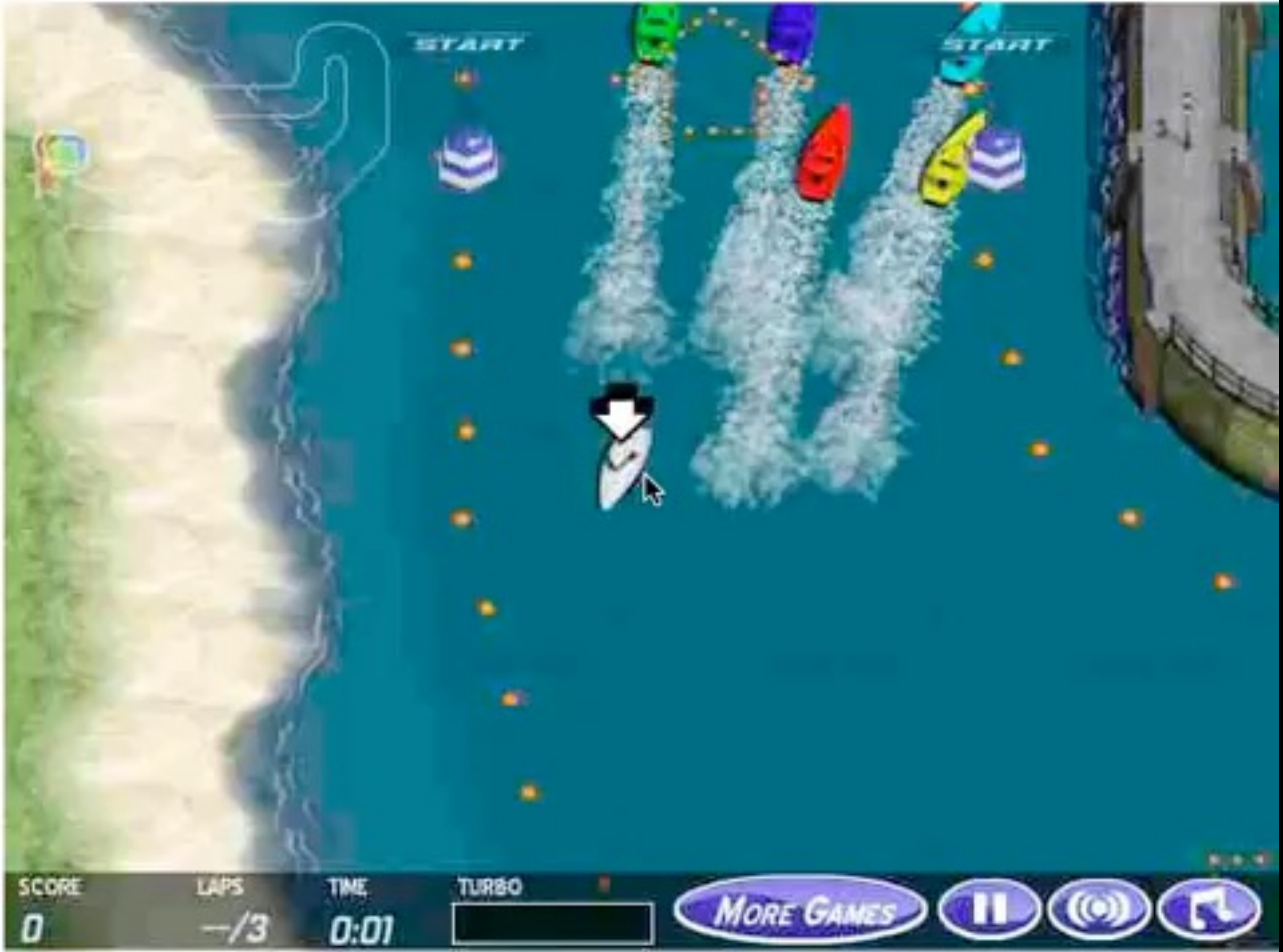
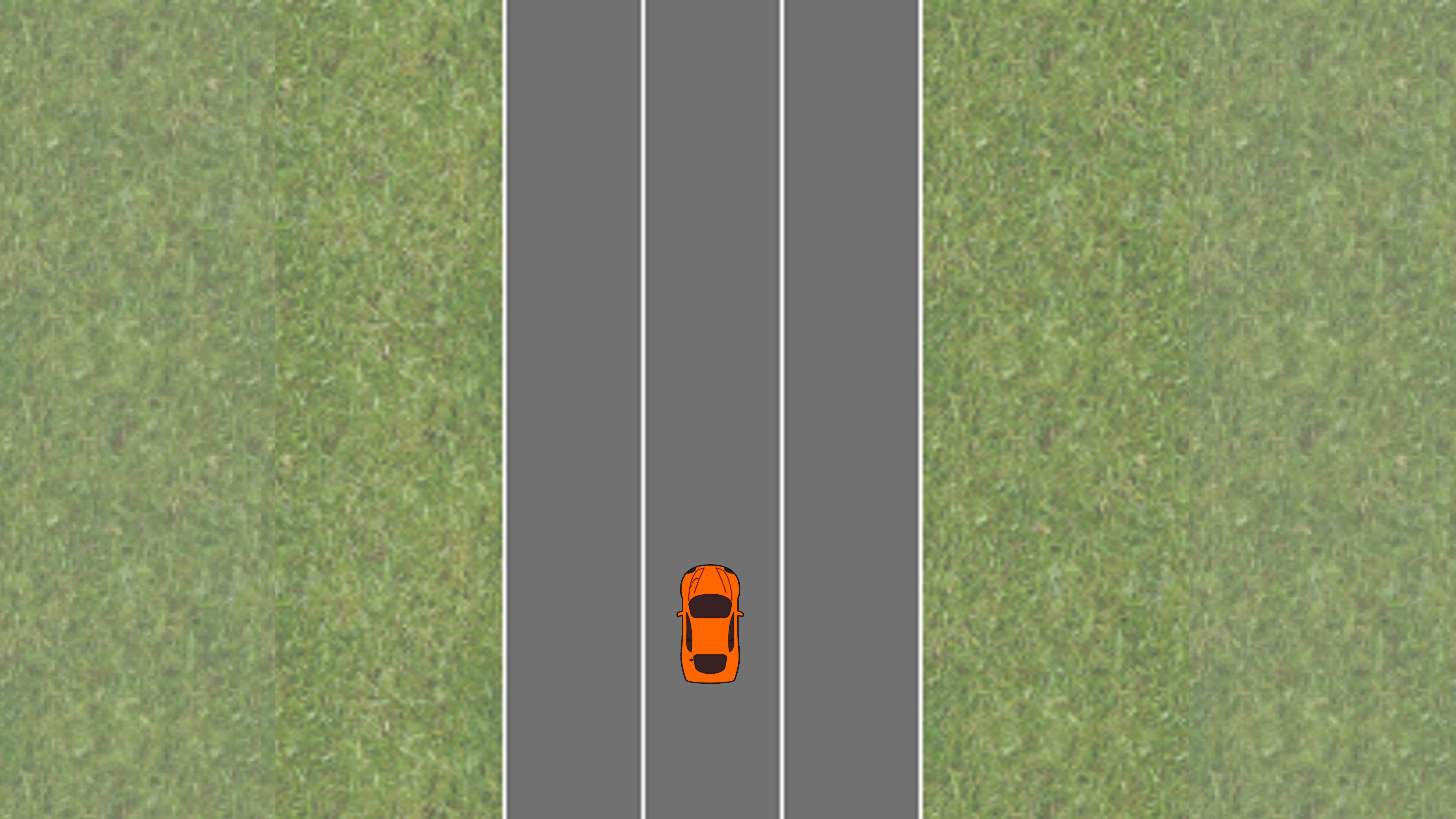
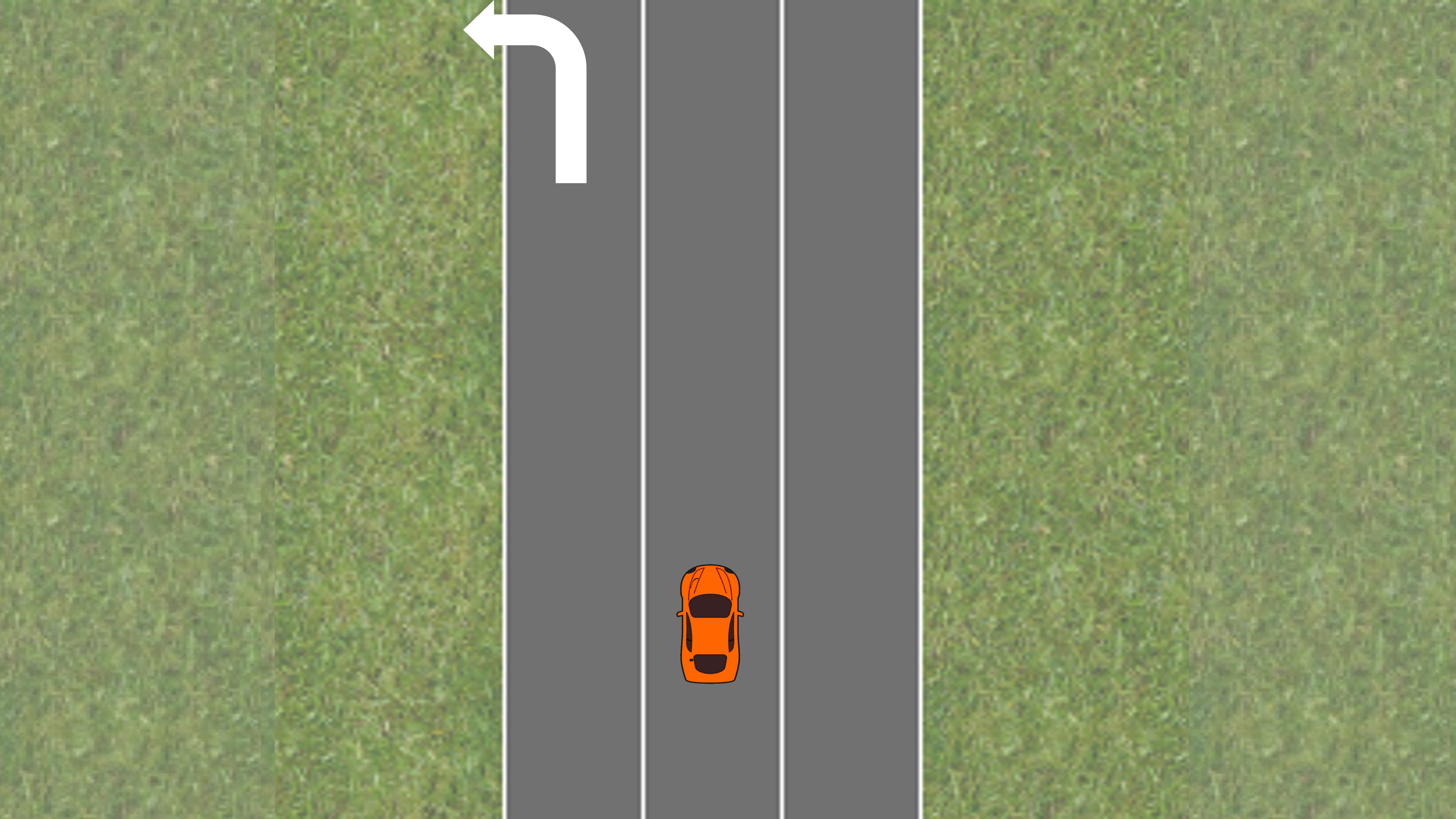
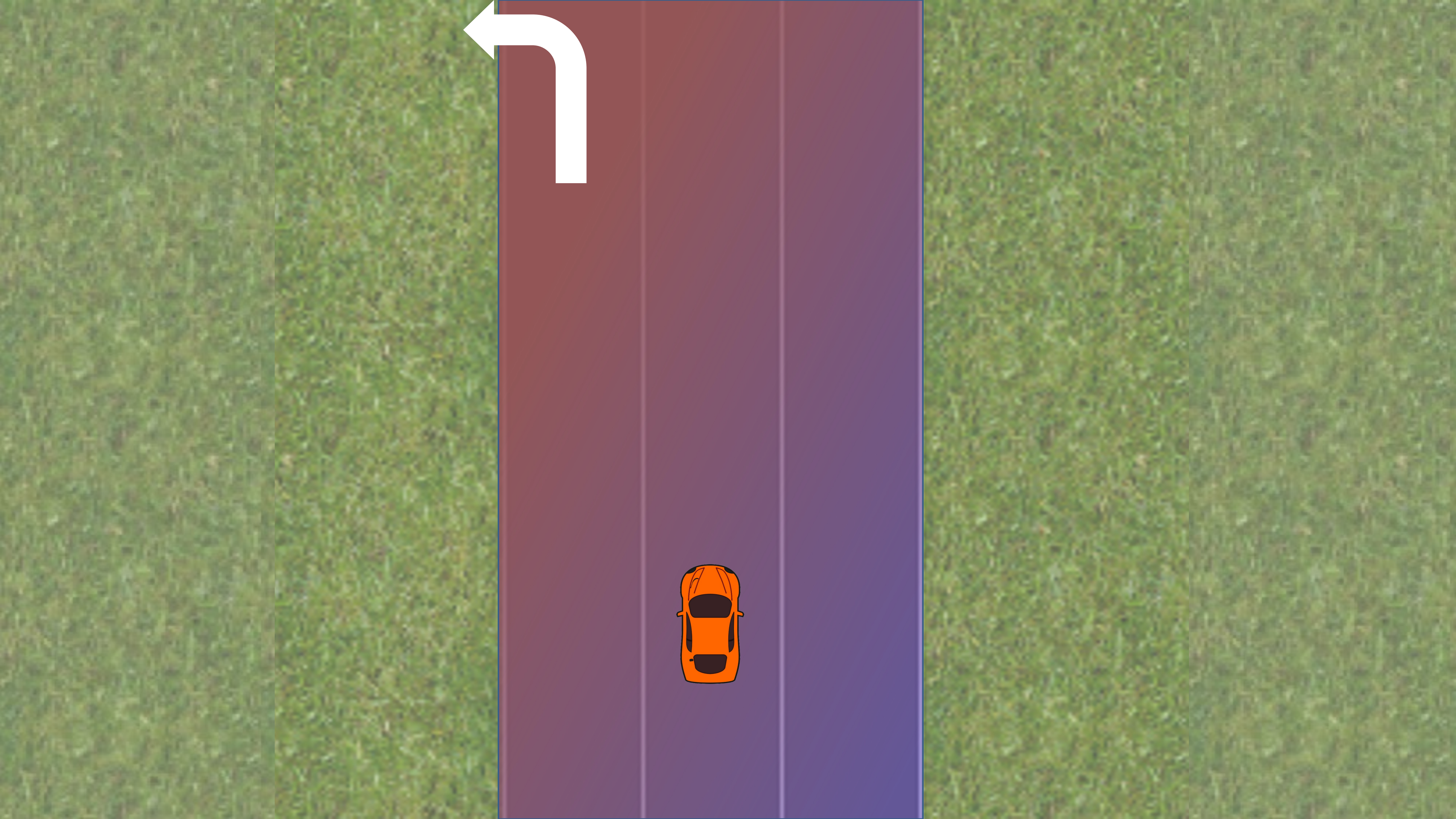
Challenges in aligning AI with human preferences and values

