Kabir Ray Malik
5/5/2025

Research Proposal: Safe Learning for Robots in a Human World

The rapid advancement of reinforcement learning (RL) and its integration into increasingly autonomous robotic systems operating alongside humans raises crucial questions about safety, reliability, and human-robot collaboration. As robots powered by RL become more commonplace in diverse environments such as workplaces, healthcare settings, public spaces, and residential areas, ensuring these systems can reliably learn and execute tasks without posing risks to human safety or damaging property has emerged as a significant historical and contemporary challenge. This research aims to explore how methodologies developed to assess and guarantee the safe implementation of embodied RL robots have evolved along with the field. By critically analyzing technical milestones and innovations, pivotal case studies involving collaborative robots, and foundational academic contributions, this study aims to highlight the effectiveness of various safety assessment practices and pinpoint where these methods have proven insufficient. The stakes involved in adequately addressing these safety and performance concerns are substantial; failures in ensuring robot safety can lead to severe physical harm, undermine public trust in technology and automation, and significantly impede the adoption and broader societal integration of robotic systems. Conversely, a thorough historical perspective on both successful and unsuccessful approaches to RL safety provides invaluable insight, guiding the development of more robust and dependable learning algorithms and safety guidelines. Understanding past mistakes and shortcomings is also essential for preempting similar future errors, protecting human welfare, and fostering a beneficial coexistence between humans and future autonomous robots in society. By ignoring these historical insights into ongoing safety and

collaborative practices, one risks repeating avoidable mistakes, potentially endangering people, hindering technological innovation, and slowing down the incorporation of advanced learning-based robotic systems for the benefit of all humanity.

Historically, the journey toward safe robot learning begins with the rigid, cage-bound manipulators of the late twentieth century. These early machines were deliberately sequestered from workers and governed by explicit movements in code, making their behaviors predictable and their hazards straightforward to isolate. The turn of the millennium, however, ushered in a wave of research that sought to more closely model human learning in robots, giving robots the capacity to adapt to novel tasks through experience. Trial-and-error, something natural to a human learner but previously disregarded for robots, became a design principle. With it arrived a host of safety dilemmas that the older, fenced-off paradigm had conveniently deferred. Engineers soon realized that an algorithm rewarded solely for task completion might discover perilous shortcuts, knocking over a pallet or veering toward a pedestrian if such actions happened to maximize its return. The dramatic televised spills of humanoid platforms in high-profile robotics contests only underscored how fragile autonomous systems could be when exploration was left unchecked. Within laboratories, the debate over whether to remove safety tethers during development captured a deeper tension: progressing quickly often meant courting danger, while proceeding cautiously risked stagnation. Over the 2000s and 2010s, these experiences catalyzed a gradual shift from ad-hoc guardrails toward systematic evaluation. Controlled testbeds, standardized scorecards that penalized unsafe maneuvers, and cross-disciplinary safety reviews emerged as stopgaps against the most conspicuous failure modes. Yet even as these tools matured, real-world deployments in warehouses, clinics, and city streets exposed new edge cases: slippery floors, unpredictable human intent, and shifting weather that laboratory metrics

rarely captured, to name a few. By the early 2020s, the community increasingly acknowledged that safety must be ingrained throughout an RL system's life-cycle, from the first exploratory learning in simulation to continuous monitoring after deployment.

Research on safe reinforcement learning has expanded swiftly over the past decade as roboticists manage the dual mandate of useful performance/ rapid learning and rigorous safety. The field now frames "safe RL" as learning policies that maximize return while keeping specified risk or cost signals within strict bounds during both training and deployment**.** García and Fernández's landmark survey formalized this agenda by distinguishing two complementary paths: reward-centric methods, which embed risk terms or heavy penalties directly into the objective, so the agent internalizes avoidance of hazardous states, and exploration-centric methods, which regulate how the agent gathers experience, for example by masking forbidden actions or consulting external risk models[1]. This taxonomy clarified that shaping the reward alone is rarely sufficient: hand-crafted penalties can be mis-specified or gamed, hence a credible safety solution typically combines reward design with explicit behavioral constraints. The dominant mathematical tool for such constraints is the Constrained Markov Decision Process (CMDP). Rather than folding every concern into one scalar reward, a CMDP lets designers specify distinct cost functions for various conditions (e.g., collision probability, energy overshoot) and require that expected costs never exceed preset thresholds. Achiam et al. operationalized this idea in Constrained Policy Optimization (CPO), an algorithm that performs trust-region updates guaranteeing near-feasible policies at every step[2]. CPO and its successors demonstrate that safe exploration can be achieved by bounding the size of each policy change so

