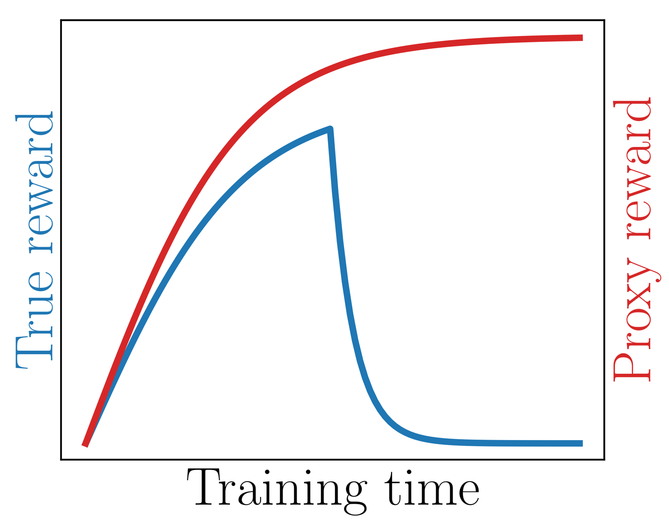
Defining and Characterizing Reward Gaming

Joar Skalse*, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, David Krueger*

Reward Gaming

A proxy reward function is ***ungameable*** if increasing expected proxy return can never decrease expected true return.

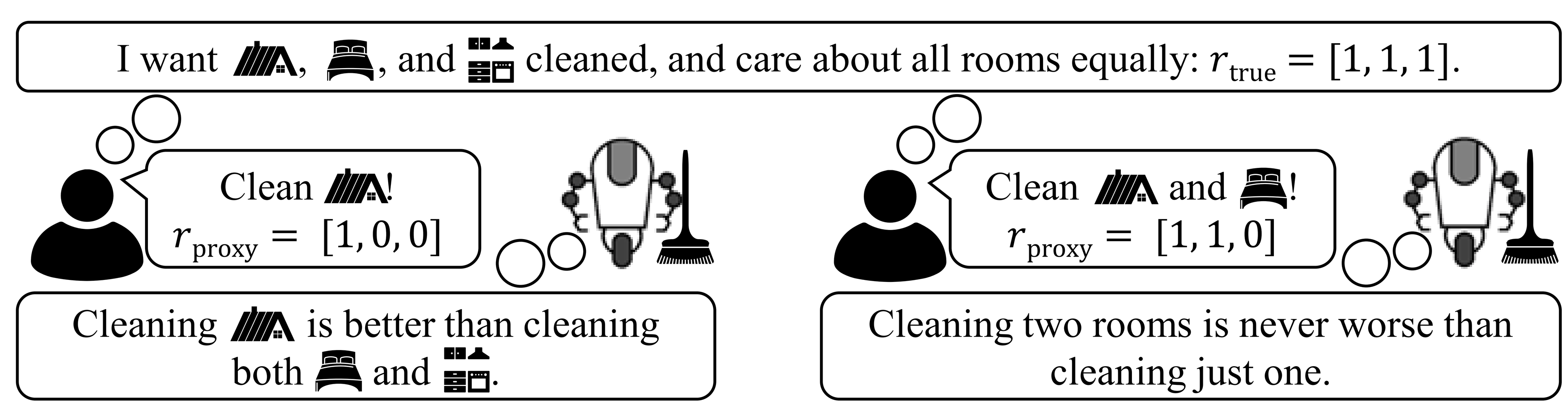


An illustration of reward gaming when optimizing a gameable proxy. Researchers observe this in practice.

Is it feasible to specify ungameable proxy rewards?

Perhaps we could make an ungameable proxy by:

- leaving some terms out of the reward function (making it "narrower")
- overlooking fine-grained distinctions between similar outcomes



(a) r_{proxy} is gameable (b) r_{proxy} is not gameable

Defining Reward Gameability and Simplification

Definition 1. A pair of reward functions R_1, R_2 are ***gameable*** relative to policy set Π and environment $(S, A, T, I, \_, \gamma)$ if there exist $\pi, \pi' \in \Pi$ such that

$$J_1(\pi) < J_1(\pi') \quad \& \quad J_2(\pi) > J_2(\pi'),$$

where $J_i(\pi)$ is the expected return of π according to R_i.