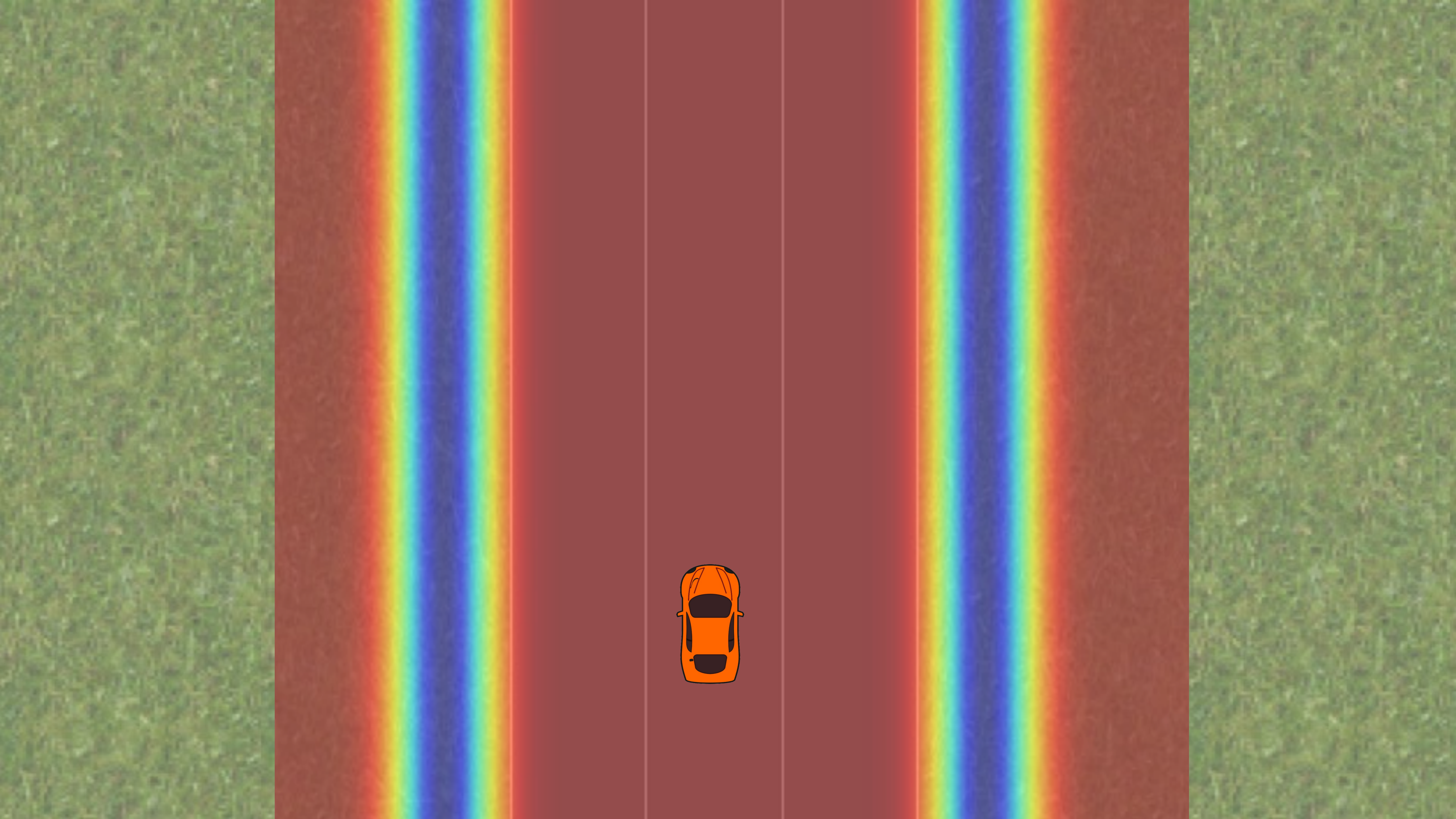
Anca Dragan

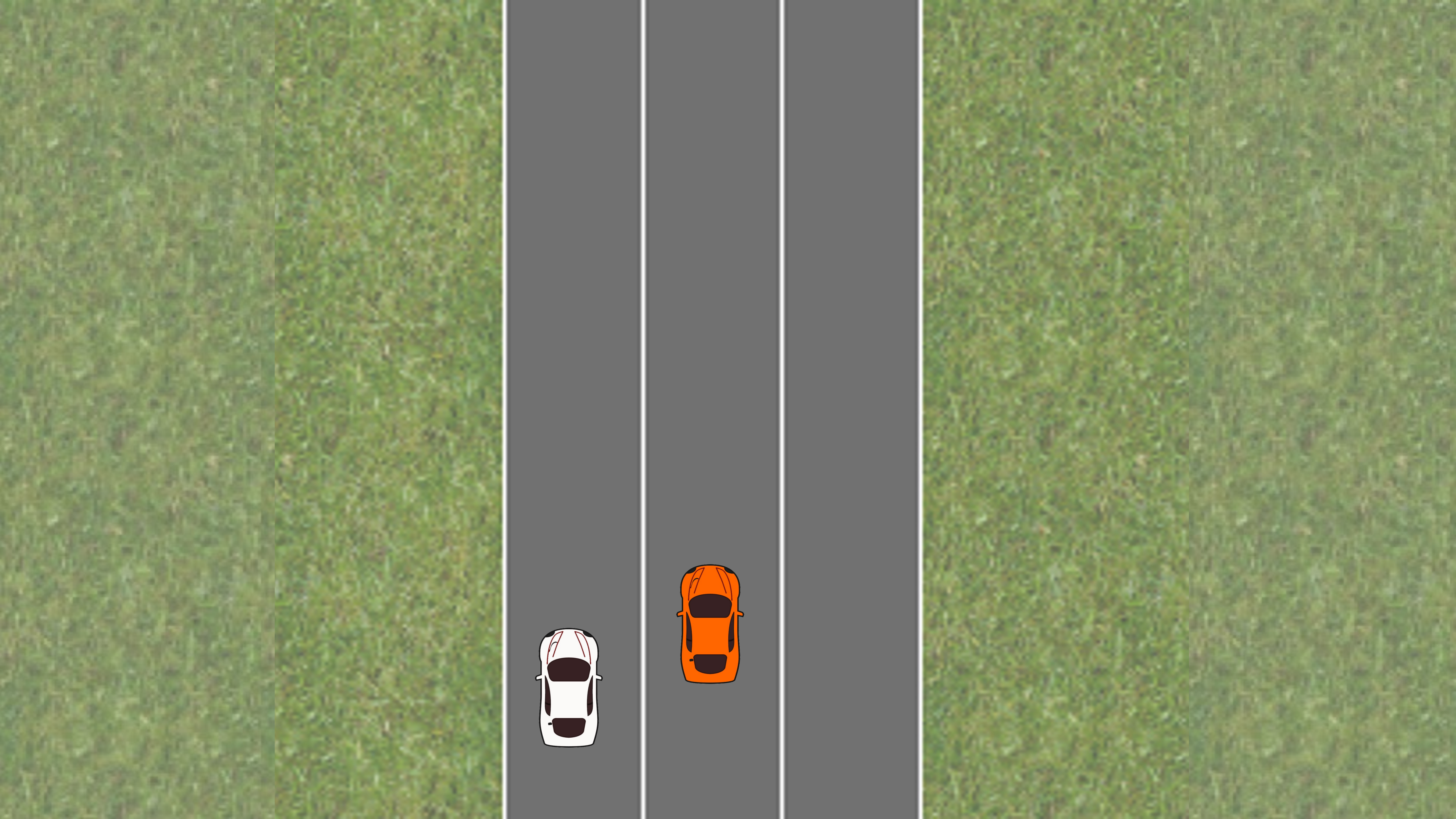

START
START
SCORE
0
LAPS
--/3
TIME
0:01
TURBO
More Games

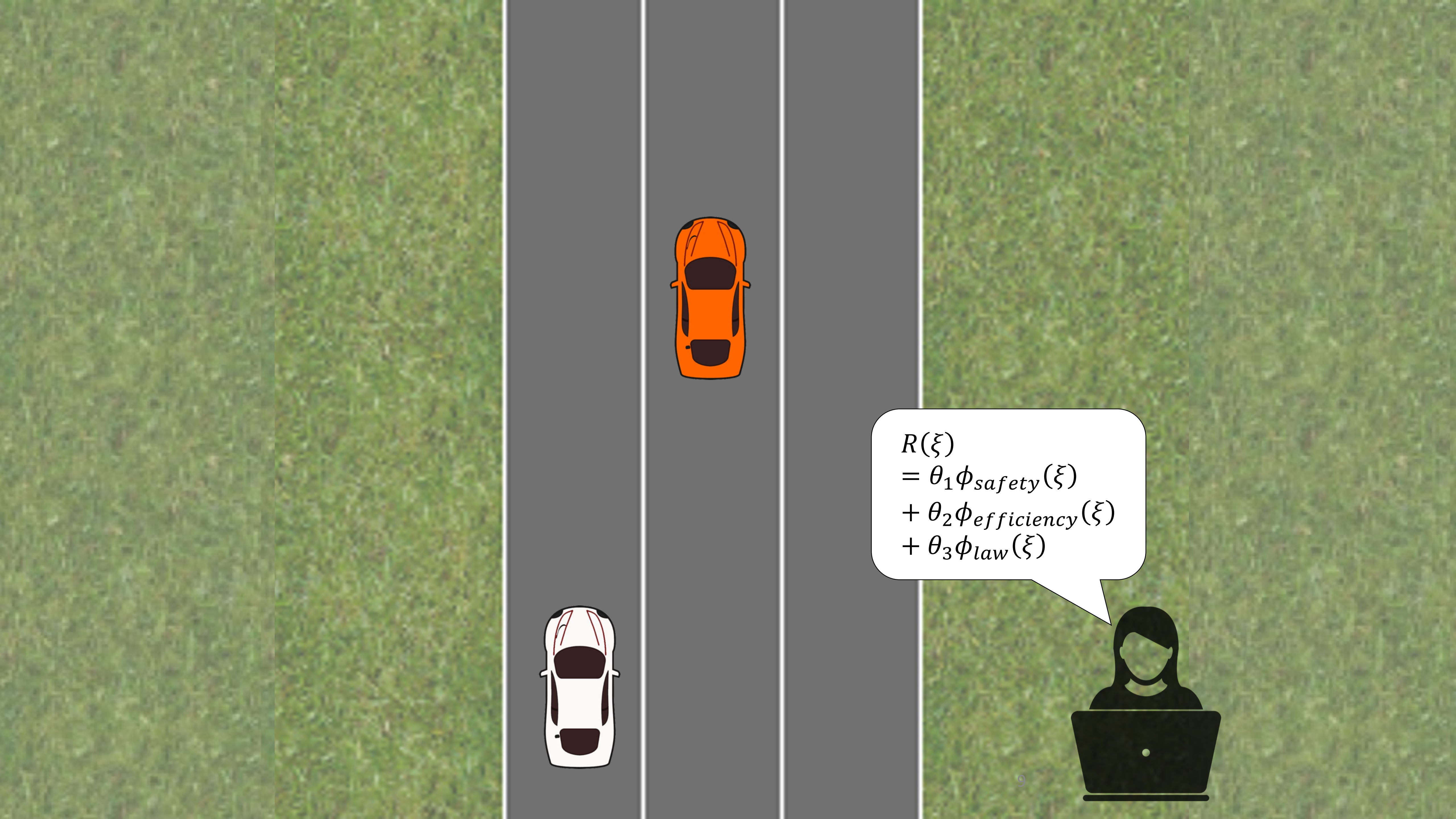




