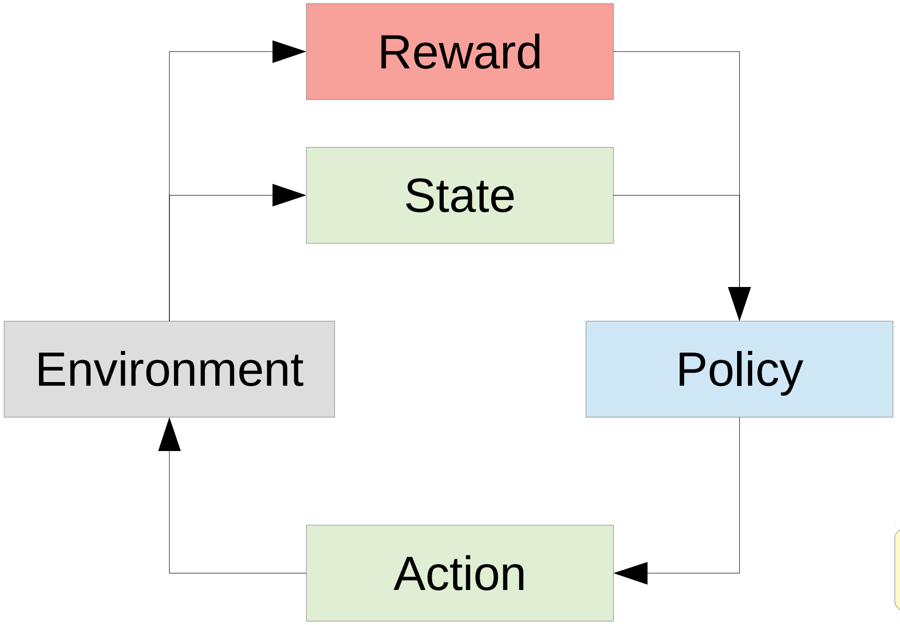
Safe and Verifiable Reinforcement Learning

Reinforcement Learning



MDP States Actions

$$M = \begin{pmatrix} s \in S \\ a \in A \\ P(s' \mid s, a) \\ r : S \times A \rightarrow \mathbb{R} \end{pmatrix}$$

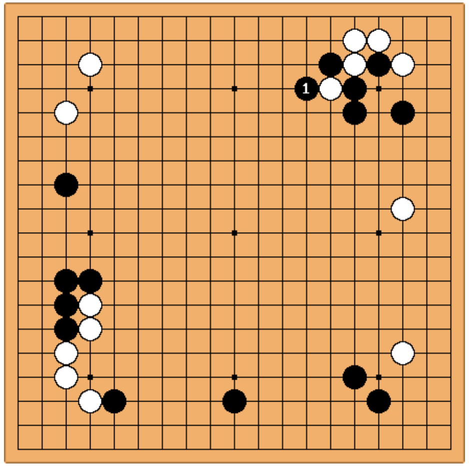
Transition Reward

Policy

$$\operatorname{argmax}_{\pi} R(\pi) = \mathrm{E}\left[\sum_{i=0}^{N} \gamma^{j} r(s_i, a_i)\right]$$

Return Discount factor

Applications



YouTube

Safety-Critical Applications

TESLA

Safety-Critical Applications

The Washington Post
Democracy Dies in Darkness

Tesla driver faces felony charges in fatal crash involving Autopilot

REUTERS® World Business Legal Markets Breakingviews Technology

September 1, 2021
4:31 PM CDT
Last Updated 8 months ago

Autos & Transportation

U.S. identifies 12th Tesla Autopilot car crash involving emergency vehicle

By David Shepardson

NTSB National Transportation Safety Board
Investigations Safety Research News & Events Advocacy Family Assistance Abo

Home › Investigations › Collision Between a Sport Util...

Northbound view of the crash scene before the Tesla was engulfed in flames. (Source: witness S. Engleman)

Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator

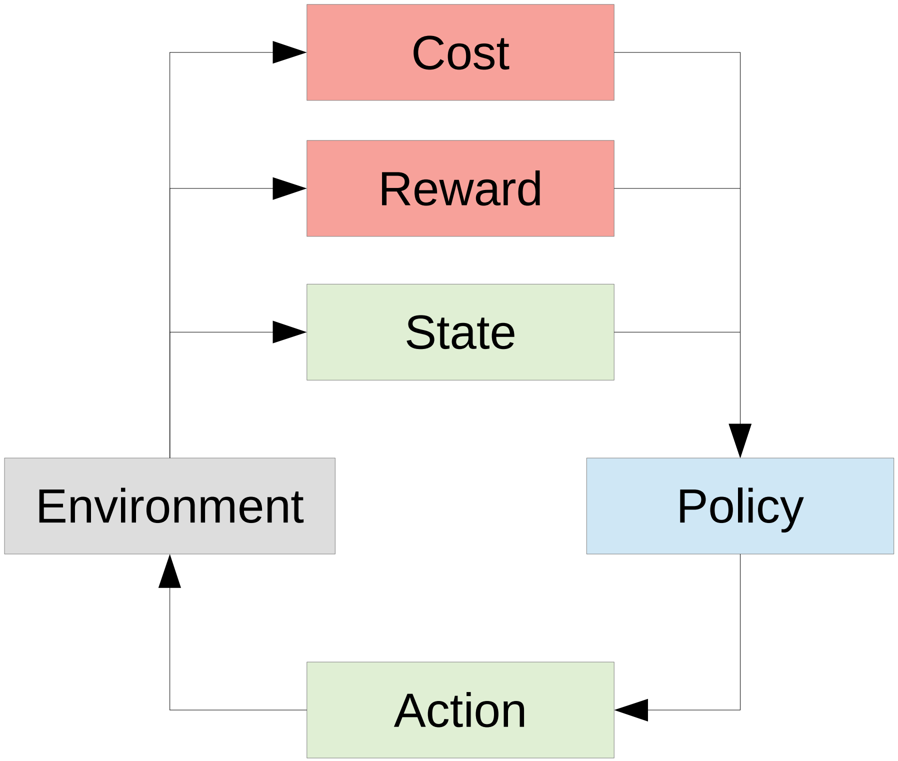
NTSB National Transportation Safety Board
Investigations Safety Research News & Events Advocacy Family Assistance Abo

Home › Investigations › Collision Between Car Operatin...

Right side of car in postcrash damaged condition.

Collision Between Car Operating with Partial Driving Automation and Truck-Tractor Semitrailer

Safe Reinforcement Learning



Constrained MDP

Cost function

Cost limit

$$M = \begin{pmatrix} s \in S \\ a \in A \\ P(s' \mid s, a) \\ r : S \times A \rightarrow \mathbb{R} \\ c : S \times A \rightarrow \mathbb{R} \\ d \in \mathbb{R} \end{pmatrix}$$

$$C(\pi) = \mathrm{E}\left[\sum_{i=0}^{N} \gamma^{i} c(s_i, a_i)\right]$$

$$C(\pi) = \min_x P\left(x \geq \sum_{i=0}^{N} \gamma^{i} c(s_i, a_i)\right) \geq \delta$$

$$\operatorname{argmax}_{C(\pi) \leq d} R(\pi)$$

Lagrange Multipliers

Convert the constrained problem to an unconstrained problem

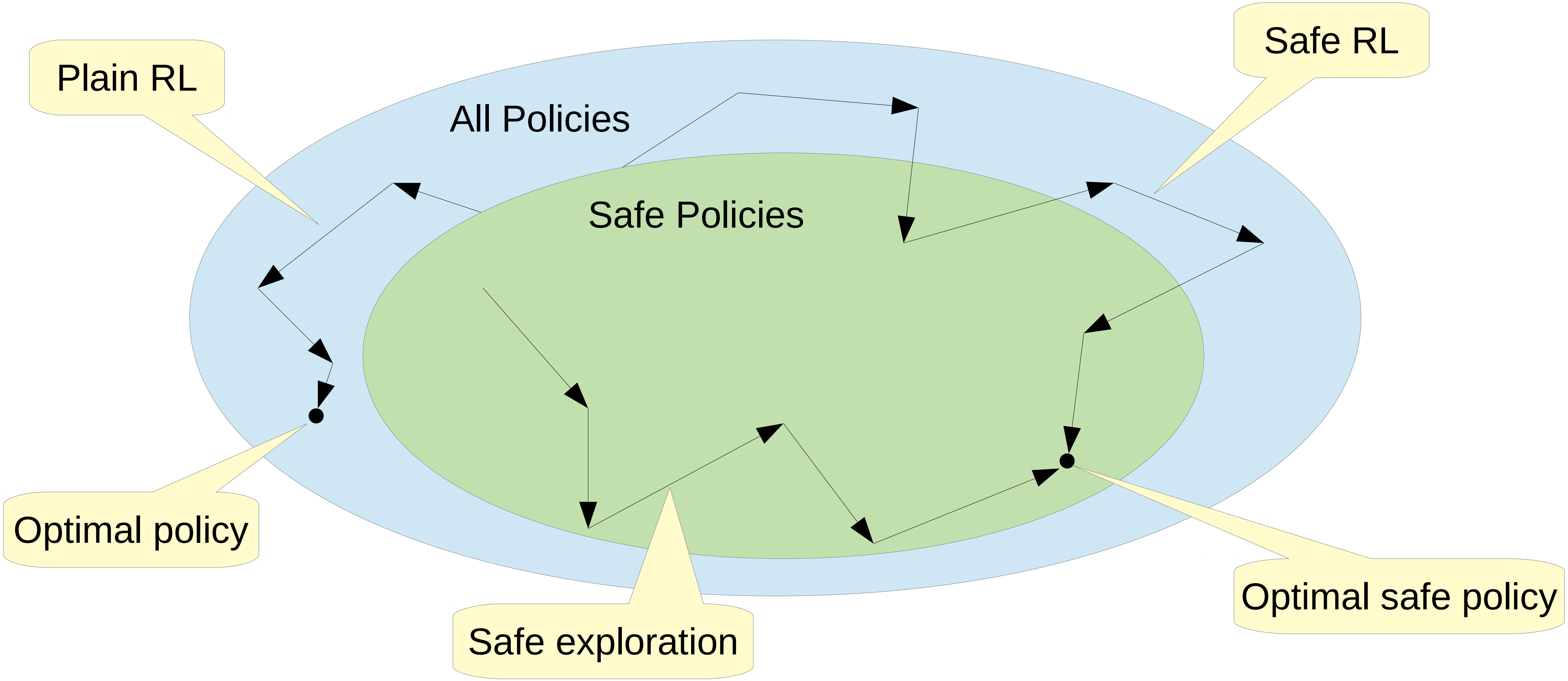
$$\operatorname{argmax}_{C(\pi)\leq d} R(\pi) = \boxed{\min_{\lambda\geq 0} \operatorname{argmax}_{\pi}\left[R(\pi)-\lambda(C(\pi)-d)\right]}$$

