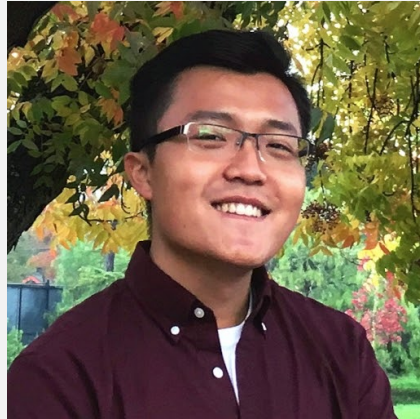


Assistive Robust Reward Design

RSS IDA Workshop 2020



Jerry Zhi-Yang He
UC Berkeley



Anca D. Dragan
UC Berkeley

AI systems treat reward functions as set in stone.

$$R_{\omega}(s, a)$$

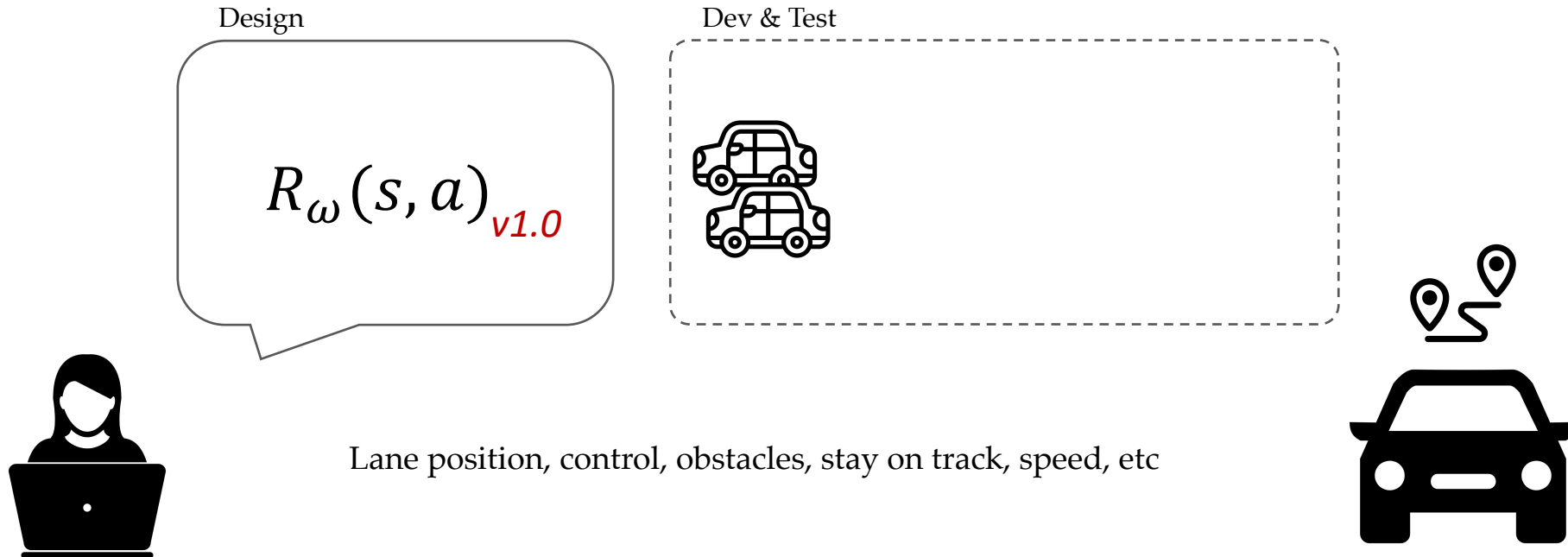


~~AI systems treat reward functions as set in stone.~~
We write reward functions iteratively.

$$R_{\omega}(s, a)$$



~~AI systems treat reward functions as set in stone.~~
We write reward functions iteratively.

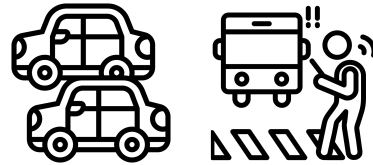


~~AI systems treat reward functions as set in stone.~~
We write reward functions iteratively.

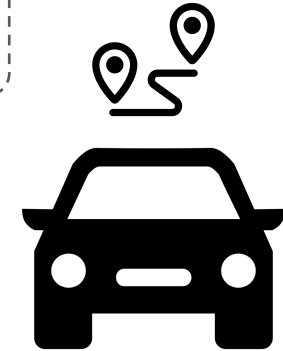
Design

$$R_{\omega}(s, a)_{v2.0}$$

Dev & Test



Lane position, control, obstacles, stay on track, speed, etc



~~AI systems treat reward functions as set in stone.~~
We write reward functions iteratively.

Design

$$R_{\omega}(s, a)_{v3.0}$$

Dev & Test



Lane position, control, obstacles, stay on track, speed, etc

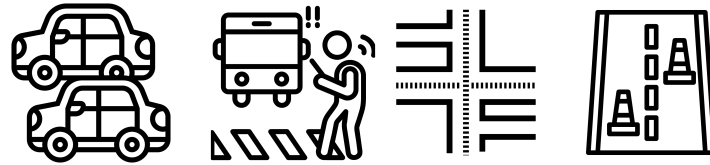


~~AI systems treat reward functions as set in stone.~~
We write reward functions iteratively.

