk profiles of the

 doi:10.20944/preprints202601.2192.v1

agent, that captures the agent's risk attitude along different dimensions of risk, such as expected return, CVaR, variance, constraint satisfaction rate etc. Such profiles allow to not just compare the algorithms by a single scalar metric, but '[they allow to compare] the overall performance of the algorithm under certain risk preferences or in certain operating regimes. Another critical area is monitoring of safety and violations during training and deployments. Safety violations can compute as frequency, cumulative penalization, or risk-weighted impact, and two ways to compare safe and non-safe learning strategies in nuanced manner. Furthermore, the consideration of trade-off curves, such as reward vs. CVaR or reward vs. constraint violation rate, may facilitate understanding and quantifying the trade-off between performance and risk mitigation, which plays a crucial role in real-world decisions.

An equally necessary complement to such assessments is the development of domain-specific as well as cross-domain benchmark suites, that consist of environments with clearly formulated risk structures, realistic boundaries and modelling of uncertainty. This reference system needs to have standardized populations, standardized risk factor models, and explicit descriptions of its evaluation strategy. Moreover, the evaluation should not only include stress testing procedures that evaluate the agent's sensitivity to out-of-distribution, delayed reward or model-misspecification scenarios to check the robustness of the solutions beyond nominal settings. The design of such benchmarks needs to be grounded on domain requirements – such as real-time performance, regulatory concerns, or data sparsity – to make sure the assessments are not only sound but also in line with the limitations of real-world use.

There are several different risk-sensitive metrics that have been proposed in the literature and do get used, but the lack of consensus and generalizability is a major obstacle to assessing risk-aware RL. Existing methods often exhibit a fragmented and environment-specialized nature and focus on simulation, which hinders systematic comparison and real world generalization. The solution of such issues calls for collaborative design of global, multi-dimensional and transparent metrics for evaluations where both theoretical and application requirements on one's level of understanding are captured. Such frameworks are necessary if the field of risk-aware RL is to move from experimental algorithms to dependable building blocks for safety-critical systems.

7. Challenges and Open Issues

Risk-sensitive decision-making using reinforcement learning (RL) has recently given rise to a variety of algorithmic developments, but several challenges hinder its theoretical development, widespread application, and ethical adoption. The challenges fall along three main dimensions: theoretical limitations, practical challenges, and ethical and policy considerations. In particular, each dimension brings specific gaps which influence the scalability, robustness and acceptability of RL systems for risk assessment. A critical and comparative assessment of these questions is necessary in order to understand where are the bottlenecks and to orientate future research.

From a theoretical perspective, one of the major open issues in the area of risk-aware RL is the absence of strong convergence guarantees, particularly under constraints and multi- objective formulations. RL theory has been well-studied for the tabular MDPs with finite state-action spaces under stationary polices where the theory has been established with the properties of strong convergence However, these convergent results do not directly apply to Constrained Markov Decision Processes (CMDPs) or Multi-objective Reinforcement Learning (MORL) problems. In CMDPs, agents have to optimize concurrently a primary reward signal and satisfy one or multiple constraints that could be at odds with each other or be even challenging to be accurately modelled. This adds extra complication to the learning procedure, in particular when the set of feasible policies is non-convex or when the constraints are stochastic and time-varying. In addition, MORL entails balancing of multiple objectives (e.g., reward-maximizing, risk-minimizing and safety) which can possibly be incommensurable and/or their exact observability may be only partial. Pareto-optimal solutions in these scenarios are typically non-unique and challenging to find computationally, and there are no generalization guarantees of existing methods with respect to different risk levels.

 doi:10.20944/preprints202601.2192.v1

Accordingly, existing algorithms often resort to approximations or heuristic compromises without formal derive, which severely undermines the robustness and interpretability of these algorithms in high-stakes applications.