Alternate π updates with λ updates

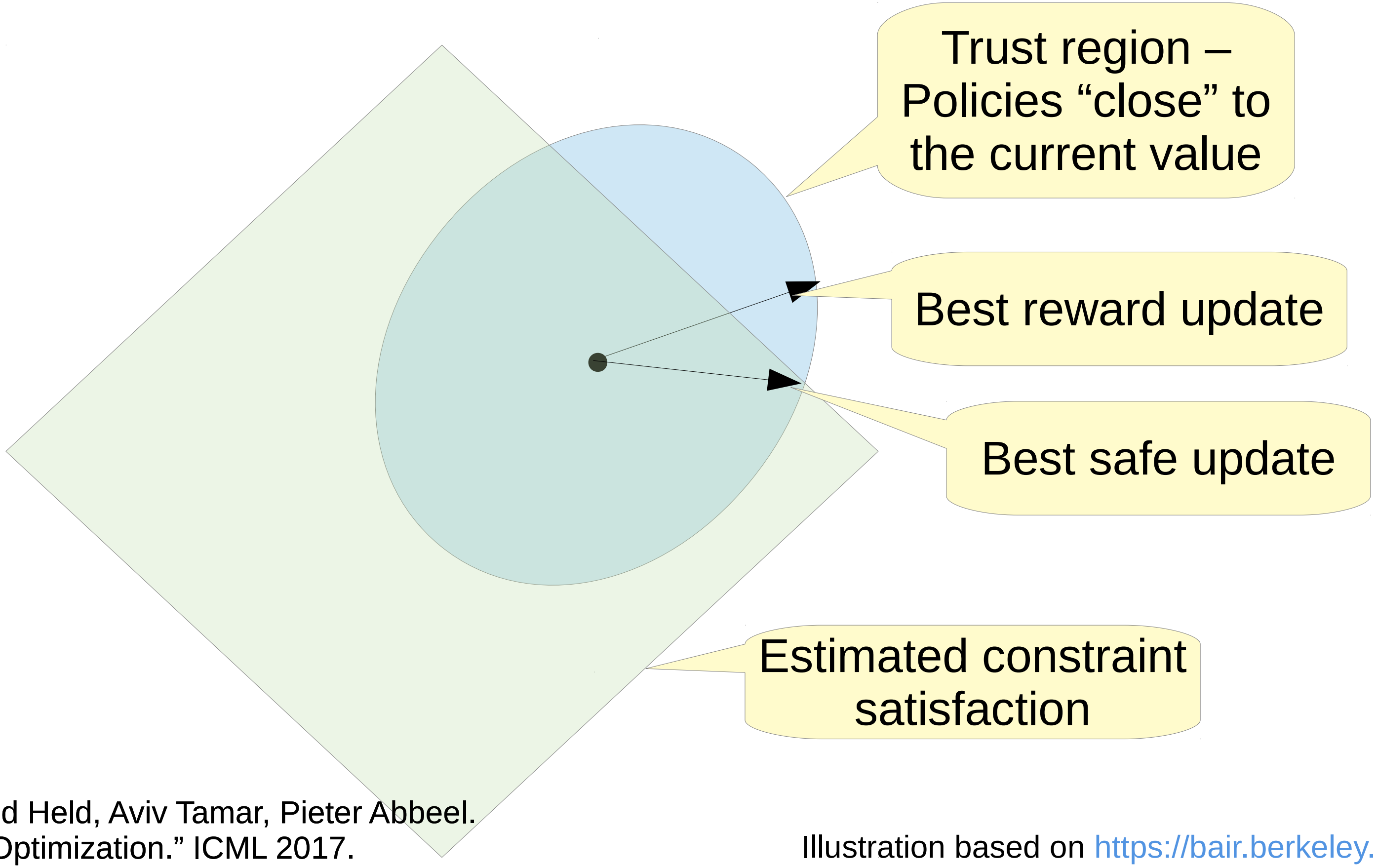
Safety at convergence

Chen Tessler, Daniel J. Mankowitz, Shie Mannor.
“Reward Constrained Policy Optimization.” ICLR 2019.

Safe Exploration