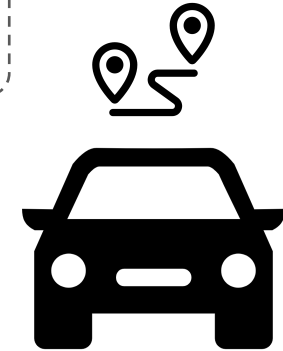
Design

$$R_{\omega}(s, a)_{v4.0}$$

Dev & Test



Lane position, control, obstacles, stay on track, speed, etc



The AI system should account for the *iterative nature* of the reward design process, rather than treat the currently specified reward as *set in stone*.

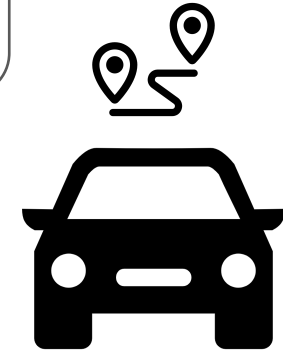
1. Realize that the current proxy reward is an “evidence”.

Current Evidence

$$R_{\omega}(s, a)_{v1.0}$$



$$\cancel{R^* = R_{\omega v1.0}}$$



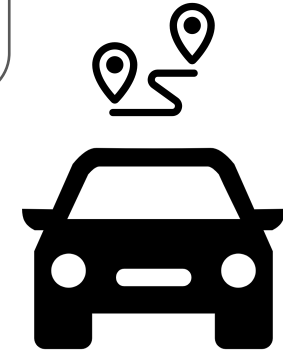
1. Realize that the current proxy reward is an “evidence”.