On the practical side, there exist several very important obstacles for a successful deployment of risk-aware RL in real systems. A key concern is the availability of data, notably for tasks where interacting with the environment comes at a cost/risk or in a constrained environment due to regulatory constraints. Unlike in synthetic benchmarks, where there is plentiful and free data, in deployed settings, such as healthcare, autonomous driving, or industrial control, gathering a large number of trajectories to train RL agents either safely or efficiently can be impractical. Then the requirement of sample efficiency in such settings is much higher, and the majority of existing RL algorithms cannot return the sample efficiency for deployment. Another important difficulty comes from the application of RL in mission-critical systems, in which the safety, reliability and transparency are necessary. Such systems are often characterised by stringent real-time requirements and thus call for decision-making to be based on deterministic and interpretable decision processes. However, a number of RL models are stochastic, high-dimensional, and do not come with formal guarantees about their behavior under unobserved circumstances. Furthermore, in real-life settings, systems are often composed of legacy systems and different types of data sources, which pose a technical challenge for integration and die risk of system-level failure. The lack of mature tools for debugging, testing, and validating RL policies in dynamic and uncertain domains limits their application to well controlled research prototypes.

In addition to algorithmic and system-level issues, there are ethical and policy considerations increasingly influencing whether RL-based risk assessment systems are deemed acceptable. When RL agents are used to make or aid decisions with societal impact, the need for responsibility, fairness, and transparency is in demand. In high-stakes settings like financial trading or autonomous systems, we need to be able to audit this decision making, and assign responsibility in the case of a bad outcome. But the adaptability and opacity of a lot of RL algorithms makes it hard. In addition, equity in decision-making is oftentimes not directly specified in traditional RL objectives, and if not well-guarded against, it is possible for RL agents to capitalize on the statistical biases in the training data, aggravating biases in the data or even breaking the law. And the fact that most deep RL models are not interpretable at all means that stakeholders cannot take comfort in understanding how these algorithms work, and in fields such as healthcare and criminal justice, we likely need (external) human oversight to trust these models to be defensible. Existing methods for explainability, such as saliency maps, or local surrogate models, are not necessarily applicable to the sequential and stochastic nature of RL policies, preventing a closer link between technical performance and societal acceptability.

To summarize these observations, a comparison between problems can be categorized in 3 dimensions: the technical, the practical, and the ethical. Technical challenges, like convergence under constraints and absence of generalization in multi-objective settings, do not limit a model's theoretical valuation. Practical obstacles such as lack of data, computational cost, and problems of integration with the system, seem to restrict their feasibility in deployment scenarios. Ethical considerations, above all responsibility, justice or transparency, affect the societal and regulatory acceptance of RL systems. These dimensions are not orthogonal; for example interpretability (a moral issue) may worsen deployment risks (a pragmatic issue), and a lack of theoretical guarantees (a technical question) may prevent certifications of compliance by regulation (a moral-practical interface).

Despite the great potential of reinforcement learning (RL) for risk-sensitive applications, its practical use is greatly limited by open theoretical, operational, and ethical problems that have yet to be tackled. Solutions to these challenges will likely require a multi-dimensional approach that involves algorithm design, systems engineering, and policy making. A more comprehensive connection between risk-aware RL research and fields like control theory, human-computer

 doi:10.20944/preprints202601.2192.v1

interaction, and regulatory science may be required in order to address these gaps and help to ensure the reliable and responsible use of RL for critical risk analysis tasks.

8. Synthesis and Future Research Opportunities

The preceding discussion emphasized how various kinds of reinforcement learning (RL) paradigms, the methodological innovations behind them, and the challenges limiting their applicability have been identified. However, there is a growing realization that cross-cutting proposals that can combine the best of the existing paradigms and bring coherence to the trade-offs between safety, robustness, adaptability, and performance is needed. This concluding section reviews the emerging patterns and calling for prospective studies organized around four main themes: cross-paradigm convergence, conceptual unification, interdisciplinary integration, methodological development. These directions are each motivated by limitations of existing work on risk-aware RL and strategies for moving toward more integrated and deployable solutions.

One of the most direct opportunities of integrating is between safe RL and robust RL - two paradigms that have proceeded largely in isolation from one another. Safe RL concerns constraint satisfaction under both learning and execution, while the focus of robust RL is on policy performance in the presence of model uncertainty and adversarial perturbations. Indeed, in many practical applications including those as diverse as autonomous navigation, financial control, or energy management, one wants to guarantee both properties at the same time: agents must not end up in states that are widely considered unsafe while achieving acceptable performance over a broad range of believable environmental dynamics. Constructing constraint-adherent yet uncertainty-robust algorithms would require a hybrid approach that mixes the CMDPs, adversarial training, dual control and the like. For instance, generalizing solid optimization approaches to accommodate safety constraints poses new mathematical problems related to feasibility and conservatism. Likewise, the safe RL algorithms need to be tailored to uncertainty sets or non-stationary environments. These efforts need new types of algorithms that can handle interactions between multiple risk sources , structural, operational, and epistemic , that do not sacrifice the efficiency and scalability needed to actually process increasingly massive amounts of data.