[1] García and Fernández, "Comprehensive Survey on Safe Reinforcement Learning."
[2] Achiam et al., "Constrained Policy Optimization."

constraint violations cannot suddenly spike: a property essential when training occurs directly on hardware. Complementary strategies place protective "shields" around the learner. Saunders et al. introduced a human-in-the-loop protocol in which a person selectively blocks or corrects catastrophic actions while the agent is still naïve; a supervised model then learns to replicate those interventions, gradually reducing human burden while preserving the zero-violation record[3]. Such oversight highlights a pragmatic trade-off: human guardianship is robust but labor-intensive, whereas automated filters demand accurate system models that are seldom available for complex robots.

Another approach is to leverage guarantees from classical control: Berkenkamp et al. combined Gaussian-process models with Lyapunov-based safety certificates, restricting exploration to the region of state space proven (within statistical confidence) to remain stable[4]. The result is a controller that improves performance without ever violating stability constraints, exemplified on an inverted-pendulum task that never allowed the pole to fall throughout learning. This is a vast difference from the typical naïve initialization of a learning agent. Similar work employs control-barrier functions or reachability analysis to compute safe sets and then projects the learned policy back into those sets at run-time. These approaches supply the formal proofs absent from purely empirical techniques, though they too rely on reasonably accurate dynamics estimates. As the methodological arsenal broadened, survey papers began mapping the terrain and identifying gaps. A recurring theme is the need for state-wise constraints: rules such as "never contact a human" that must hold at every instant, not just on average. Zhao et al. catalogue emerging algorithms that honor such instantaneous bounds, from policy-gradient

[3] Saunders et al., "Trial Without Error: Towards Safe Reinforcement Learning via Human Intervention."
[4] Berkenkamp et al., "Safe Model-Based Reinforcement Learning with Stability Guarantees."

variants with hard masks to model-predictive shields that veto unsafe commands in real time[5]. Their review also underscores a persistent gab between simulated algorithm performance and real life (sim2real gap)**:** algorithms validated in perfect simulators often falter on physical robots because of sensor noise, latency, or unmodelled interactions with humans. Consequently, recent literature stresses continuous risk monitoring and logging once a policy is deployed. By streaming safety metrics, recording near-misses, and retraining on those edge cases, practitioners create a feedback loop that gradually tightens safety margins in the messy real world. Taken together, these contributions reveal a maturing yet still fragmented discipline. Reward shaping, CMDP optimization, human supervision, Lyapunov certification, and real-time shielding each address slices of the safety problem, but no single approach fully resolves the tension between rapid learning and guaranteed protection. Moreover, many studies evaluate safety only in simulation or on narrowly scoped hardware tasks, leaving open questions about scalability to high-degree of freedom (DOF) systems like collaborative arms, mobile robots, or autonomous vehicles. Future work, including this research proposal, builds directly on this literature by (i) examining how different assessment practices perform when robots share space with people, (ii) analyzing which combinations of constraints, oversight, and certificates have historically prevented accidents, and (iii) identifying metrics that capture not just average reward but tail-risk events most relevant to human safety.