So, gameability occurs when two reward functions rank two policies differently.

Definition 2. R_2 is a ***simplification*** of R_1 relative to policy set Π if for all $\pi, \pi' \in \Pi$

$$\begin{aligned} & J_1(\pi) < J_1(\pi') \Rightarrow J_2(\pi) \leq J_2(\pi') \\ \& \quad & J_1(\pi) = J_1(\pi') \Rightarrow J_2(\pi) = J_2(\pi') \end{aligned}$$

and there exist $\pi, \pi' \in \Pi$ such that

$$J_2(\pi) = J_2(\pi') \quad \& \quad J_1(\pi) \neq J_1(\pi').$$

One reward function is a simplification of another when it induces the same policy ordering, but sets at least two adjacent policies equal, while the original reward sets them not equal.

A reward function R is ***trivial*** if it sets the values of all policies equal:

$$J_1(\pi) = J_1(\pi') \quad \forall\, \pi, \pi' \in \Pi.$$

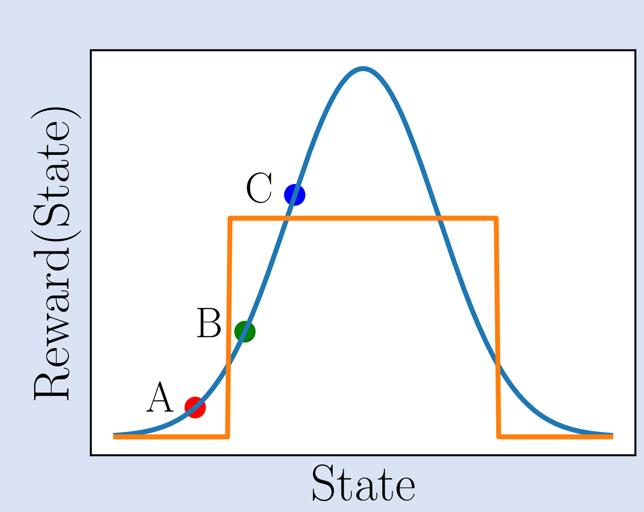
Results

Theorem 1. If the policy set contains an open set, then all nontrivial reward functions are gameable with respect to all other nontrivial reward functions.

Theorem 2. Given a finite policy set and a reward function R, we can always find a different, nontrivial reward function which is ungameable with respect to R.

Theorem 3. Given a finite policy set and a reward function R, we provide necessary and sufficient conditions for existence of a nontrivial simplification of R.

Examples



Despite the step function seeming like a simplification of the Gaussian, these reward functions are gameable.