Another important direction of research is to better understand to what extent the experience-based algorithms can exploit limited or adversarial feedback, which is a central challenge for RL when tackling risk-sensitive tasks. Many of the most effective present day algorithms assume dense and safe interaction with the environment, assumptions that fail when exploration is costly or dangerous. Sample efficiency is important and the exploration behavior is required to be steered tightly to prevent domain violations or catastrophic failure. Exploration via offline RL, batch-constrained learning, and simulation-to-reality transfer is promising in this sense. Such approaches allow for policy learning using historical or synthetic data, with less dependence on real-world trial-and-error. Also, meta-learning and few-shot adaptation would facilitate transfer of knowledge over similar tasks, decreasing the dependence on unsafe exploration. Furthermore, integrating uncertainty-aware exploration policies, namely Bayesian or ensemble-based approaches, also helps in handling risk in learning. Progressing along this line of inquiry will involve a number of formalisms to assess and bound the quality of learning under the feedback constraint, such as confidence calibration, policy regularization, and performance certificates under partial information.

In order to structure and consolidate the variety of techniques presented in the review, a conceptual unification of risk-aware reinforcement learning in a unified theoretical framework is urgently required. Existing paradigms are typically studied in isolation, with disparate assumptions pertaining to risk modelling, optimisation objectives, and learning dynamics. One promising avenue is in formalizing a “risk-adjusted RL” framework, that does not decouple risk as external constraint or secondary objective but rather incorporate it in the core policy evaluation and update mechanisms. This unified framework would extend classical MDPs and CMDPs by enabling flexible risk measures (from variance-based measures to distributional constraints and adversarial dynamics). And, it would allow for multi-criteria optimization, where agents can trade-off between reward

 doi:10.20944/preprints202601.2192.v1

maximization and varying types of risk based on task-level constraints, or user preferences. From a methodological standpoint this means to introduce new policy gradient expressions, value functions, and Bellman operators, where risk measures are involved explicitly. An integrated conceptualization would enable cross-algorithm comparisons, ease of implementation for hybrid models, and an increased level of transparency in the empirical literature and benchmarking.

In addition, advances in risk-aware RL rely on inter-disciplinary contributions, with, in particular, neighbouring disciplines providing mature tools for the modelling of risk, constrained optimization, as well as uncertainty handling. Control theory is a well-grounded scaffold for the design of robust and provably safe decision-making systems, as supported by tools there such as Lyapunov stability analysis, barrier functions, and reachability analysis among others. Such mechanisms can be incorporated into RL to make the learned policies more reliable and interpretable. Economics and decision theory provide formal models of utility, preference under risk, and bounded rationality which might guide the design of richer reward and risk functions. Operations research provides mature optimization methods for multi-objective and constrained planning, and behavioural science offers models for a human's preferences, trust, and risk perception, which become more important in human-in-the-loop RL systems. Cross-disciplinary collaboration in these areas can result in new formulations, hybrid methods, and rigorous evaluation criteria that can adequately meet the needs of risk assessment in complex socio-technical systems.

From these strategic directions, a systematic research agenda can be developed. A first, very important future direction is that of creating benchmarks for risk-aware RL that are standardized. These benchmarks should also have a wide variety of different environments with distinct risk structures, constraint sets, and forms of uncertainty, as well as well-established evaluation protocols and metrics to ensure cross-comparison. Second, theoretical tool development is required to obtain convergence, policy optimality and performance guarantees under risk-sensitive objectives. Such tools will need to handle limited, non-convex, multi-objective formulations and include realistic model errors and data constraints. Third, focus on application-specific customization and adaptability where the algorithms are customised to the structural and operational characteristics of their applications should be the emphasis. This involves defining domain-specific safety constraints, incorporating expert knowledge, as well as evaluating policies through simulation and formal verification techniques. Lastly, open-source platform-agnostic and modular frameworks can aid in reproducibility and cross-pollination of collaboration between different communities focused on risk-aware decision making.