Understanding how safety assessment practices migrated from ad-hoc safeguards to formal, learning-aware protocols requires tracing the historical record. A touchstone is the DARPA Robotics Challenge (DRC, 2012–2015), which thrust semi-autonomous humanoids into disaster-recovery tasks such as climbing ladders and turning valves. Although few teams used

[5] Zhao et al., “State-Wise Safe Reinforcement Learning: A Survey.”

modern RL, the competition nevertheless exposed the limits of 2010s autonomy under real-world time pressure: robots toppled, stalled, or simply timed out, all while the internet watched and laughed. Atkeson and colleagues' post-mortem tallied dozens of falls and noted that public perception became "robots are slow and fall down," crystallizing the performance versus safety tension that still suspends learning robots today[6]. The detailed team technical reports, available in the *Journal of Field Robotics* special issue and DARPA's archive, explain how operators traded speed for stability, how tethers and emergency stops were deployed, and which fallback behaviors mitigated disaster-level failures[7]. For safe-RL research these documents are more than colorful anecdotes: they catalogue concrete failure modes (unexpected ground contact, joint saturation, sensor dropout) that any future RL controller must detect and avoid.

A second line of understanding can be drawn from the formal safety standards that codify best practice for robots working near humans. *ISO 10218* sets baseline requirements for industrial arms: interlocked cages, emergency-stop circuits, and detailed validation tests, reflective of the decades of hard-won factory experience[8]. The rise of collaborative robots prompted the companion specification *ISO/TS 15066*, which introduces force limits, speed-and-separation monitoring, and power-and-force control for scenarios where humans and machines share the same workspace[9]. These documents supply numerical thresholds that an RL policy must ultimately satisfy, such as maximum allowable contact forces, sensor reaction times, and fault-tolerance levels. They also highlight verification gaps: standards typically assume deterministic controllers, whereas learning policies are stochastic and may drift as they adapt.

---

[6] Atkeson et al., "What Happened at the DARPA Robotics Challenge, and Why?"
[7] DARPA, "DARPA Robotics Challenge (DRC) Program."
[8] ISO, *ISO 10218-1:2011*.
[9] ISO, *ISO/TS 15066:2016*.

Bridging that mismatch, perhaps by embedding ISO-derived constraints into the reward or by certifying the runtime shield against standard-mandated fault models, is an open problem that anchors the present proposal. Case studies of real deployments further enrich this landscape. The first fatal industrial-robot accident in 1979 forced manufacturers to adopt rigorous interlocks, demonstrating that a single mishap can reshape regulation culture. More recently, warehouse fleets and assistive service robots have begun using RL to optimize routing, gripping, or navigation. Although detailed incident logs are scarce, as many firms regard them as proprietary information, conference papers and IEEE archives occasionally surface instructive snippets: near-miss counts during pilot roll-outs, rates of human intervention, wear-and-tear metrics after prolonged learning. Even blog-style disclosures from the right sources are revealing. OpenAI's account of its Rubik's Cube-solving robotic hand discusses force-threshold resets and frequent human restarts during policy fine-tuning, hinting at the engineering overhead required to keep an ambitious RL system within safe bounds.

Curating such literature will help identify which hazards persist despite simulation success and which mitigation patterns generalize across platforms. Benchmark datasets and simulators supply yet another historical waypoint. OpenAI's *Gym* (2016) unified performance-centric evaluation across RL algorithms, but left safety implicit. In response, *Safety Gym* (2019) introduced high-dimensional continuous-control tasks that log both reward and a cost signal for unsafe events (collisions, boundary violations). The accompanying study by Ray et al. quantified progress via metrics such as violation-free episode rate and worst-case return[10], while a companion blog post disseminated code and baseline results to the broader

[10] Ray et al., "Benchmarking Safe Exploration in Deep Reinforcement Learning."

community[11]. Safety Gym's popularity reflects the field's recognition, circa 2019, that repeatable, open benchmarks are essential to move safety from anecdote to science, paralleling the impact ImageNet had on computer vision. Importantly, the benchmark's designers emphasized *online* safety: costs accrue throughout training, not just at convergence, mirroring real-robot constraints where every trial can cause damage.