Current Evidence

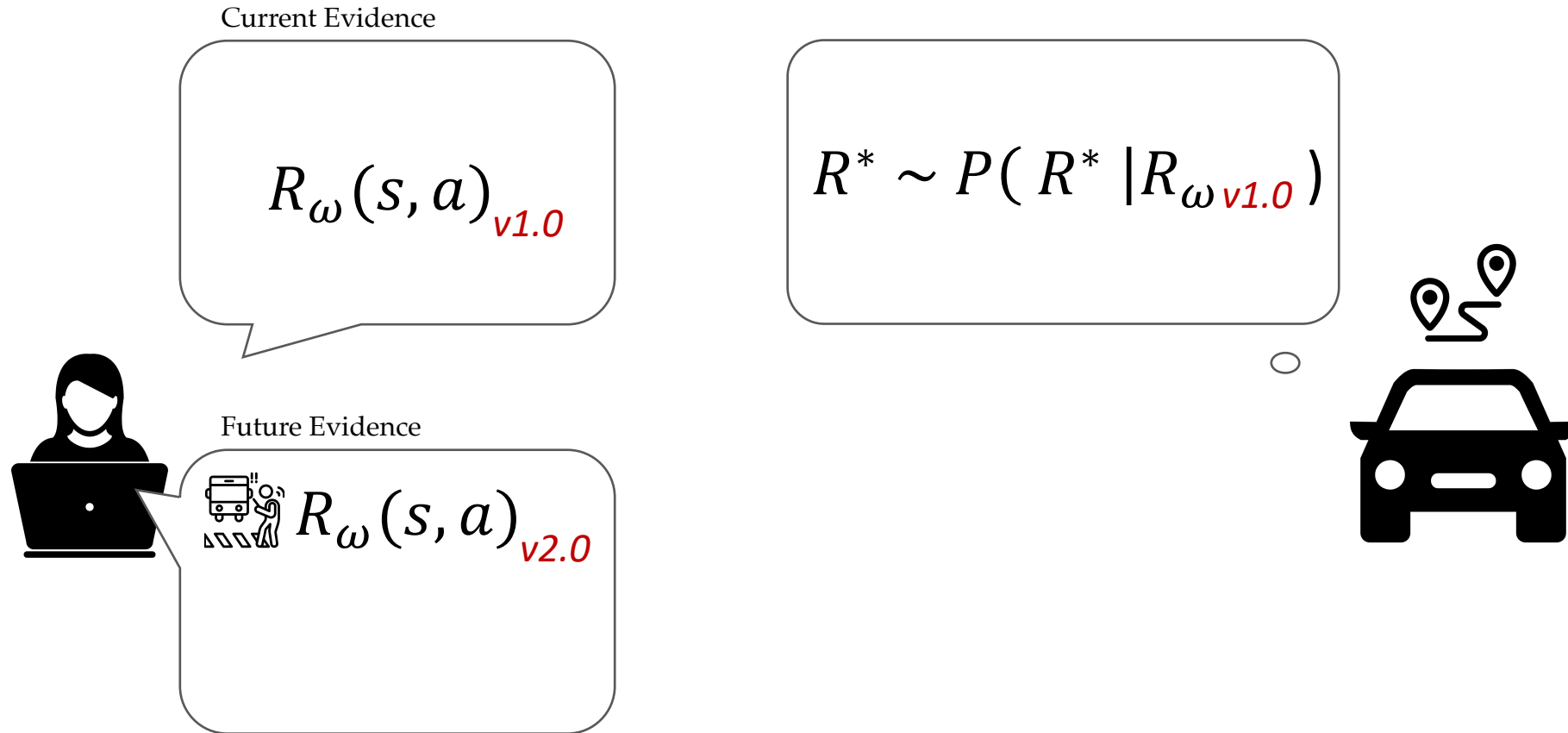
$$R_{\omega}(s, a)_{v1.0}$$



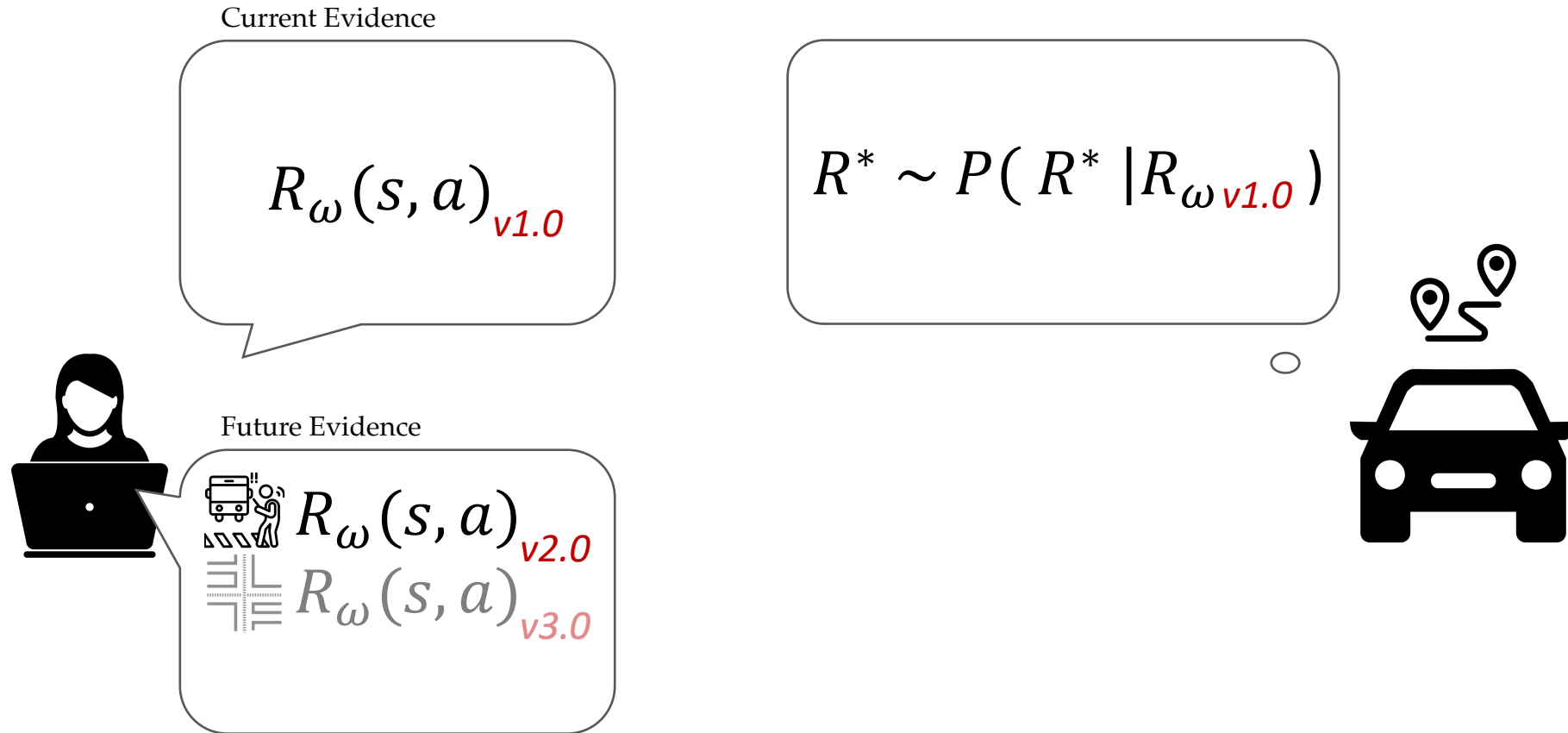
$$\begin{aligned} & \cancel{R^* = R_{\omega v1.0}} \\ R^* & \sim P(R^* | R_{\omega v1.0}) \end{aligned}$$