The next generation of reinforcement learning for risk assessment is that it becomes a coherent, flexible and reliable frame that supports critical decision-making in many fields. Theoretical unification, cross-paradigm synthesis, and interdisciplinary cooperation will be necessary in order to overcome present fragmentation and to achieve this potential. As risk-sensitive RL increasingly matures, it must not only synchronize with algorithmic formalism but with the structural, operational, and ethical constraints of the environments where these learning agents are to operate.

9. Conclusion

This review identifies a field that is both very promising and still in its infancy. Reinforcement Learning provides a paradigm change from static, retrospective risk assessment to sequential, adaptive, and dynamic decision-making under uncertainty, as the analysis shows. There are significant opportunities in the fields of finance, healthcare, and autonomous systems because to RL agents' capacity to learn through interaction, optimize for long-term results, and continually improve from streaming data. Nonetheless, the discipline is fragmented, with domain-specific solutions devoid of a common theoretical foundation. There are still significant gaps in establishing interpretability, guaranteeing safety, and closing the sim-to-real transfer gap, all of which make it more difficult to get from research prototypes to reliable, practical implementation.

The comparative analysis indicates that no single RL paradigm is universally superior; rather, the choice of approach is highly context-dependent. Risk-sensitive RL holds the most promise for

 doi:10.20944/preprints202601.2192.v1

domains with well-defined, quantifiable risk metrics, such as financial portfolio optimization where tail risks (e.g., CVaR) can be explicitly optimized. Safe RL, particularly Constrained MDPs, is indispensable for safety-critical applications like autonomous driving and medical treatment planning, where hard constraints must not be violated. Robust RL is best suited for environments with significant model uncertainty or adversarial conditions, such as cybersecurity. The most promising future direction lies not in a single approach, but in hybrid models that can integrate the risk-awareness of the first, the safety guarantees of the second, and the resilience of the third.

The transformative potential of RL in risk assessment is undeniable, yet its path forward demands a concerted, interdisciplinary effort. We issue a call to action toward the development of responsible, risk-aware AI. This requires moving beyond pure performance metrics to create algorithms that are not only powerful but also safe, interpretable, and aligned with human values. Future research must prioritize the creation of standardized benchmarks, the establishment of rigorous theoretical guarantees for safety and convergence, and the fostering of collaboration between RL researchers, domain experts, and ethicists. Only by confronting these methodological and practical challenges head-on can we unlock the full potential of Reinforcement Learning to build more intelligent, resilient, and trustworthy risk assessment systems for the complex challenges of the future.

Author Contributions: Conceptualization, C.A. and H.H.; methodology, H.H.; software, C.A. and H.H.; validation, C.A. and H.H; formal analysis, C.A. and H.H; investigation, C.A.; resources, C.A. and H.H.; data curation, C.A.; writing—original draft preparation, C.A. and H.H.; writing—review and editing, H.H.; visualization, C.A. and H.H.; supervision, H.H.; project administration, C.A. and H.H.; funding acquisition, C.A. and H.H..

Funding: Please add: This research received no external funding

Acknowledgments: The authors extend their sincere gratitude to everyone that contributed to the success of this research

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CMDP	Constrained Markov Decision Process
CVaR	Conditional Value-at-Risk
DRM	Distortion Risk Measure
MDP	Markov Decision Process
MORL	Multi-Objective Reinforcement Learning
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RL	Reinforcement Learning
VaR	Value-at-Risk

References

1. Hans, A., Schneegaß, D., Schäfer, A. M., & Udluft, S. (2008, April). Safe exploration for reinforcement learning. In *ESANN* (pp. 143-148).
2. Jaimungal, S., Pesenti, S. M., Wang, Y. S., & Tatsat, H. (2022). Robust risk-aware reinforcement learning. *SIAM Journal on Financial Mathematics*, *13*(1), 213-226.
3. Chow, Y., Nachum, O., Faust, A., Duenez-Guzman, E., & Ghavamzadeh, M. (2019). Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*.

