Safe Reinforcement Learning 01: A Brief Introduction

Shangding Gu
https://people.eecs.berkeley.edu/~shangding.gu

EECS, UC Berkeley

February 2025

Content

- What is safe reinforcement learning?
- Growth of safe reinforcement learning.
- Principle of safe reinforcement learning.
- Application of safe reinforcement learning.
- Navigation applications.
- Recommendation literature.
- Questions and homework tasks.

What is safe reinforcement learning?

Why do we need to care about safe reinforcement learning?

- Safe Reinforcement Learning refers to the design and implementation of RL algorithms that explicitly incorporate *safety* constraints during learning and deployment. Safety can be defined in terms of hard constraints (e.g., avoiding dangerous states, satisfying physical limitations) or soft constraints (e.g., minimizing risk, ensuring fairness or ethical behavior) [5].

What is safe reinforcement learning?

Why do we need to care about safe reinforcement learning?

- Safe RL [5] is a subfield of RL [6] that focuses on ensuring the safety of an agent while it explores and optimizes its behavior. It aims to prevent the agent from taking actions that could lead to catastrophic failures, violations of constraints, or unsafe outcomes in real-world applications, such as robotics [4], autonomous driving [2], and LLMs [5].

Figure: Autonomous driving Safety.

What is safe reinforcement learning?

Why do we need to care about safe reinforcement learning?

Safety Definition

- Safety as the ability of agents to operate without causing harm to their environments or themselves across diverse real-world settings, where "harm" refers to any physical, psychological, or environmental damage that may adversely affect agents' environments or themselves [1].

(a) Unsafe

(b) Safe

Figure: Robot-Human Interaction.

Growth of Safe Reinforcement Learning

Increasing Importance of Safety in RL

- Traditional RL focuses on maximizing rewards but ignores safety concerns.
- Safe RL ensures reliability in high-stakes applications (e.g., robotics, healthcare, autonomous systems).

Growth of Safe Reinforcement Learning

Rise in Research and Publications

- Growing interest in Safe RL in top ML/AI conferences (NeurIPS, ICML, ICLR).
- Increased research on constrained RL, risk-aware decision-making, and formal safety guarantees.

Principle of Safe Reinforcement Learning

Safe RL aims to develop AI systems that learn optimal policies while maintaining safety constraints during both training and deployment.

Safety Constraints in Learning

Constrained Optimization Framework

Risk-Aware Decision Making

Safe Exploration

Human-in-the-Loop Safety Mechanisms

Principle of Safe Reinforcement Learning

Safety Constraints in Learning

- RL agents must respect predefined safety constraints while optimizing rewards.
- Constraints can be hard (strict rules) or soft (risk-aware penalties).
- Example: In autonomous driving, Safe RL ensures the car never exceeds speed limits or violates traffic rules.

Principle of Safe Reinforcement Learning

Constrained Optimization Framework

- Safe RL is often modeled as a Constrained Markov Decision Process (CMDP).
- Balances reward maximization with constraint satisfaction.
- Example: A robotic arm optimizes efficiency while ensuring no collisions with humans.

Principle of Safe Reinforcement Learning

Risk-Aware Decision Making

- Optimizes policies to minimize potential catastrophic failures.
- Uses risk-sensitive objectives (e.g., CVaR, worst-case optimization).
- Example: In finance, Safe RL prevents high-risk investments that could cause massive losses.

Principle of Safe Reinforcement Learning

Safe Exploration

- Prevents RL agents from taking unsafe actions while exploring new strategies.
- Uses uncertainty estimation, shielding, and reward shaping.
- Example: In healthcare, Safe RL ensures that experimental treatments do not exceed safe dosage levels.

Principle of Safe Reinforcement Learning

Human-in-the-Loop Safety Mechanisms [3]

- Integrates human oversight to guide RL training and prevent unsafe actions.
- Uses human feedback, expert demonstrations, or intervention mechanisms.
- Example: In robotic surgery, Safe RL relies on surgeon intervention if the AI suggests risky actions.