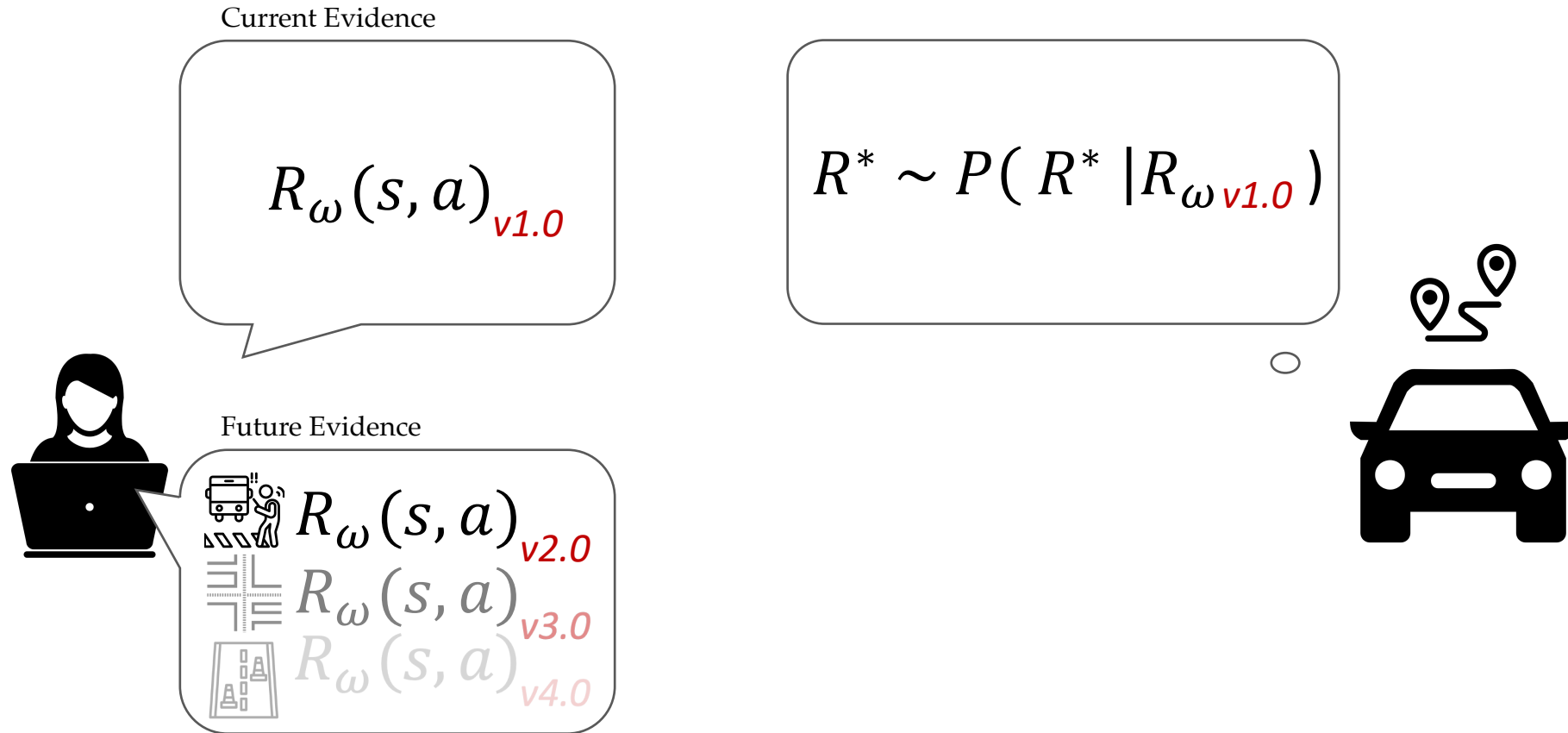
1. Realize that the current proxy reward is an “evidence”.
2. Account for “future evidence”.