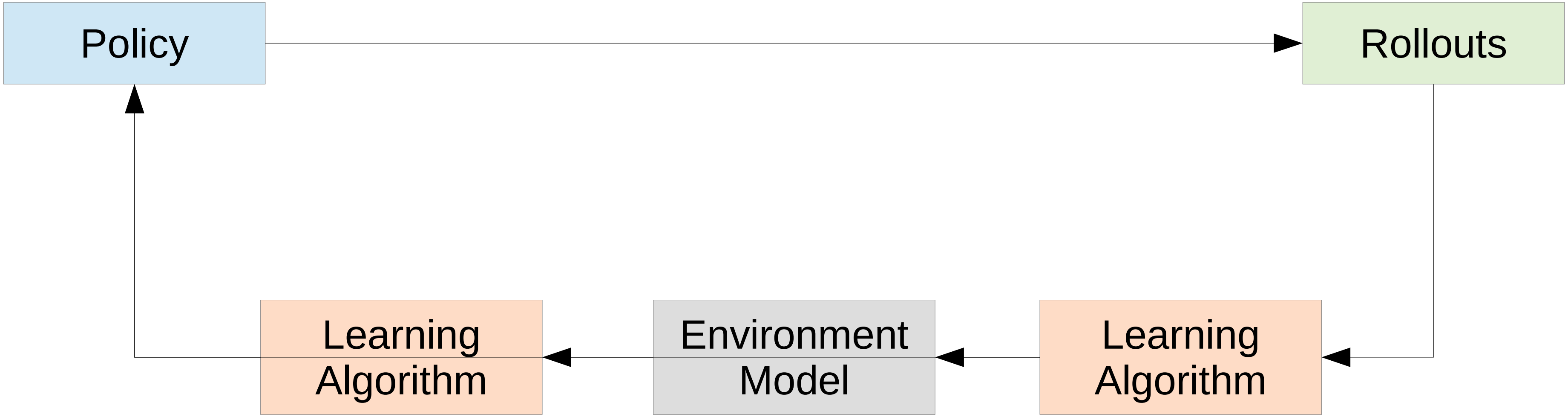
Constrained Policy Optimization



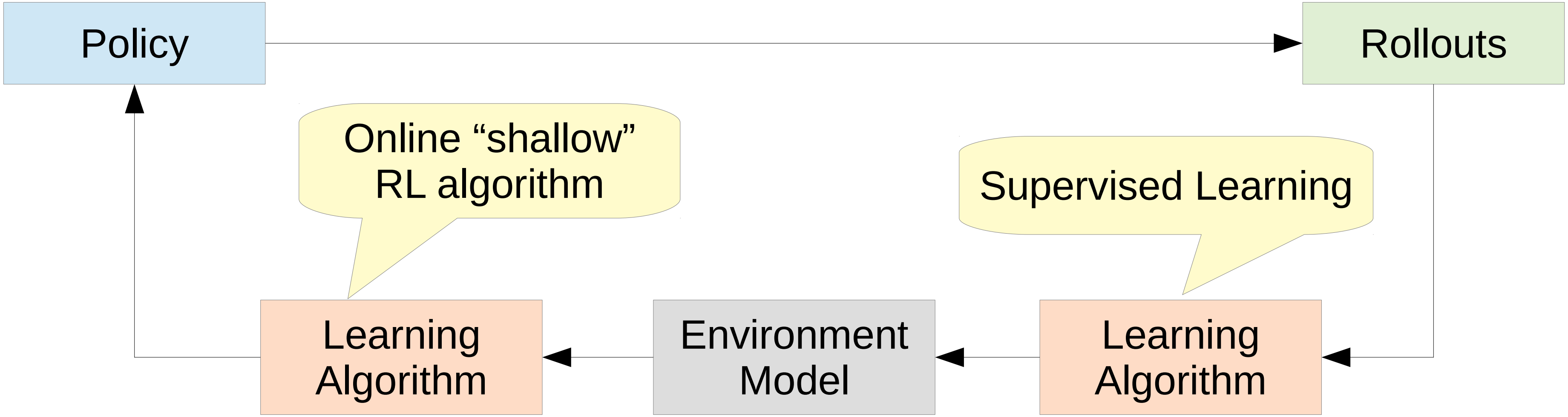
Joshua Achiam, David Held, Aviv Tamar, Pieter Abbeel.
“Constrained Policy Optimization.” ICML 2017.