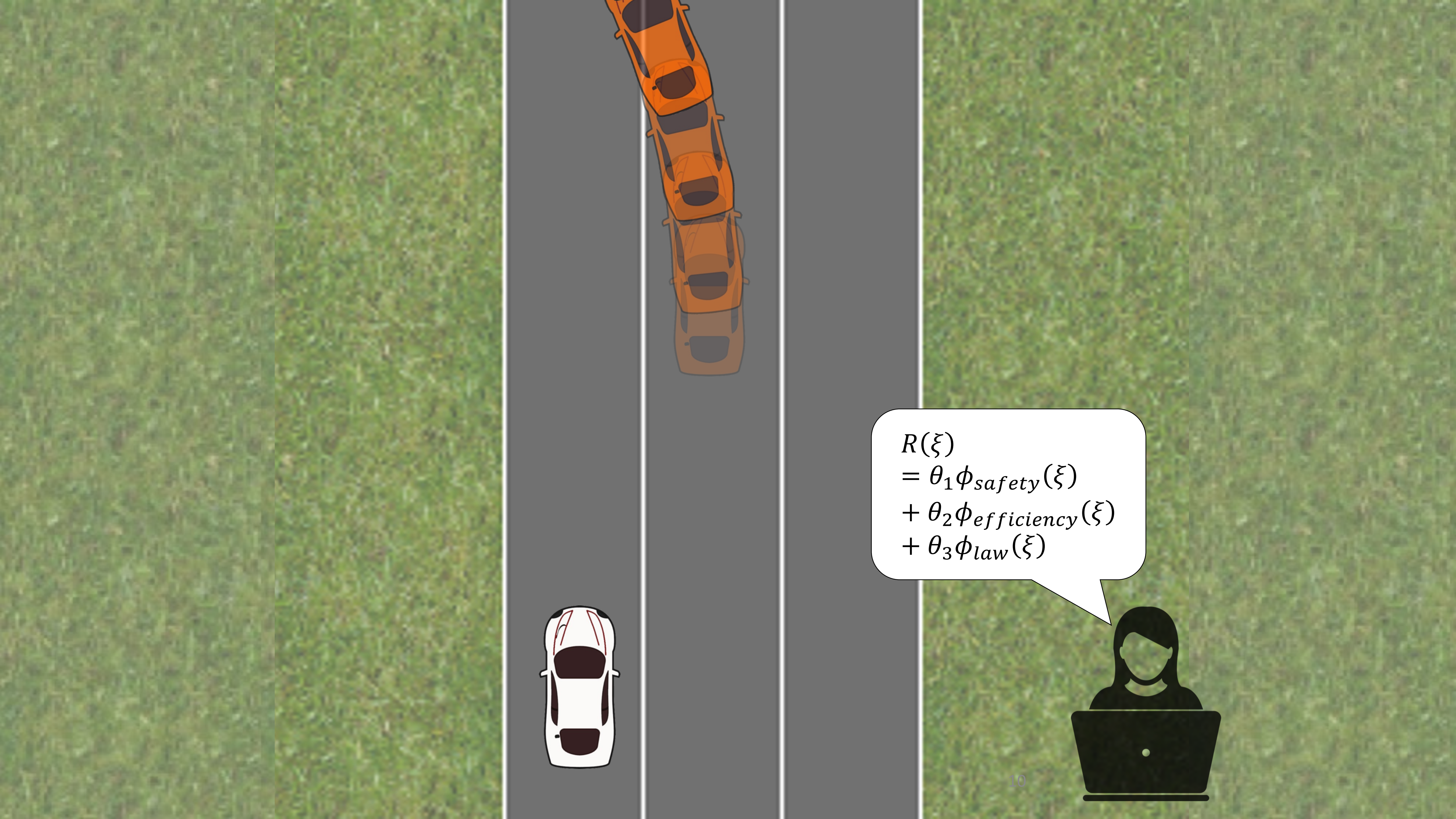




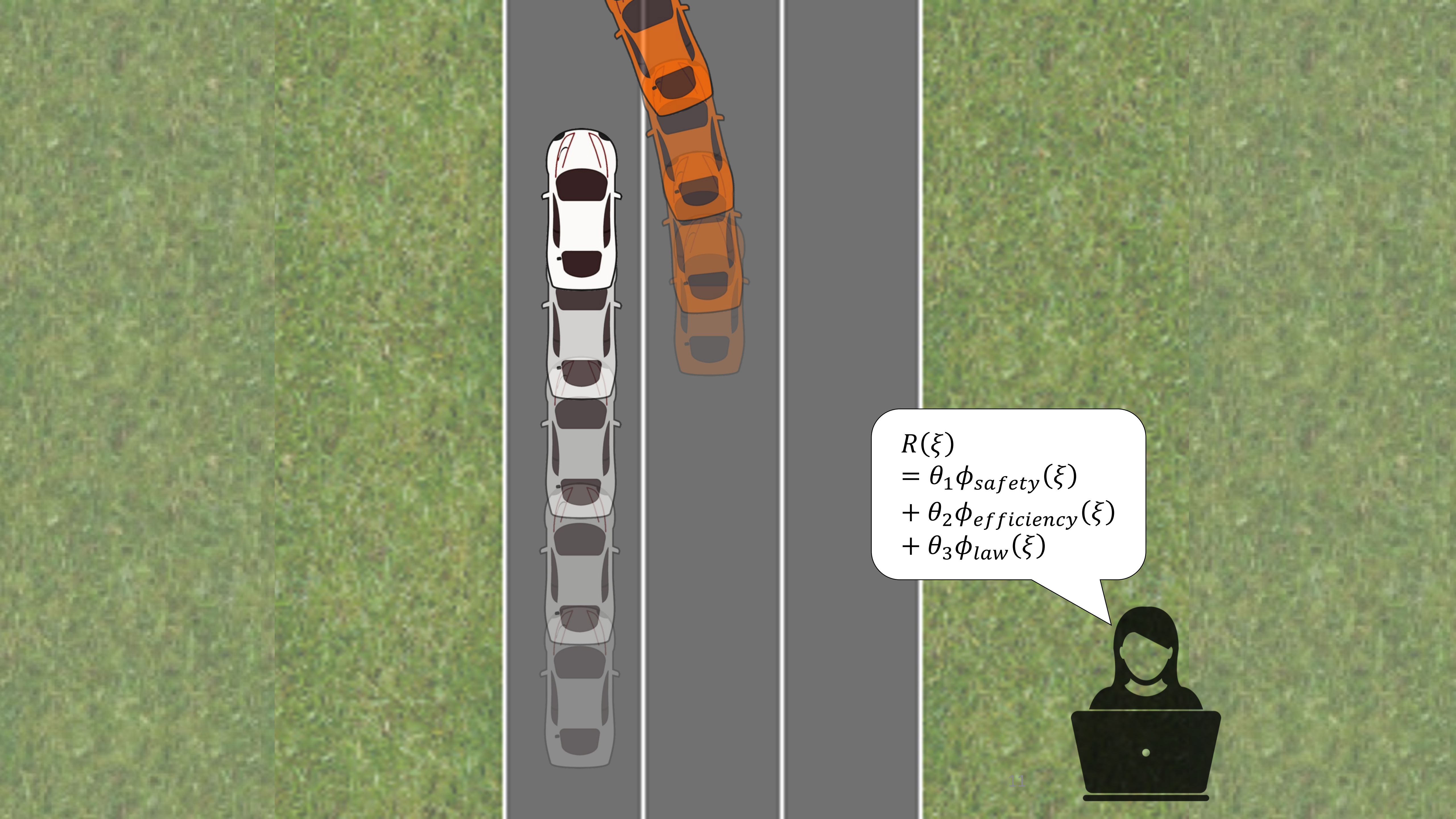




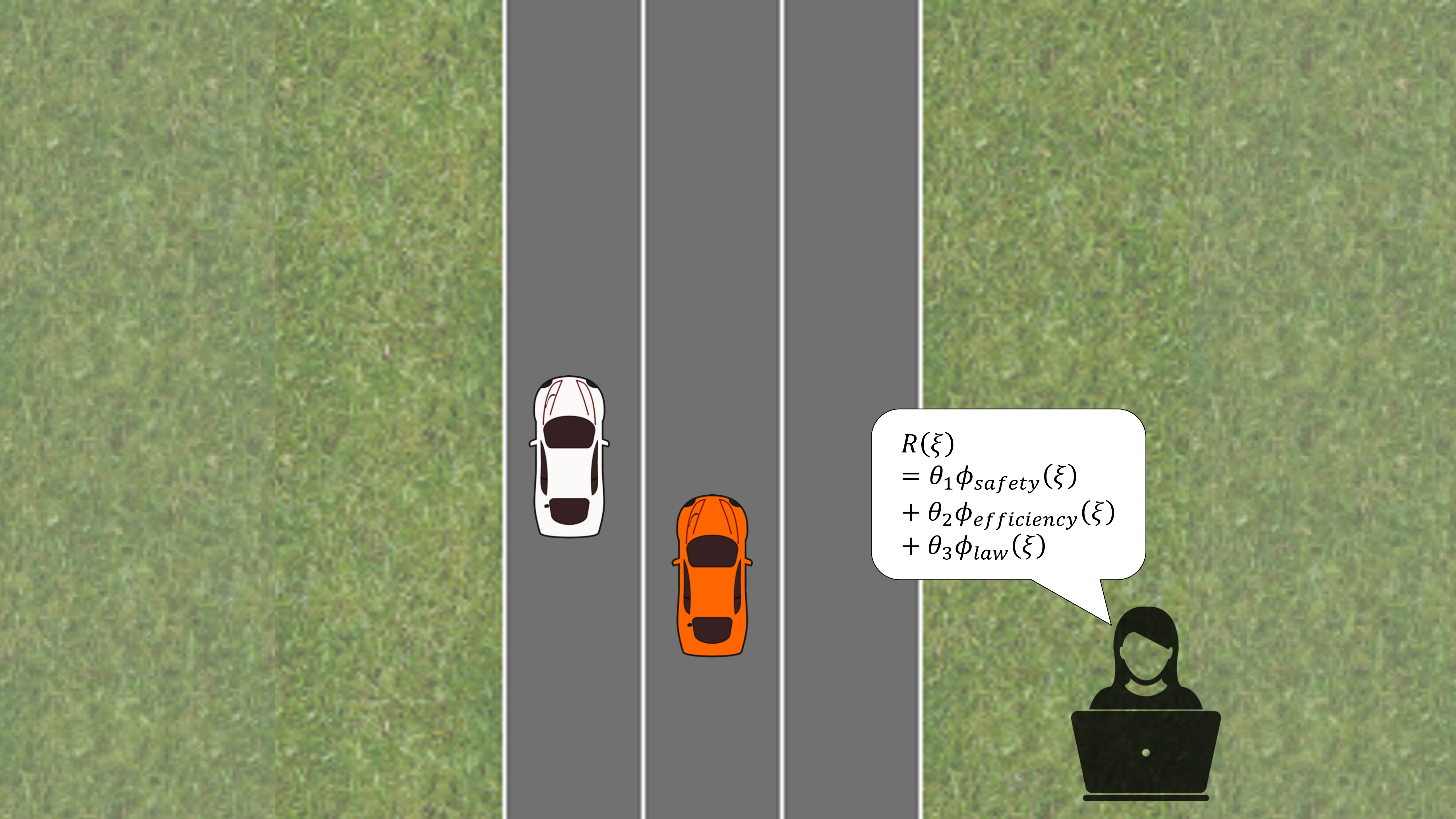
$$\begin{aligned}&R(\xi)\\&= \theta_1\phi_{safety}(\xi)\\&+ \theta_2\phi_{efficiency}(\xi)\\&+ \theta_3\phi_{law}(\xi)\end{aligned}$$



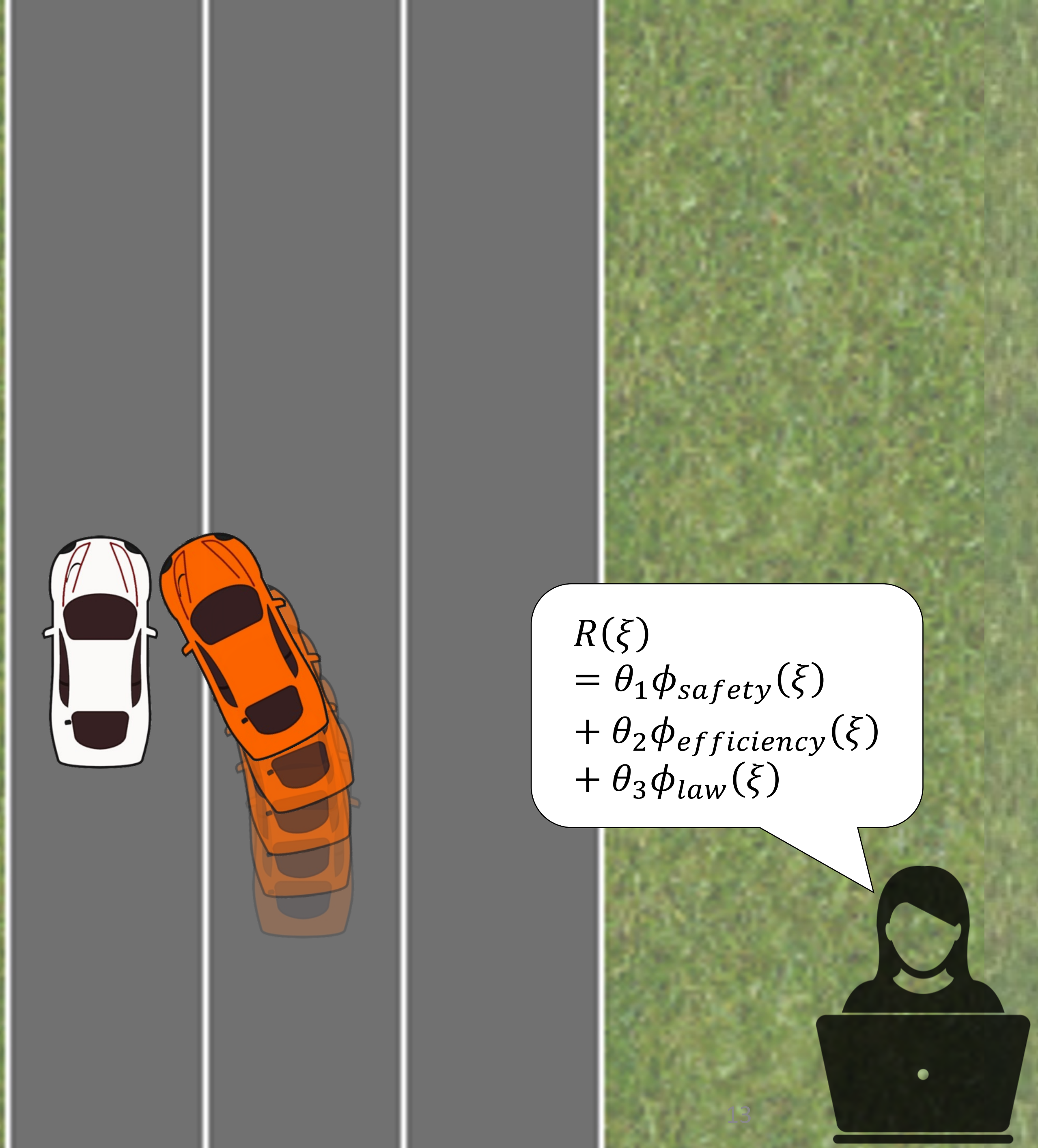
$$
\begin{aligned}
&R(\xi) \\
&= \theta_1 \phi_{safety}(\xi) \\
&+ \theta_2 \phi_{efficiency}(\xi) \\
&+ \theta_3 \phi_{law}(\xi)