Survey papers and archival reviews tie these threads together. García and Fernández's 2015 survey, for instance, documents a 2009 autonomous-helicopter challenge in which teams used demonstrations to avoid catastrophic exploration errors: evidence that concerns over the "sim2real gap" predate the deep-learning era. By collating competitions, standards, and benchmark suites, one can trace a clear trajectory: from rigid physical isolation, through human-in-the-loop supervision, toward formalized constraints and algorithmic shields capable of verifying an RL policy in situ. Each stage represents a ratcheting of trust: only when one layer proves inadequate does industry or research add the next. Recognizing gaps in the historical record is equally instructive. Because many commercial RL deployments remain proprietary, publicly available useful datasets of robot learning sessions in human environments are limited. Sharing anonymized logs: frequency of human takeover, distribution of near-miss severities, perhaps in a standardized reporting form, would let researchers benchmark algorithms under authentic risk. Another gap is the absence of a procedural bridge between ISO standards and RL certification: auditors lack clear guidelines for evaluating policies that adapt over time. Documenting experimental attempts at such certification, whether in regulatory sandboxes or industry pilot projects, is vital. The historical sources outlined above, like competition after-action reports, formal standards, bespoke benchmarks, and scattered industrial anecdotes

[11] OpenAI, "Safety Gym."

therefore supply not only cautionary tales but also concrete metrics, thresholds, and failure patterns against which new safe-RL methods can be stress-tested. Grounding the present research in this evidence base ensures that proposed techniques address not just theoretical safety but the pragmatic, historically observed realities of robots and humans coexisting in the same environment.

Securing a clear historical picture of safe reinforcement learning is difficult precisely because the most instructive evidence is often hidden, fragmented, or contested. Proprietary deployment logs stay behind corporate firewalls; the study therefore triangulates open reports such as the DARPA Robotics Challenge after-action dossiers that catalogue every fall and tether-save[12][13], and practitioner interviews to extract trends without breaching confidentiality. A second obstacle is the absence of a shared benchmark: research groups report everything from "collision cost" in Safety Gym[14] to "constraint return" in Constrained Policy Optimization[15], thwarting apples-to-apples comparisons. By rescaling these metrics to a violation-per-1 000-steps baseline and calibrating them against ISO force-and-speed thresholds[16][17], the project proposes a composite safety score that travels across domains. Defining "unsafe" remains slippery, so incidents are partitioned into catastrophic harm, near-misses, and minor infractions following industrial taxonomy and state-wise-constraint theory[18], validated through expert video annotation: echoing the initial human-in-the-loop shielding that slashed catastrophe occurrence

[12] Atkeson et al., "What Happened at the DARPA Robotics Challenge, and Why?"
[13] DARPA, "DARPA Robotics Challenge (DRC) Program."
[14] Ray et al., "Benchmarking Safe Exploration in Deep Reinforcement Learning."
[15] Saunders et al., "Trial Without Error: Towards Safe Reinforcement Learning via Human Intervention."
[16] ISO, *ISO 10218-1:2011*.
[17] ISO, *ISO/TS 15066:2016*.
[18] Zhao et al., "State-Wise Safe Reinforcement Learning: A Survey."

in RL training[19]. Interdisciplinary tensions add yet another layer: engineers focus on dynamics, ethicists on values, regulators on liability, (and computer scientists on why their hardware isn't working). Aligning these lenses requires mapping corporate goals onto theoretical guarantees[20] onto legal standards[21], and enabling the cross-disciplinary conversation needed to achieve that. Critics meanwhile might contend that classic control suffices or that safety cripples learning speed; however historical record shows tasks like dexterous manipulation already depend on RL, and constraint-aware algorithms have cut violation rates with minimal reward loss[22] while human oversight has even accelerated convergence[23].