Illustration based on https://bair.berkeley.edu/blog/2017/07/06/cpo/

Model-Based Reinforcement Learning



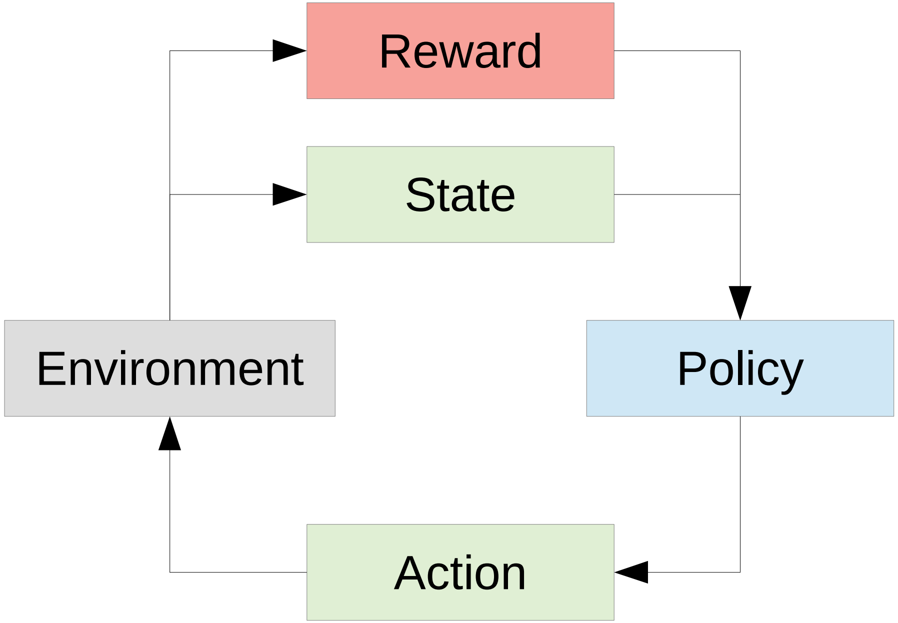
Safe MBRL



Zuxin Liu, Hongyi Zhou, Baiming Chen, Sicheng Zhong, Martial Hebert, Ding Zhao. "Constrained Model-based Reinforcement Learning with Robust Cross-Entropy Method." arXiv 2021.

Guanya Shi, et al. "Neural Lander: Stable Drone Landing Control Using Learned Dynamics." ICRA 2019.

Verified Reinforcement Learning



$$M = \begin{pmatrix} s \in S \\ a \in A \\ P(s' \mid s, a) \\ r : S \times A \rightarrow \mathbb{R} \end{pmatrix}$$