\end{aligned}
$$



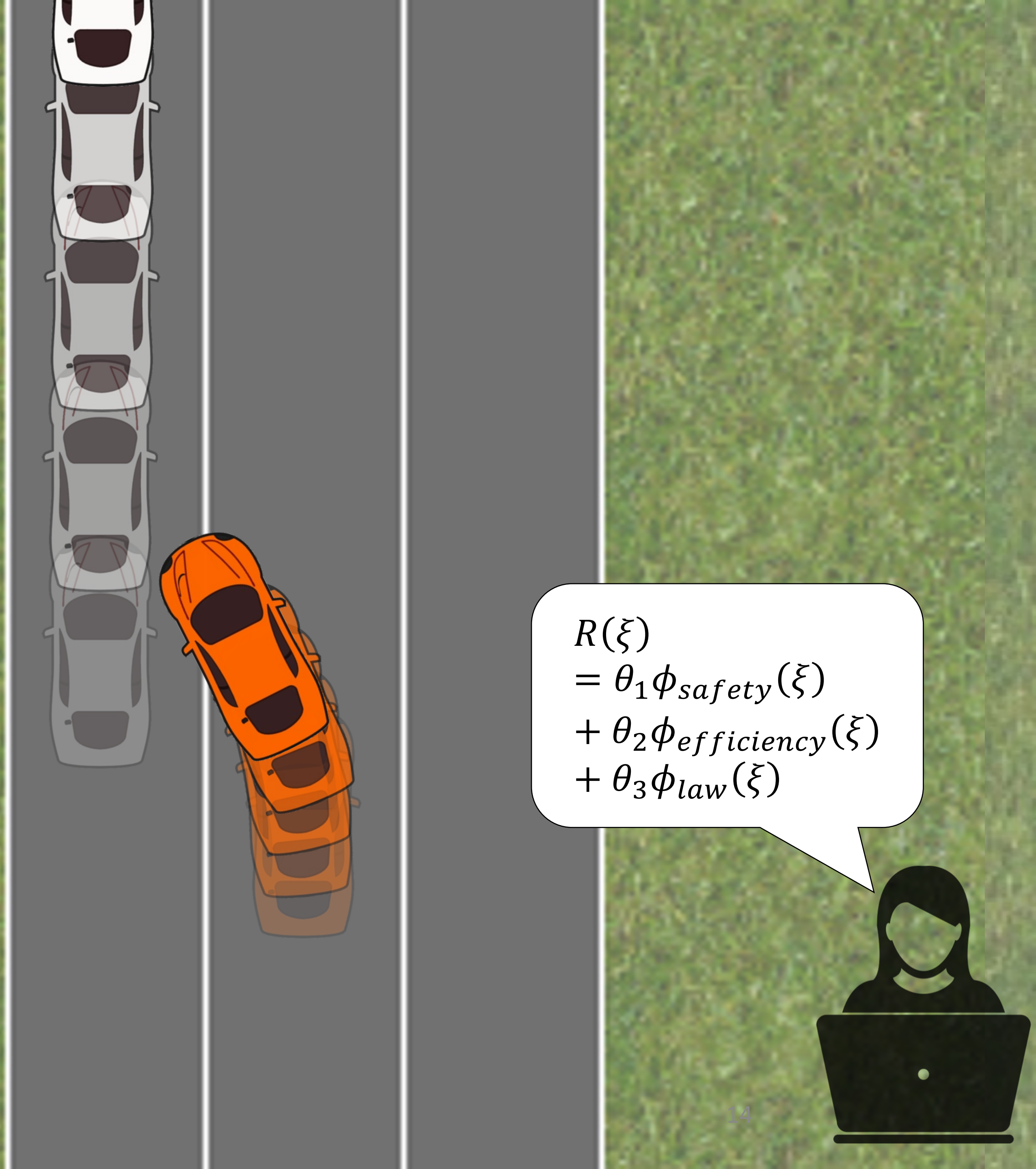
$$R(\xi) = \theta_1\phi_{safety}(\xi) + \theta_2\phi_{efficiency}(\xi) + \theta_3\phi_{law}(\xi)$$



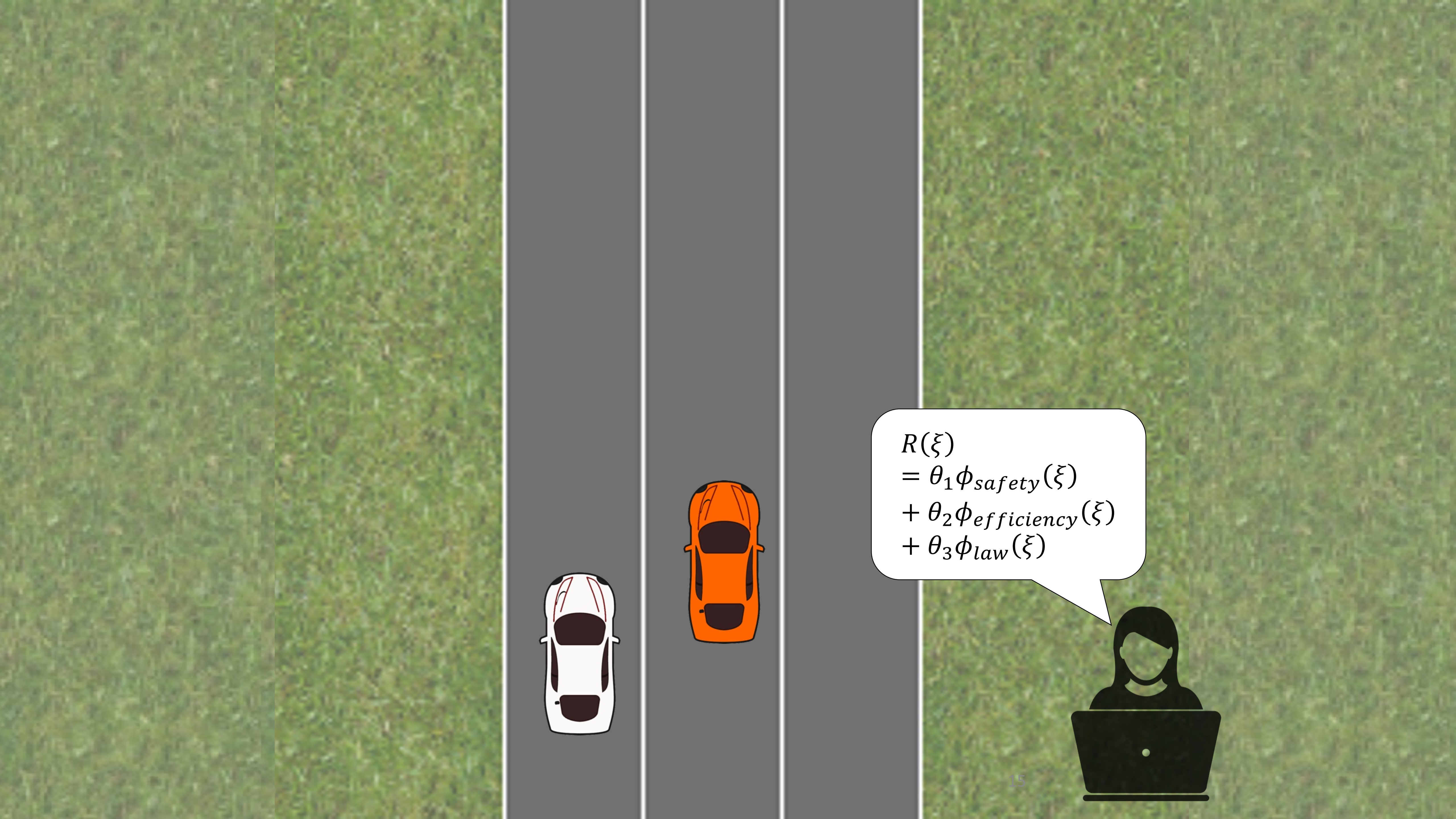