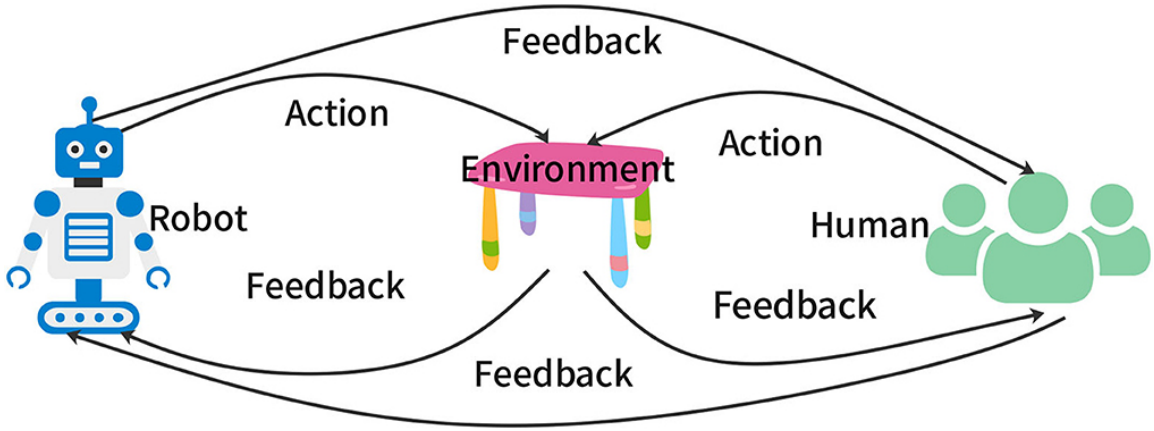

Figure: Human in the loop Safe RL.

Applications of Safe Reinforcement Learning

Safe RL is applied in autonomous vehicles, robotics, healthcare, finance, energy management, and space exploration to ensure safe decision-making and risk-aware optimization. From self-driving cars avoiding collisions to medical AI optimizing treatments without harm, Safe RL enables AI systems to operate reliably in safety-critical environments.

Applications of Safe Reinforcement Learning

Sample applications of safe reinforcement learning:

Autonomous Vehicles

- Ensures safe decision-making in self-driving cars.
- Helps prevent collisions and obey traffic laws.
- Companies like Waymo, Tesla, and Cruise use Safe RL for real-world deployment.
- Example: RL models are trained with safety constraints to handle unexpected pedestrians and adverse weather conditions safely.

Applications of Safe Reinforcement Learning

Sample applications of safe reinforcement learning:

Robotics

- Enables safe interaction between robots and humans in industrial and service environments.
- Used in robotic arms, warehouse automation, and assistive robots.
- Example: Boston Dynamics and Amazon Robotics use Safe RL to prevent robotic failures and ensure human safety.

Applications of Safe Reinforcement Learning

Sample applications of safe reinforcement learning:

Healthcare & Medical AI

- Safe RL optimizes medical treatments while avoiding harmful interventions.
- Applied in personalized medicine, drug dosing, and robotic surgery.
- Example: Safe RL in radiation therapy adjusts dosage levels while minimizing risks to healthy tissues.

Applications of Safe Reinforcement Learning

Sample applications of safe reinforcement learning:

Finance & Risk Management

- Ensures safe investment strategies by managing risk in financial decision-making.
- Used in algorithmic trading, portfolio optimization, and fraud detection.
- Example: Hedge funds apply Safe RL to balance risk and reward in automated trading strategies.

Applications of Safe Reinforcement Learning

Sample applications of safe reinforcement learning:

Energy & Smart Grid Management

- Optimizes energy consumption while maintaining grid stability.
- Applied in power grids, renewable energy distribution, and industrial systems.
- Example: Safe RL is used in wind farm management to optimize power output without exceeding mechanical stress limits.