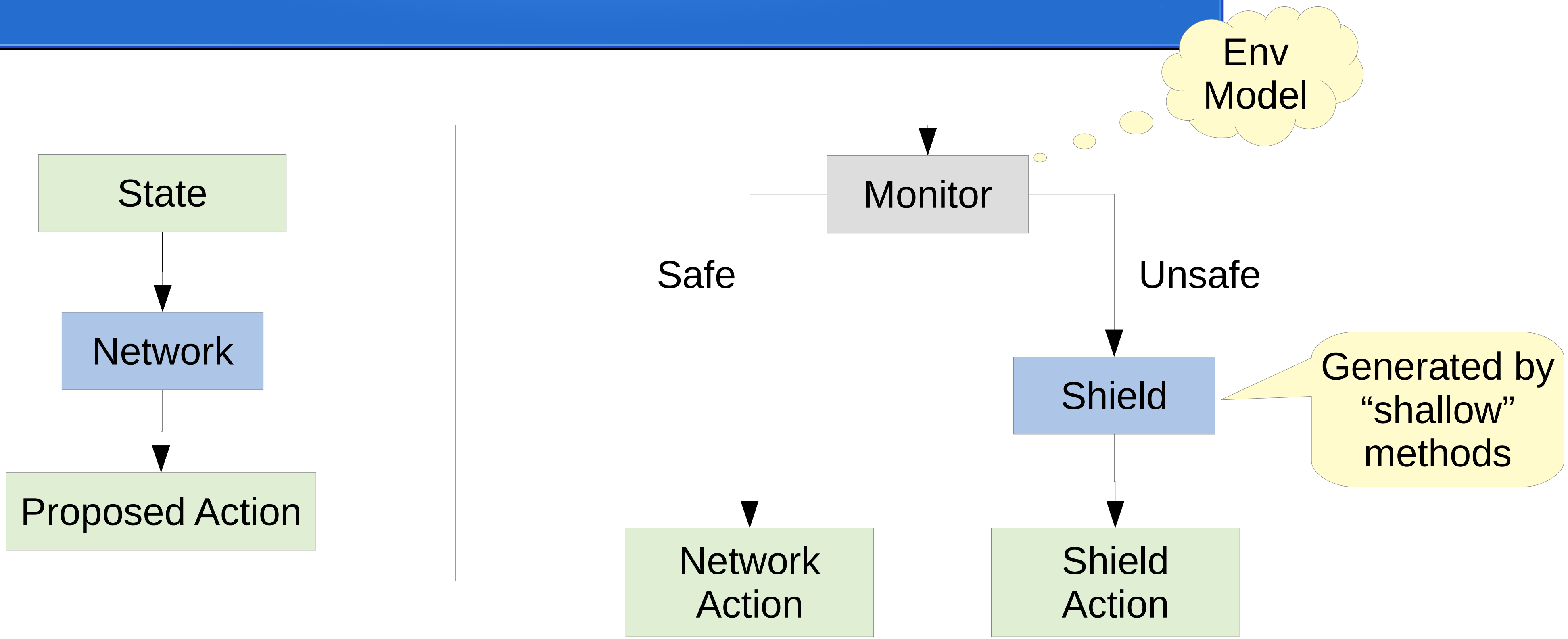
Unsafe states

$$S_U \subset S$$

$$\mathrm{Safe}(\pi) \;:=\; \forall i, P_\pi(s_i \in S_U) = 0$$

$$\operatorname{argmax}_{\mathrm{Safe}(\pi)} R(\pi)$$

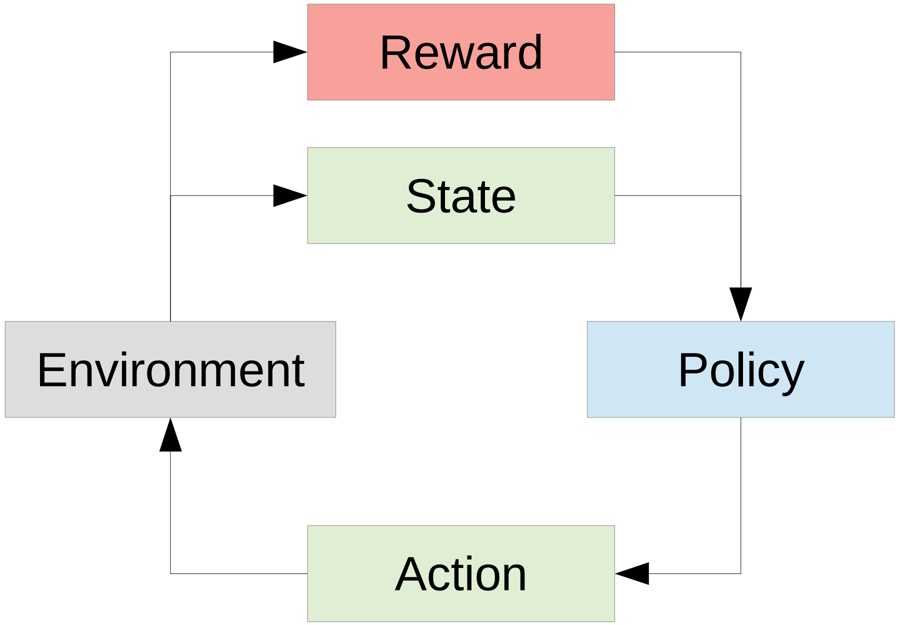
Shielding

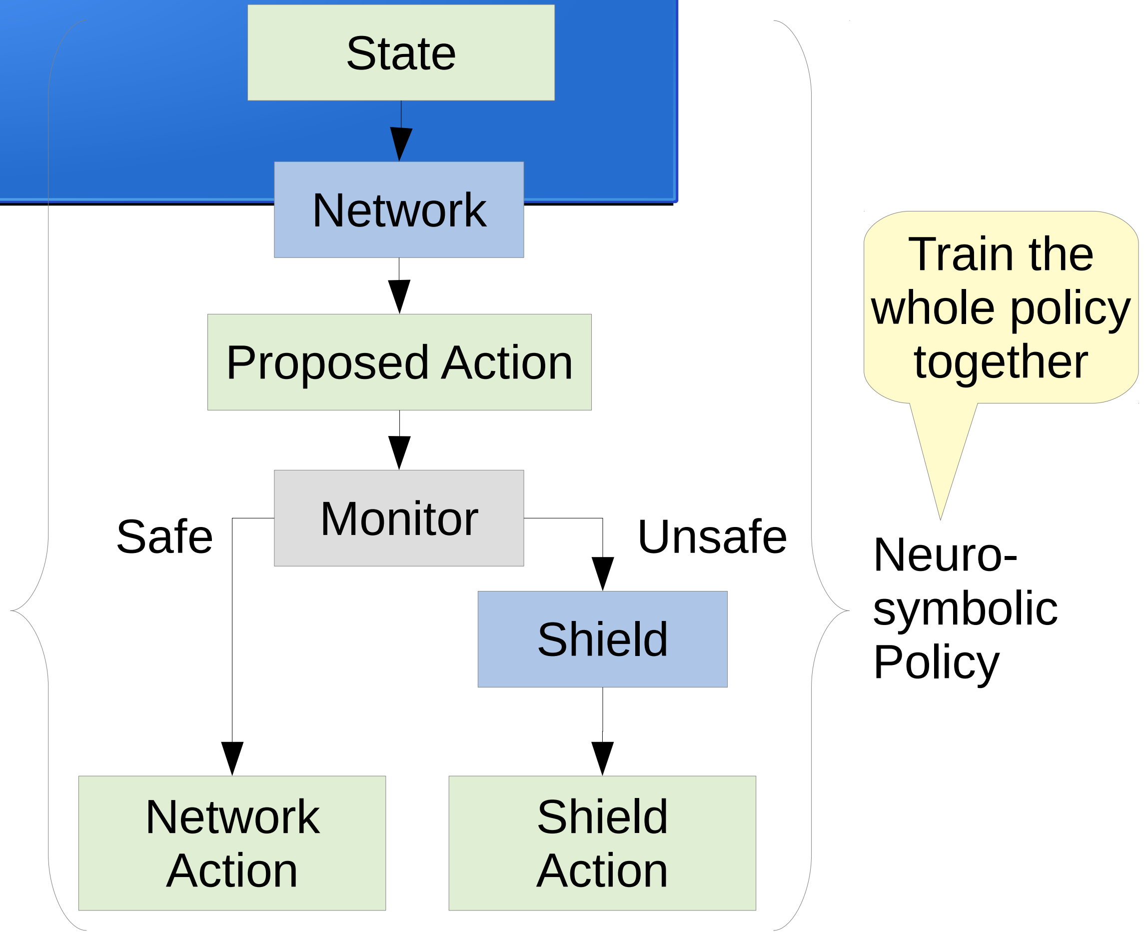


Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, Ufuk Topcu. "Safe Reinforcement Learning via Shielding." AAAI 2018.