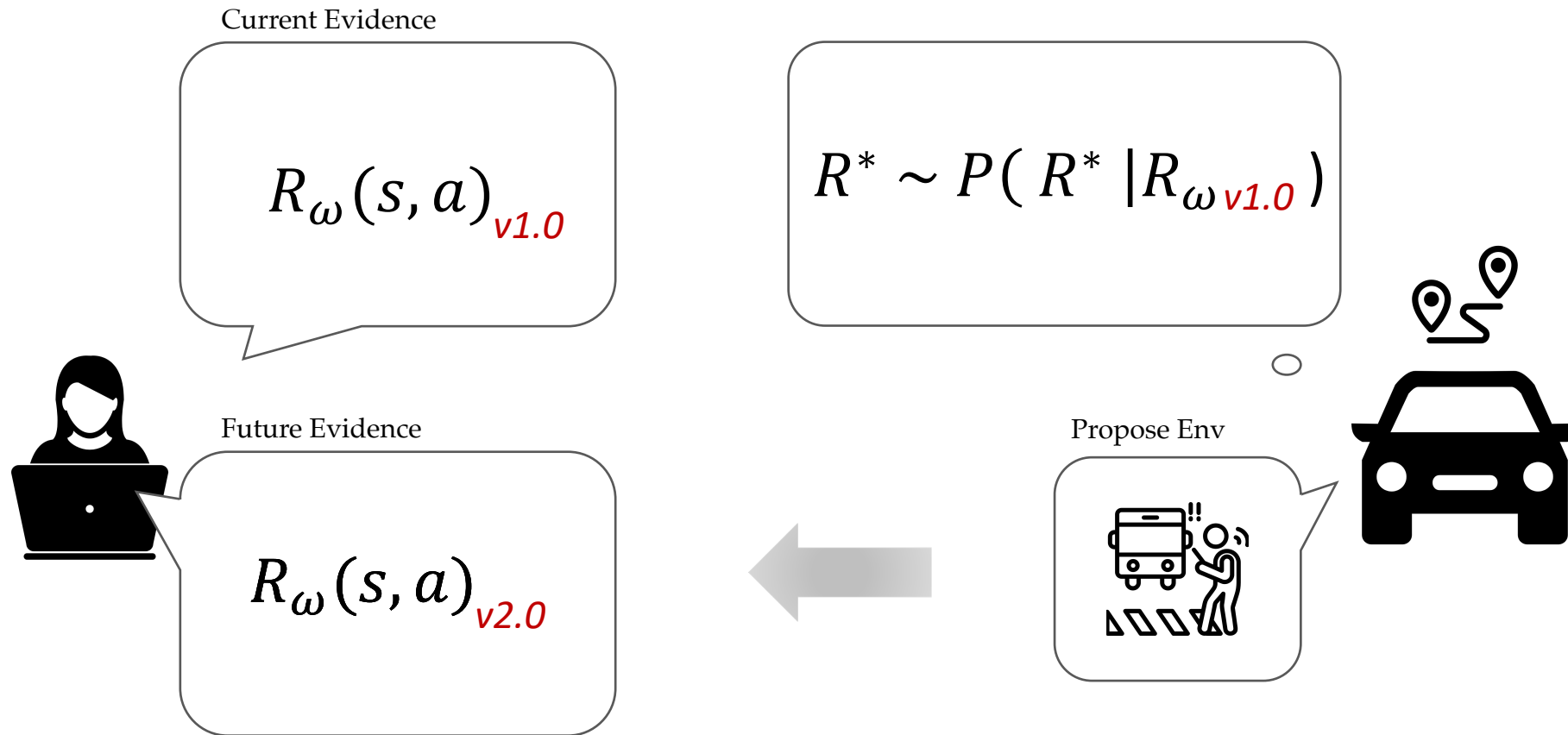
1. Realize that the current proxy reward is an “evidence”.
2. Account for “future evidence”.



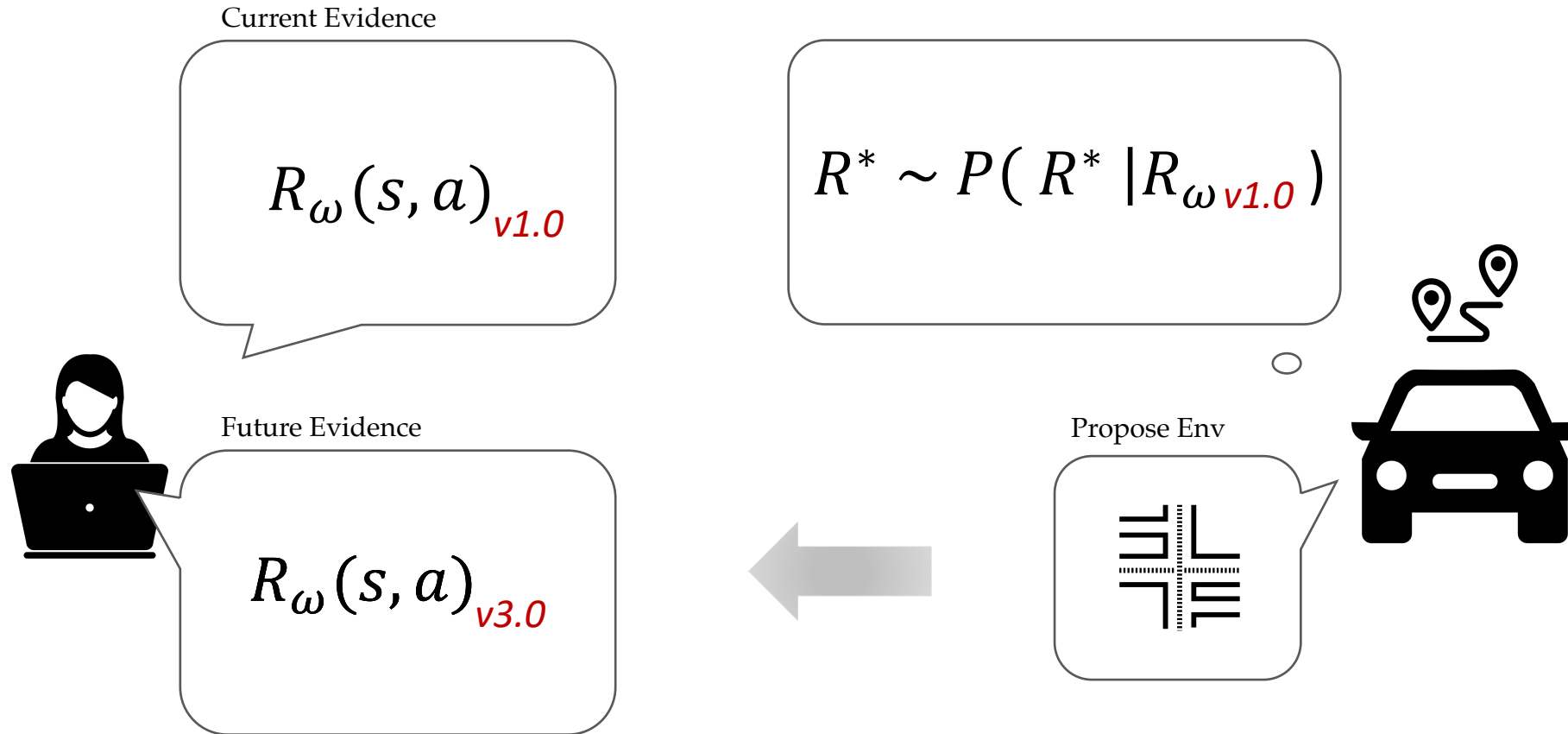
1. Realize that the current proxy reward is an “evidence”.
2. Account for “future evidence”.