Applications of Safe Reinforcement Learning

Sample applications of safe reinforcement learning:

Space Exploration

- Enables autonomous control of space robots and planetary rovers.
- Ensures safe navigation and operation in unknown environments.
- Example: NASA uses Safe RL for Mars rover path planning, avoiding hazardous terrain while maximizing exploration efficiency.

Navigation Applications

Safe Reinforcement Learning is widely used in warehouse robots, delivery drones, and search & rescue robots. Example: Amazon Robotics uses Safe RL for warehouse navigation to prevent collisions with humans and objects.

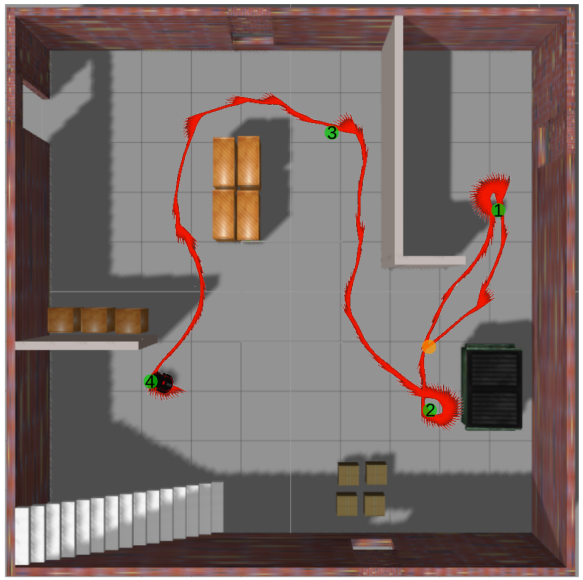

Figure: Navigation Task [1].

[1] https://github.com/reiniscimurs/DRL-robot-navigation

Summary of Safe Reinforcement Learning Applications

The value of safe RL learning technology has been recognized by companies across several industries.

- Safe Robot Control
- Safe Foundation Models, such as LLMs and VLMs
- Power System
- Traffic Management
- Autonomous Driving
- Ship Path Planning
- Semiconductor Manufacturing
- Financial Analysis
- Healthcare

Recommendation Literature.

- Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT Press.
- 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.
- Altman, E. (2021). Constrained Markov decision processes. Routledge.

Questions and Homework Tasks.

- What is the difference between RL and Safe RL?
- Why do we need CMDP for Safe RL?

The End


Shangding Gu.
Safe reinforcement learning to make decisions in robotics.
Technical University of Munich, 2024.


Shangding Gu, Guang Chen, Lijun Zhang, Jing Hou, Yingbai Hu, and Alois Knoll.
Constrained reinforcement learning for vehicle motion planning with topological reachability analysis.
Robotics, 11(4):81, 2022.


Shangding Gu, Alap Kshirsagar, Yali Du, Guang Chen, Jan Peters, and Alois Knoll.
A human-centered safe robot reinforcement learning framework with interactive behaviors.
Frontiers in Neurorobotics, 17:1280341, 2023.


Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang.
Safe multi-agent reinforcement learning for multi-robot control.
Artificial Intelligence, 319:103905, 2023.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll.
A review of safe reinforcement learning: Methods, theories and applications.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

Richard S Sutton and Andrew G Barto.
Reinforcement learning: An introduction.
MIT Press, 2018.

Types of Learning?

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning: Uses

Prediction of future cases

Use the rule to predict the output for future inputs

Knowledge extraction

The rule is easy to understand

Compression

The rule is simpler than the data it explains

Outlier detection

Exceptions that are not covered by the rule, e.g., fraud

Unsupervised Learning

- Learning "what normally happens"
- No output
- Clustering: Grouping similar instances
- Other applications: Summarization, Association Analysis.

Example applications

- Customer segmentation in CRM
- Image compression: Color quantization
- Bioinformatics: Learning motifs

Reinforcement Learning

- No supervised output but delayed reward
- Policies: what actions should an agent take in a particular situation.Utility estimation: how good is a state (used by policy)
- Credit assignment problem (what was responsible for the outcome)

Applications

- Game playing
- Robot in a maze
- Multiple agents, partial observability