He Zhu, Zikang Xiong, Stephen Magill, Suresh Jagannathan. "An Inductive Synthesis Framework for Verifiable Reinforcement Learning." PLDI 2019.

REVEL



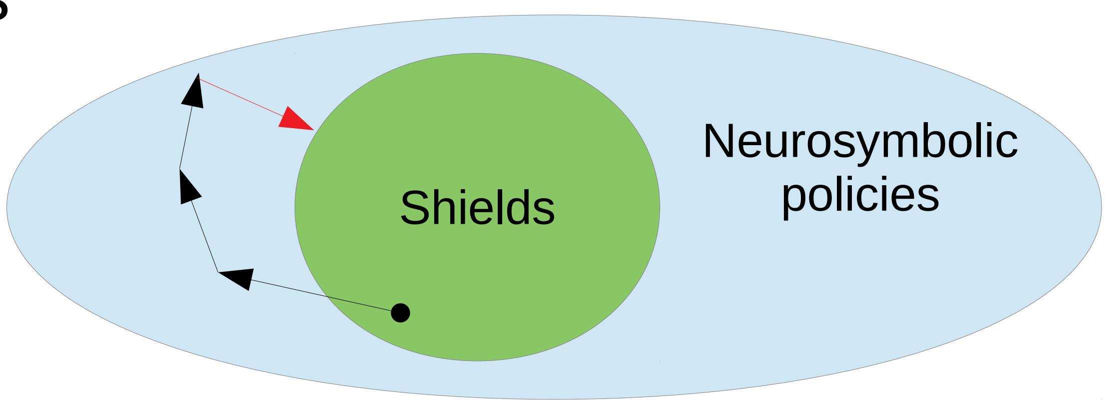


$$\pi(s) = \text{if } \phi(s, f(s)) \text{ then } f(s) \text{ else } g(s)$$

Greg Anderson, Abhinav Verma, Isil Dillig, Swarat Chaudhuri. “Neurosymbolic Reinforcement Learning with Formally Verified Exploration.” NeurIPS 2020.