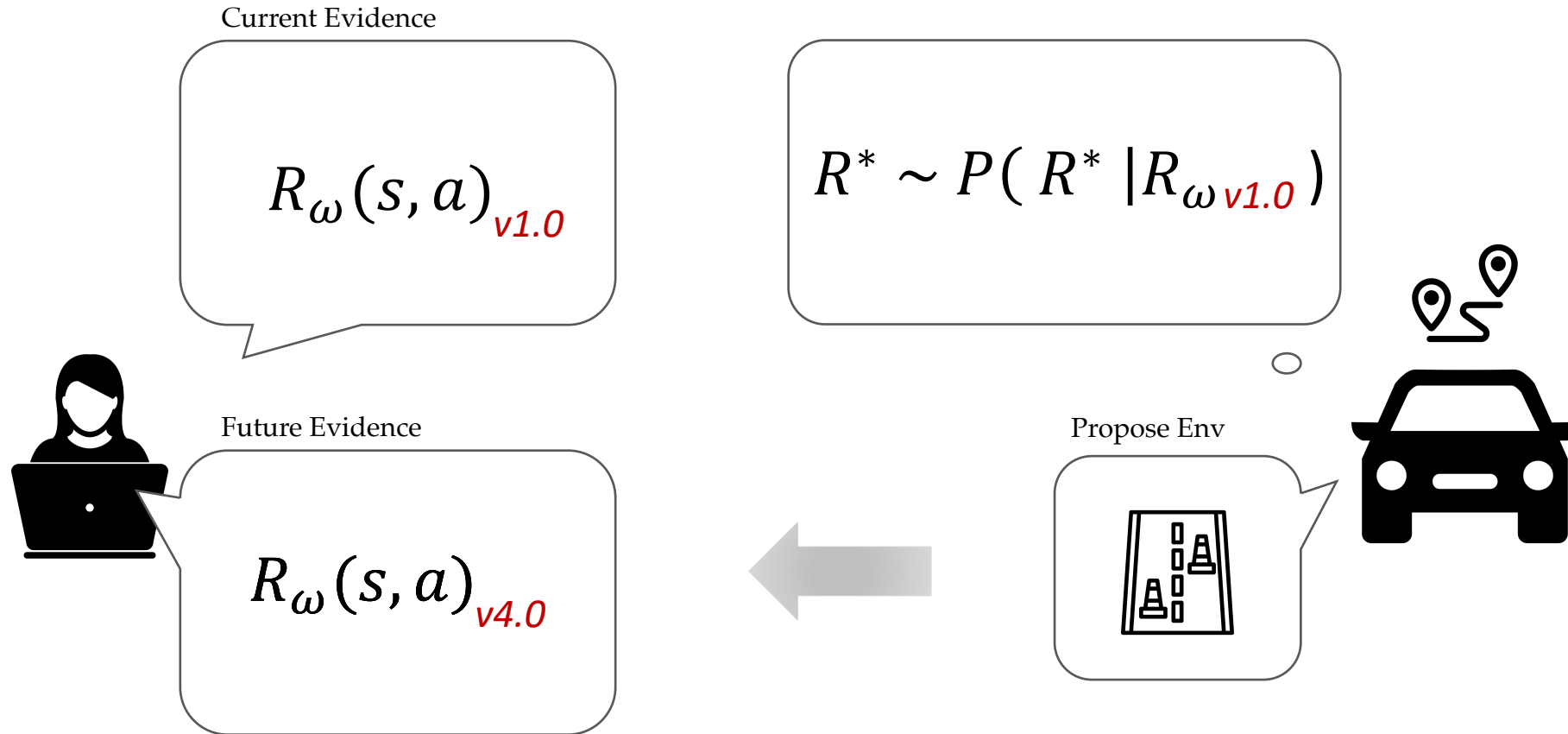
1. Realize that the current proxy reward is an “evidence”.
2. Account for “future evidence”.
3. Act to influence the designer.