Looking to the future, the study argues that embedding safety criteria inside the learning loop rather than retrofitting them afterwards offers the most reliable path to true human-friendly robotics. Early evidence shows a clear trajectory: heuristics in the 2000s gave way to constraint-aware optimization[24] and Lyapunov-certified exploration[25], while structured benchmarks like Safety Gym[26] standardized evaluation and accelerated progress. Yet unresolved questions loom: How do single-agent guarantees scale to fleets? How can reward functions be better aligned with nuanced human risk preferences? What certification process will regulators accept or need for adaptive policies? And how can safety generalize under distribution shift? Answering these will require richer public datasets, tighter integration of formal verification with learning, and metrics that capture tail-risk events as faithfully as they capture average reward. By

[19] Saunders et al., "Trial Without Error: Towards Safe Reinforcement Learning via Human Intervention."
[20] Achiam et al., "Constrained Policy Optimization."
[21] ISO, *ISO 10218-1:201*.
[22] Achiam et al., "Constrained Policy Optimization."
[23] Saunders et al., "Trial Without Error: Towards Safe Reinforcement Learning via Human Intervention."
[24] Achiam et al., "Constrained Policy Optimization."
[25] Berkenkamp et al., "Safe Model-Based Reinforcement Learning with Stability Guarantees."
[26] Ray et al., "Benchmarking Safe Exploration in Deep Reinforcement Learning."

grounding new proposals in the documented successes and failures of the past, the research remains tightly coupled to its central concern: ensuring that as learning robots proliferate across human environments performing useful work, they do so not just with competence, but with care.

Bibliography

Achiam, Joshua, David Held, Aviv Tamar, and Pieter Abbeel. "Constrained Policy Optimization." In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 22:22–31. Sydney, Australia: PMLR, 2017. Accessed May 2025. https://arxiv.org/abs/1705.10528.

Atkeson, Christopher G., Peter W. Bluethmann, Benjamin Stephens, et al. "What Happened at the DARPA Robotics Challenge, and Why?" *Journal of Field Robotics* 34, no. 2 (2017): 229–243. Accessed May 2025. https://doi.org/10.1002/rob.21683.

Berkenkamp, Felix, Matteo Turchetta, Angela Schoellig, and Andreas Krause. "Safe Model-Based Reinforcement Learning with Stability Guarantees." In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, edited by I. Guyon et al., 908–918. Long Beach, CA: Curran Associates, Inc., 2017. Accessed May 2025. https://proceedings.neurips.cc/paper_files/paper/2017/file/7eb3c8beff5c5c42a6d408d7b5d9b100-Paper.pdf.

DARPA. "DARPA Robotics Challenge (DRC) Program." Defense Advanced Research Projects Agency. Accessed May 2025. https://www.darpa.mil/program/darpa-robotics-challenge.

García, Javier, and Fernando Fernández. "A Comprehensive Survey on Safe Reinforcement Learning." *Journal of Machine Learning Research* 16, no. 1 (2015): 1437–1480. Accessed May 2025. http://jmlr.org/papers/v16/garcia15a.html.

International Organization for Standardization (ISO). *ISO 10218-1:2011: Robots and Robotic Devices — Safety Requirements for Industrial Robots — Part 1: Robots*. Geneva: ISO, 2011. Accessed May 2025. https://www.iso.org/standard/51330.html.

International Organization for Standardization (ISO). *ISO/TS 15066:2016: Robots and Robotic Devices — Collaborative Robots*. Geneva: ISO, 2016. Accessed May 2025. https://www.iso.org/standard/62996.html.

OpenAI. “Safety Gym.” Blog post, November 21, 2019. Accessed May 2025. https://openai.com/research/safety-gym.

Ray, Alex, Joshua Achiam, and Dario Amodei. “Benchmarking Safe Exploration in Deep Reinforcement Learning.” OpenAI Technical Report, 2019. Accessed May 2025. https://arxiv.org/abs/1910.01708.

Saunders, William, Girish Sastry, Andreas Stuhlmüller, and Owain Evans. “Trial Without Error: Towards Safe Reinforcement Learning via Human Intervention.” *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2018. Accessed May 2025. Extended version: https://arxiv.org/abs/1707.05173.

Zhao, Weiye, Daqing Yi, and Xiangyu Kong. “State-Wise Safe Reinforcement Learning: A Survey.” In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, 2023. Accessed May 2025. https://arxiv.org/abs/2301.11086.