 doi:10.20944/preprints202601.2192.v1

4. Wang, Y., & Zou, S. (2021). Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, *34*, 7193-7206.
5. Eldeeb, E., Sifaou, H., Simeone, O., Shehab, M., & Alves, H. (2024). Conservative and Risk-Aware Offline Multi-Agent Reinforcement Learning. *IEEE Transactions on Cognitive Communications and Networking*.
6. Moldovan, T. M., Abbeel, P., Jordan, M., & Borrelli, F. (2014). *Safety, risk awareness and exploration in reinforcement learning* (Doctoral dissertation, University of California, Berkeley).
7. Laroche, R., Trichelair, P., & Des Combes, R. T. (2019, May). Safe policy improvement with baseline bootstrapping. In *International conference on machine learning* (pp. 3652-3661). PMLR.
8. Dabney, W., Rowland, M., Bellemare, M., & Munos, R. (2018, April). Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).
9. Shen, Y., Tobia, M. J., Sommer, T., & Obermayer, K. (2014). Risk-sensitive reinforcement learning. *Neural computation*, *26*(7), 1298-1328.
10. [Brown, D. S., Cui, Y., & Niekum, S. (2018, October). Risk-aware active inverse reinforcement learning. In *Conference on Robot Learning* (pp. 362-372). PMLR.
11.] Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
12. Garcıa, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, *16*(1), 1437-1480.
13. Moos, J., Hansel, K., Abdulsamad, H., Stark, S., Clever, D., & Peters, J. (2022). Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, *4*(1), 276-315.
14. Wang, D., Li, L., Wei, W., Yu, Q., Hao, J., & Liang, J. (2025, April). Improving Generalization in Offline Reinforcement Learning via Latent Distribution Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 20, pp. 21053-21061).
15. Wang, D., Li, L., Wei, W., Yu, Q., Hao, J., & Liang, J. (2025, April). Improving Generalization in Offline Reinforcement Learning via Latent Distribution Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 20, pp. 21053-21061).
16. Yang, Z., Jin, H., Tang, Y., & Fan, G. (2024, May). Risk-Aware Constrained Reinforcement Learning with Non-Stationary Policies. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (pp. 2029-2037).
17. Mankowitz, D. J., Levine, N., Jeong, R., Shi, Y., Kay, J., Abdolmaleki, A., ... & Riedmiller, M. (2019). Robust reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:1906.07516*.
18. Clements, W. R., Van Delft, B., Robaglia, B. M., Slaoui, R. B., & Toth, S. (2019). Estimating risk and uncertainty in deep reinforcement learning. *arXiv preprint arXiv:1905.09638*.
19. Borkar, V. S. (2010, July). Learning algorithms for risk-sensitive control. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS* (Vol. 5, No. 9).
20. Villalobos-Arias, L., Martin, D., Krishnan, A., Gagné, M., Potts, C. M., & Jhala, A. (2023). Modeling Risk in Reinforcement Learning: A Literature Mapping. *arXiv preprint arXiv:2312.05231*.
21. Rowland, M., Dadashi, R., Kumar, S., Munos, R., Bellemare, M. G., & Dabney, W. (2019, May). Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning* (pp. 5528-5536). PMLR.
22. Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2015). Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, *8*(5-6), 359-483.
23. Pinto, L., Davidson, J., Sukthankar, R., & Gupta, A. (2017, July). Robust adversarial reinforcement learning. In *International conference on machine learning* (pp. 2817-2826). PMLR.
24. Ni, X., & Lai, L. (2024, November). Robust Risk-Sensitive Reinforcement Learning with Conditional Value-at-Risk. In *2024 IEEE Information Theory Workshop (ITW)* (pp. 520-525). IEEE.
25. Garcıa, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, *16*(1), 1437-1480.
26. Ni, X., & Lai, L. (2024, November). Robust Risk-Sensitive Reinforcement Learning with Conditional Value-at-Risk. In *2024 IEEE Information Theory Workshop (ITW)* (pp. 520-525). IEEE.

 doi:10.20944/preprints202601.2192.v1

27. Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017, July). Constrained policy optimization. In *International conference on machine learning* (pp. 22-31). PMLR.
28. Gu, S., Sel, B., Ding, Y., Wang, L., Lin, Q., Knoll, A., & Jin, M. (2025). Safe and balanced: A framework for constrained multi-objective reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
29. Moos, J., Hansel, K., Abdulsamad, H., Stark, S., Clever, D., & Peters, J. (2022). Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, *4*(1), 276-315.
30. Allouch, A., Koubaa, A., Khalgui, M., & Abbes, T. (2019). Qualitative and quantitative risk analysis and safety assessment of unmanned aerial vehicles missions over the internet. *Ieee Access*, *7*, 53392-53410.