Mirror Descent (for RL)

- *Lift* a shield to a neurosymbolic policy
- *Update* the policy in the neurosymbolic space
- *Project* the resulting neurosymbolic policy back onto the space of shields



Abhinav Verma, Hoang M. Le, Yisong Yue, Swarat Chaudhuri. "Imitation-Projected Policy Gradient for Programmatic Reinforcement Learning." NeurIPS 2019.

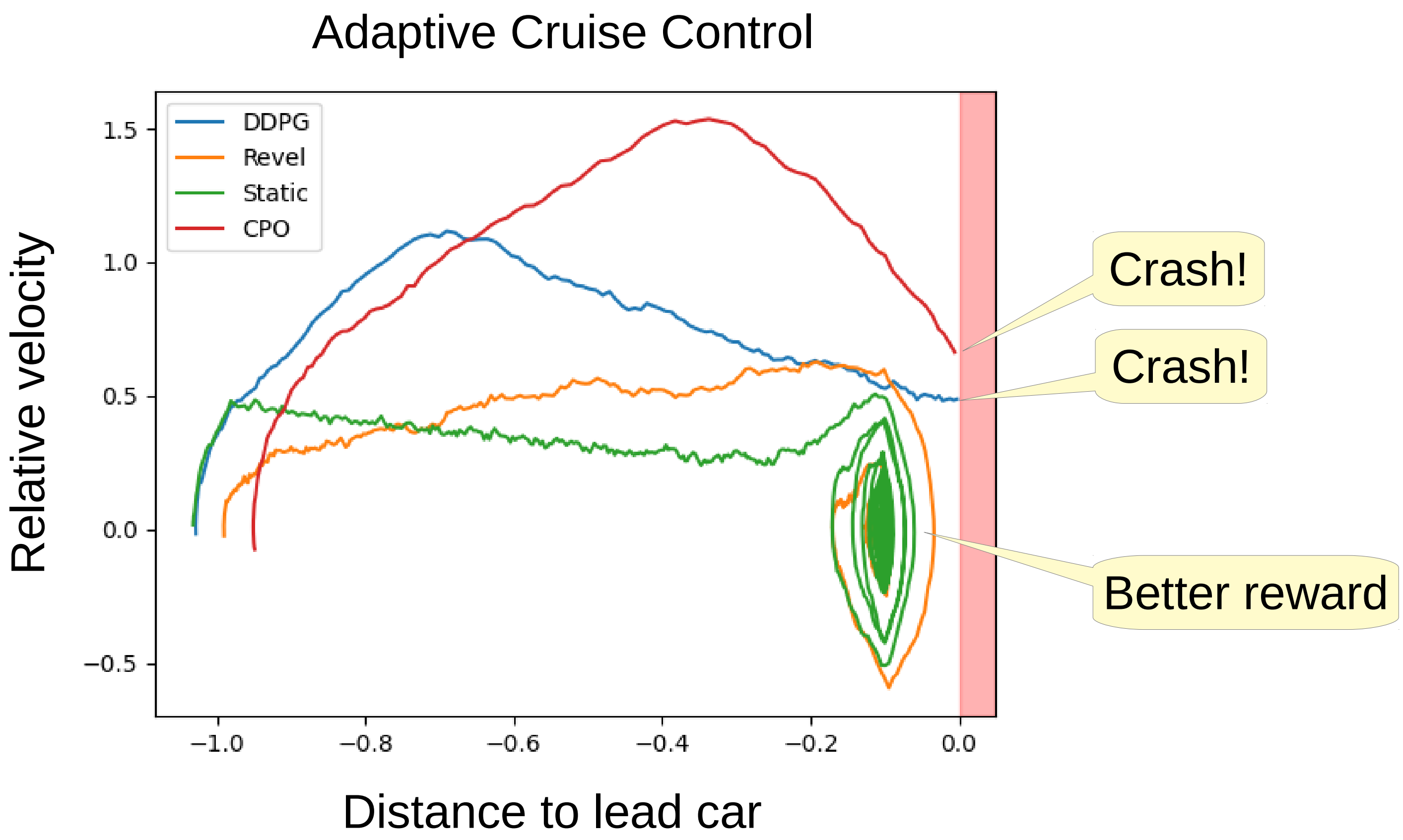
Mirror Descent in REVEL

- *Lift* a shield to a neurosymbolic policy
  - Imitation learning: $g(s) \rightarrow \text{if } \phi(s, f_g(s)) \text{ then } f_g(s) \text{ else } g(s)$
- *Update* the policy in the neurosymbolic space
  - Gradients descent on the neural component
- *Project* the resulting neurosymbolic policy back onto the space of shields
  - Imitation learning once again

Neural networks are universal approximators

Lots of theory here

Results



Adaptive Cruise Control
DDPG
Revel
Static
CPO
1.5
1.0
0.5
0.0
−0.5
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
Relative velocity
Distance to lead car
Crash!
Crash!
Better reward