$$R(\xi) \\ = \theta_1 \phi_{safety}(\xi) \\ + \theta_2 \phi_{efficiency}(\xi) \\ + \theta_3 \phi_{law}(\xi)$$



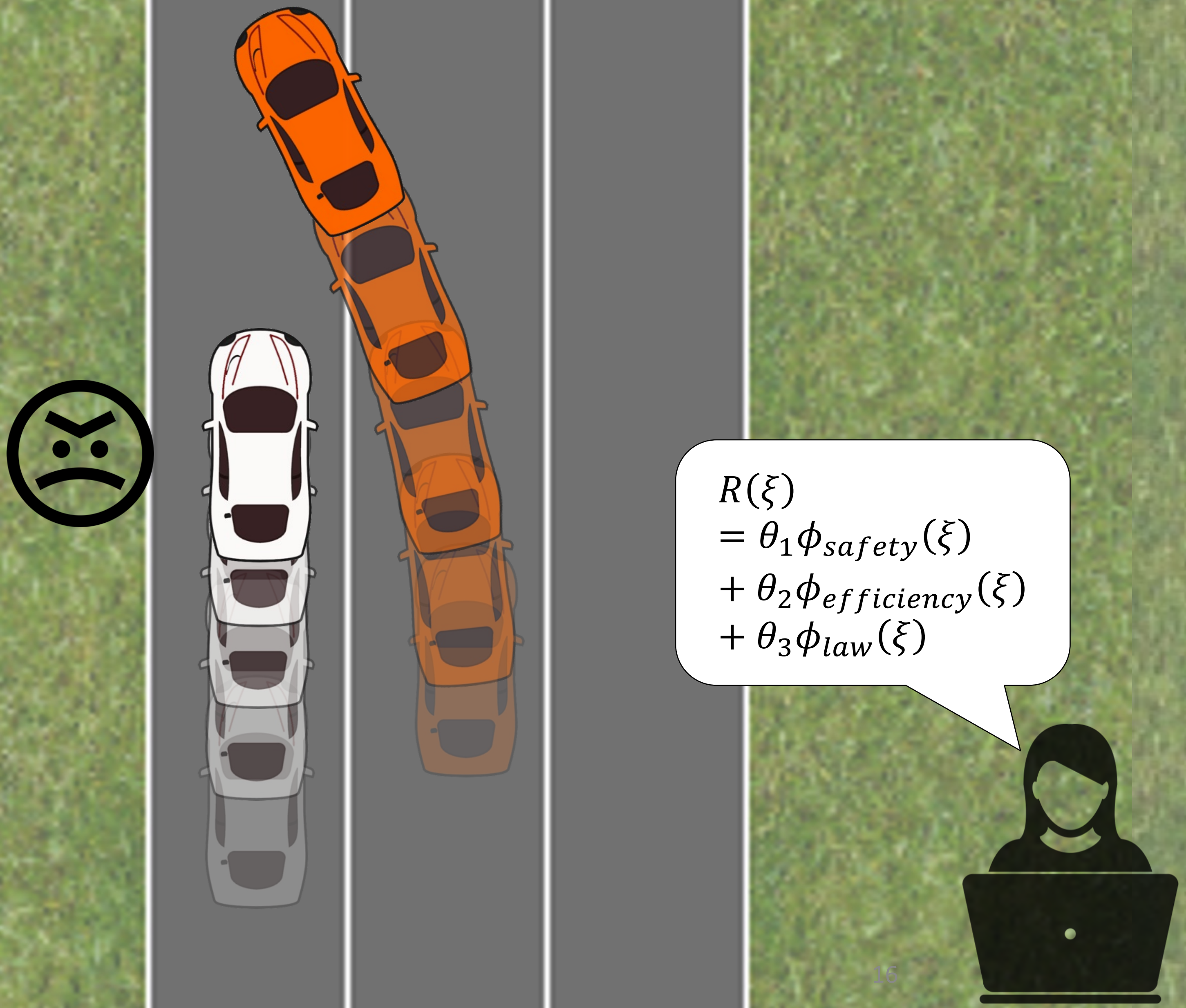
$$
\begin{aligned}
&R(\xi)\\
&= \theta_1 \phi_{safety}(\xi)\\