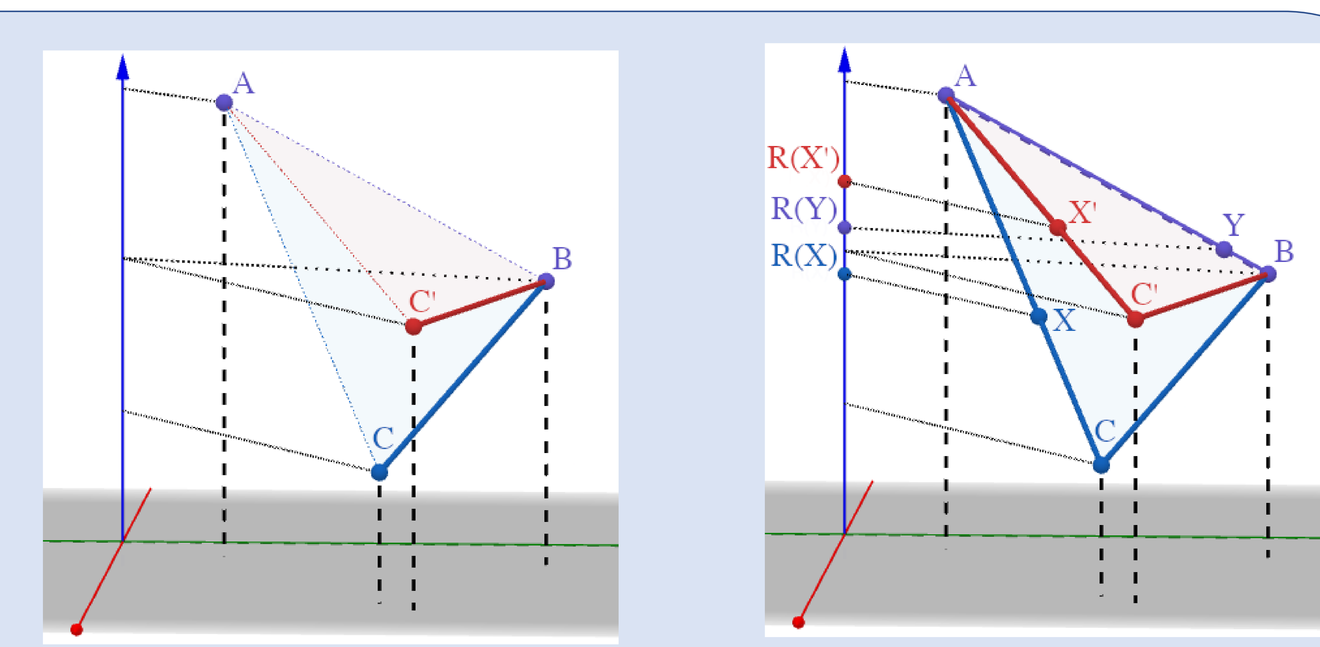
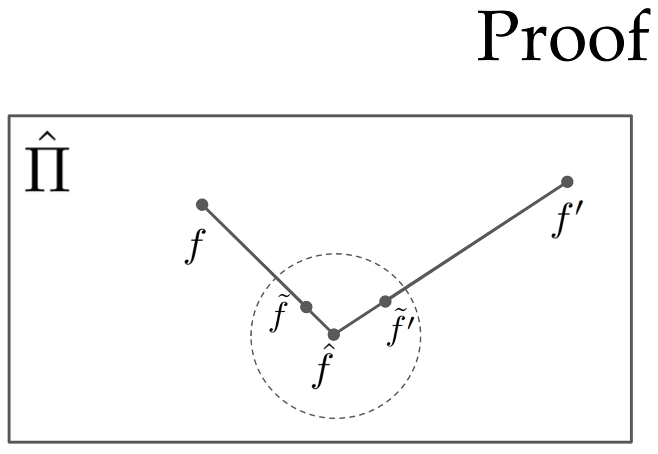


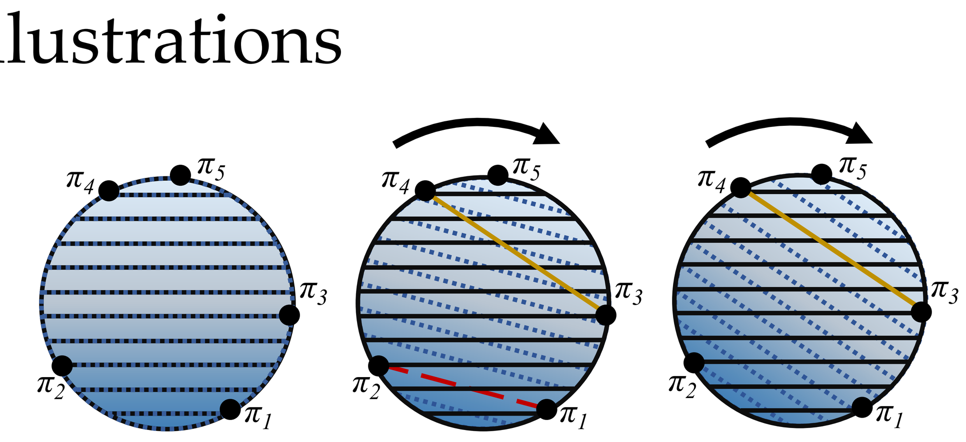
Illustration of two results of simplification on infinite policy sets.

- Left: nontrivial simplification is possible by keeping policies A and BC at different heights.
- Right: attempting the same simplification results in gameability; the only possible simplification is the trivial one.

Proof illustrations



Theorem 1



Theorem 2

Limitations

- Definition may be **too strict**: gameability is far from a guarantee of gaming.
- Definition is **symmetric**, but behaviors with low proxy reward and high true reward are much less concerning than the reverse (our agent probably won't solve climate change while learning to wash dishes)