1. Realize that the current proxy reward is an “evidence”.
2. Account for “future evidence”.
3. Act to influence the designer.



1. Realize that the current proxy reward is an “evidence”.
2. Account for “future evidence”.
3. Act to influence the designer.



Key Insight

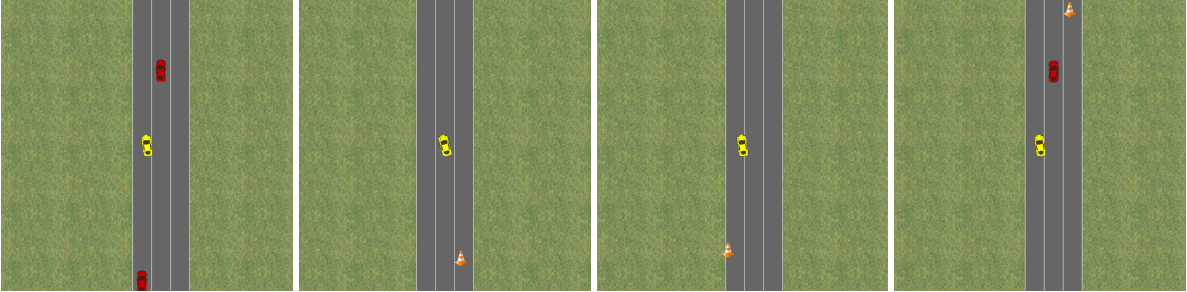
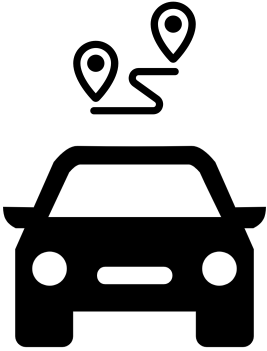
An assistive reward design system should **actively expose** the designer to the environments that have the **most potential to narrow down** what the reward should be.



Finding Informative Edge Cases

Current Evidence

$$R_{\omega}(s, a)_{v1.0}$$



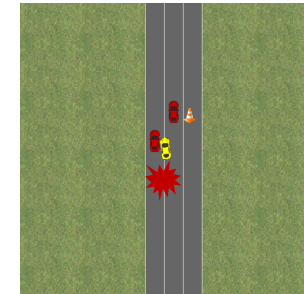
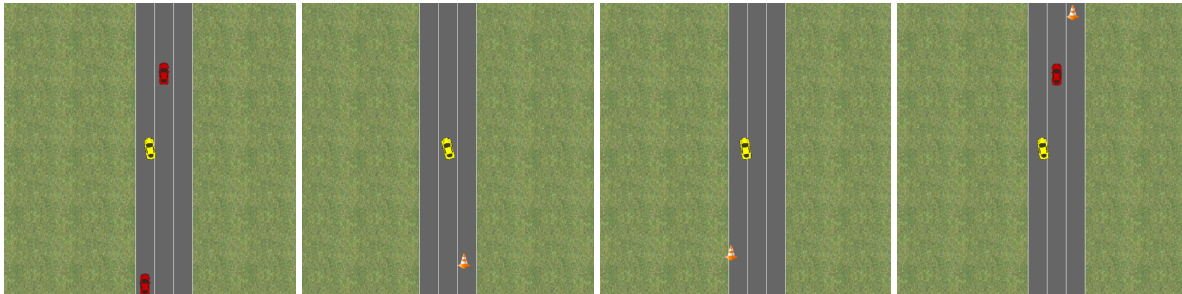
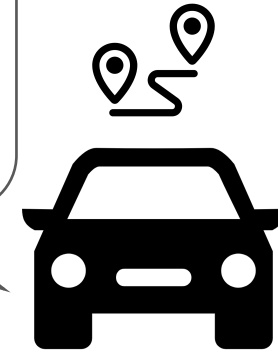
Finding Informative Edge Cases