&+ \theta_2 \phi_{efficiency}(\xi)\\
&+ \theta_3 \phi_{law}(\xi)
\end{aligned}
$$



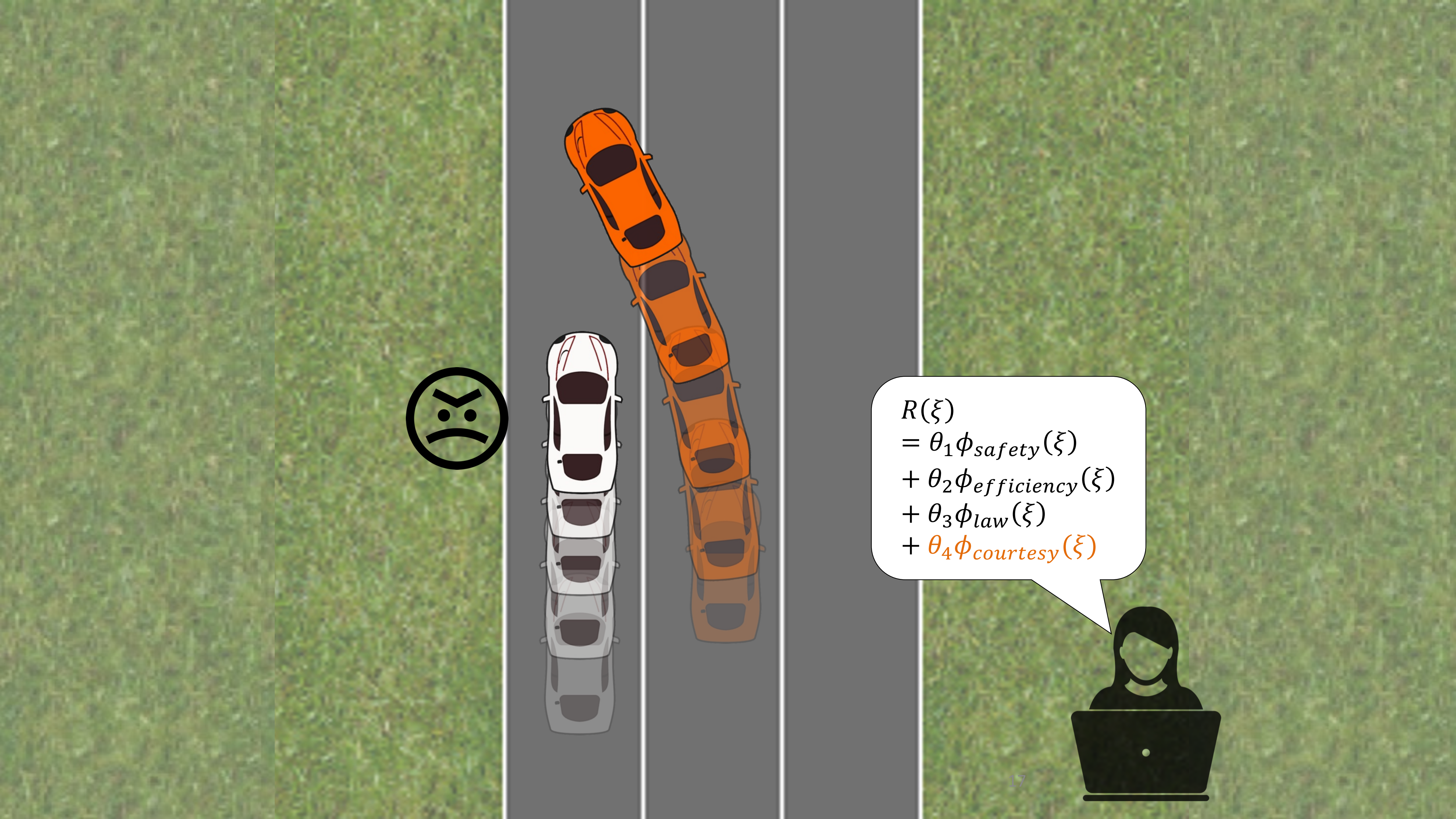
$$
\begin{aligned}
&R(\xi) \\
&= \theta_1 \phi_{safety}(\xi) \\
