# Enhancing Safety in Reinforcement Learning Training through Transformers Filtering

Mario Fiorino

Doctorate in Artificial Intelligence, 38th cycle (2022-2025), Politecnico di Torino

**Abstract** Artificial Intelligence (AI) refers to the field of study and the collection of techniques aimed at developing machine systems capable of reproducing advanced cognitive processes such as learning, reasoning, and decision-making. Over recent decades, AI has evolved rapidly, emerging as a transformative discipline with profound impacts on science, industry, and society. Among its paradigms, Reinforcement Learning (RL) has played a fundamental role in advancing AI capabilities, enabling systems to learn optimal behaviors through interactions with their environments. RL has been successfully applied across diverse domains, including robotics, healthcare, resource management, and large language model tuning, where it has, in several instances, demonstrated superhuman performance. More specifically, RL focuses on designing algorithms that enable an agent (an autonomous software entity) to make sequential decisions aimed at achieving optimal performance. Through iterative interactions with its environment, the agent seeks to maximize a feedback signal, commonly referred to as the reward, which evaluates the effectiveness of its actions.

The interaction between the agent and the environment—particularly in model-free approaches, where systems lack an explicit prior model of the environment and acquire knowledge solely through trial-and-error exploration—has highlighted both the strengths and limitations of the RL paradigm. Strengths include its flexibility in learning complex behavior without requiring explicit models. Significant limitations, including sample inefficiency and safety risks, arise from the uncontrolled exploration inherent to model-free learning. In particular, in safety-critical domain, such exploration can lead to severe or irreversible consequences due to the absence of predictive safeguards. For instance, trial-and-error exploration in an autonomous driving system might entail executing maneuvers that risk collisions and serious accidents. Similarly, in an automated medical diagnostic system, such exploration could involve suggesting treatments that potentially result in adverse patient outcomes. Traditional model-free RL algorithms are not suited to handle such high-stakes scenarios, as they lack mechanisms to guarantee safe behavior during learning.

In response to these challenges, the field of Safe Reinforcement Learning (Safe RL) has emerged, aiming to incorporate explicit safety considerations into the learning process, seeking to optimize performance rewards while ensuring that the agent operates within predefined safety constraints. Since its inception, several approaches have been proposed, notably Constraint-Based Methods and Shielding Methods. Although these methodologies have shown promising results, they exhibit several limitations. Both rely heavily on the explicit formulation of safety constraints, which are frequently partially unknown, highly complex, or impractical to define, interpret, and implement. Moreover, verifying the correctness and completeness of these constraints often poses additional challenges. Providing additional insight, in the context of model-free RL, Constraint-Based Methods face a fundamental limitation: during training, the agent inevitably violates constraints due to the inherent trial-and-error mechanism, thereby restricting their applicability primarily to simulated environments. Shielding Methods, while offering mechanisms to prevent unsafe actions, require an a priori model of the environment to be effective, rendering them unsuitable for model-free scenarios.

This dissertation addresses these limitations by introducing a novel framework, Safe Reinforcement Learning via Transformer Filtering, which leverages the representational power of Transformer neural networks to implicitly learn safety constraints directly from data. Once trained on sequences of expert safe trajectories, the Safety Transformer acts as a dynamic filter over the agent's action space, constraining exploration to safety-consistent behaviors. Fundamentally, this framework enables safe online learning in model-free RL settings without requiring the explicit modeling of constraints.

Empirical validation was conducted across two distinct domains: a deterministic navigation task and a stochastic biomedical control problem. The results demonstrate that the proposed approach effectively eliminates unsafe interactions during training in deterministic settings and significantly mitigates them in stochastic environments, without the filtering mechanism adversely affecting reward accumulation. In summary, the experimental findings indicate that the proposed framework maintains safety without compromising the agent's learning performance, permitting convergence toward optimal or near-optimal policies.

**Keywords:** Safe Reinforcement Learning, Model-Free Reinforcement Learning, Safety Transformers Filtering.