Current Evidence

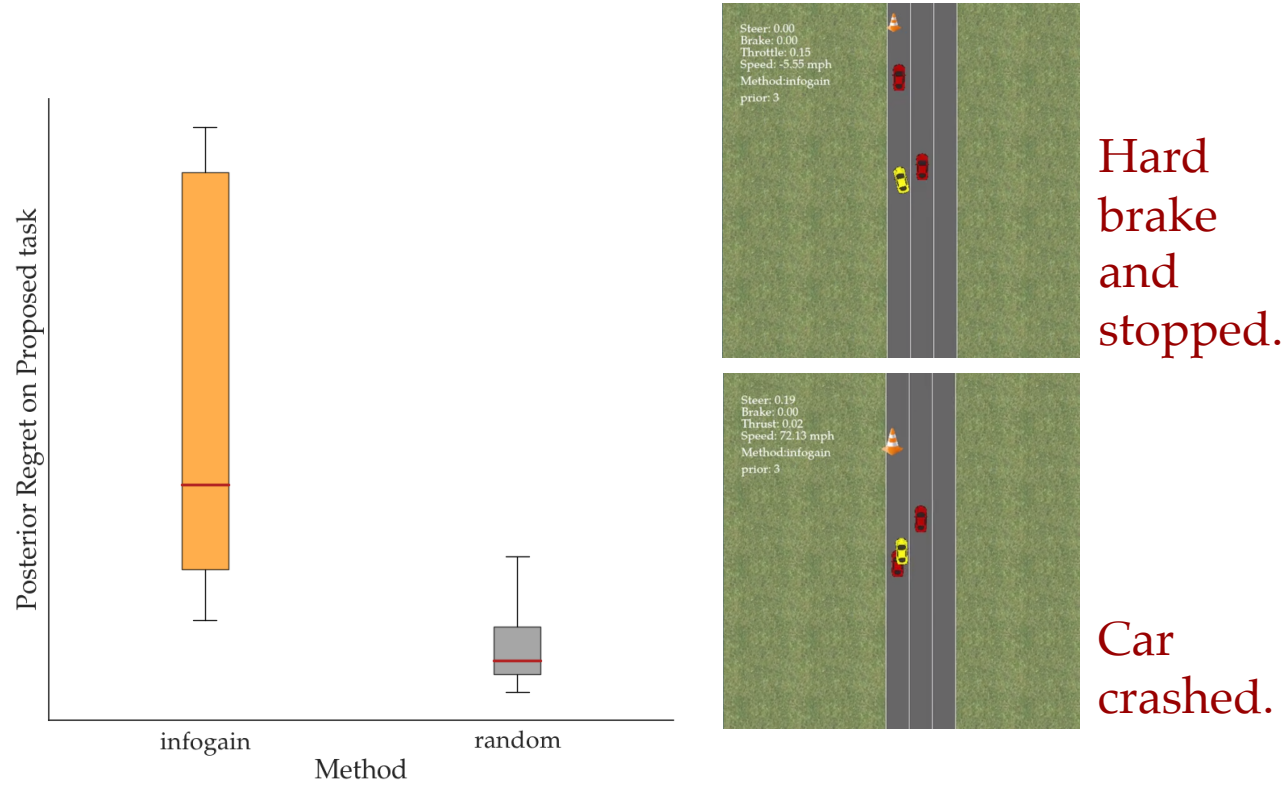
$$R_{\omega}(s, a)_{v1.0}$$



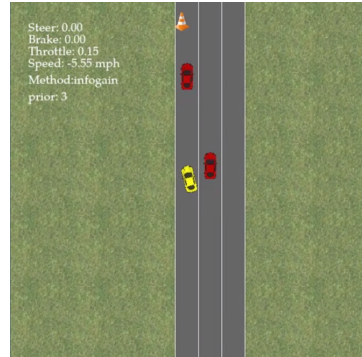
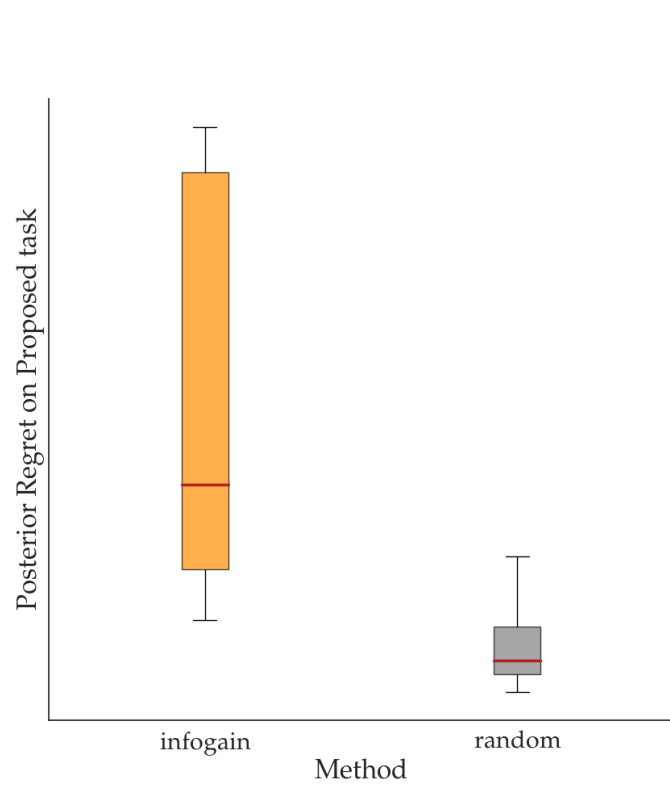
$$\arg \max_m \mathbb{E}_b [\mathbb{H}(b) - \mathbb{H}(b'|m)]$$



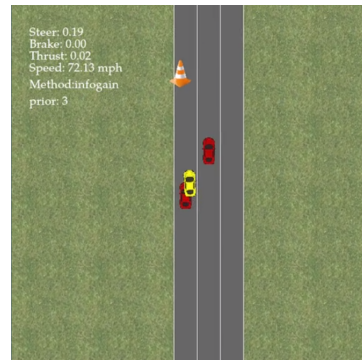
Finding Informative Edge Cases



Finding Informative Edge Cases



Hard
brake
and
stopped.



Car
crashed.