&+ \theta_2 \phi_{efficiency}(\xi) \\
&+ \theta_3 \phi_{law}(\xi)
\end{aligned}
$$



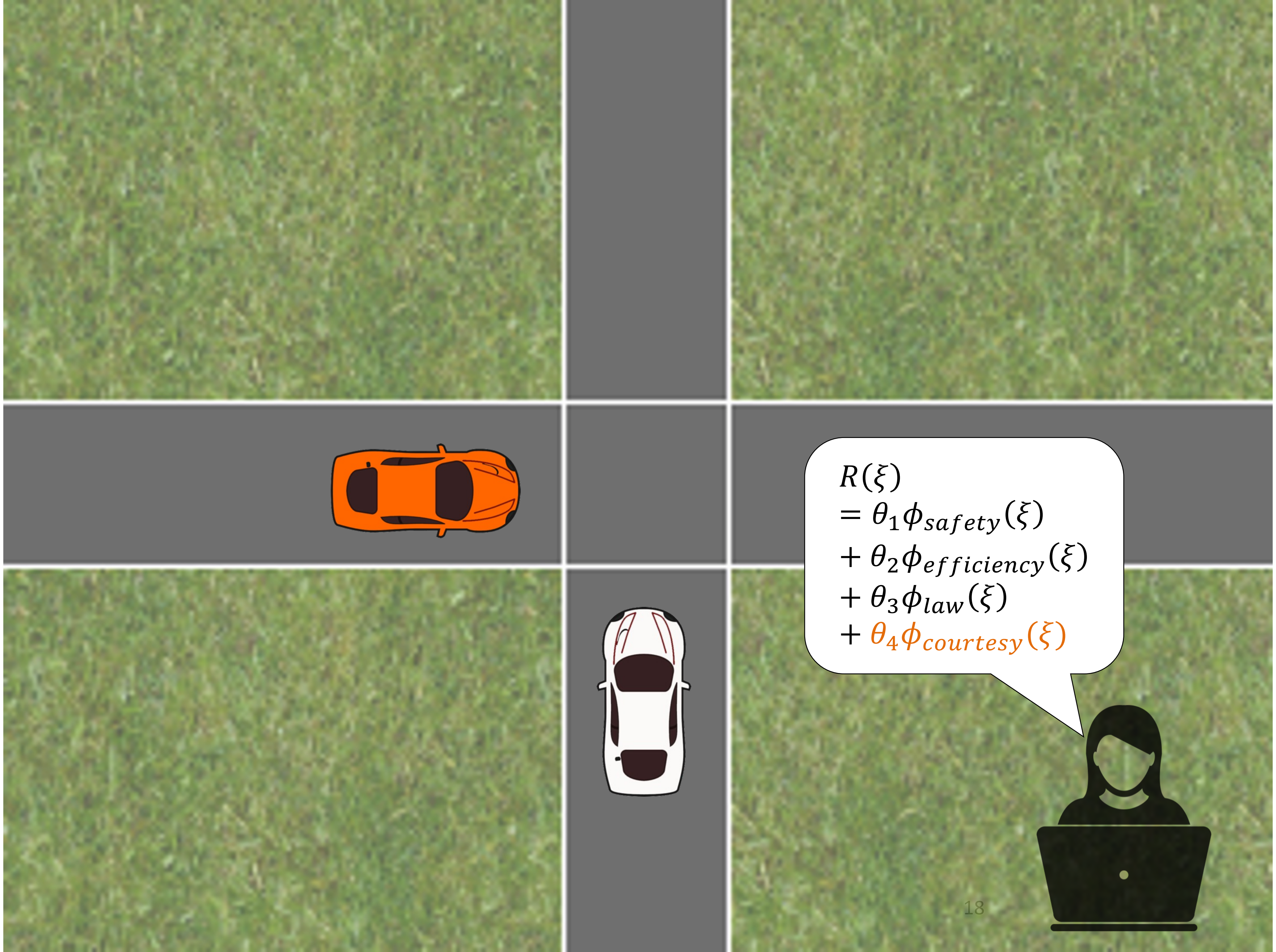
$$R(\xi) = \theta_1\phi_{safety}(\xi) + \theta_2\phi_{efficiency}(\xi) + \theta_3\phi_{law}(\xi)$$



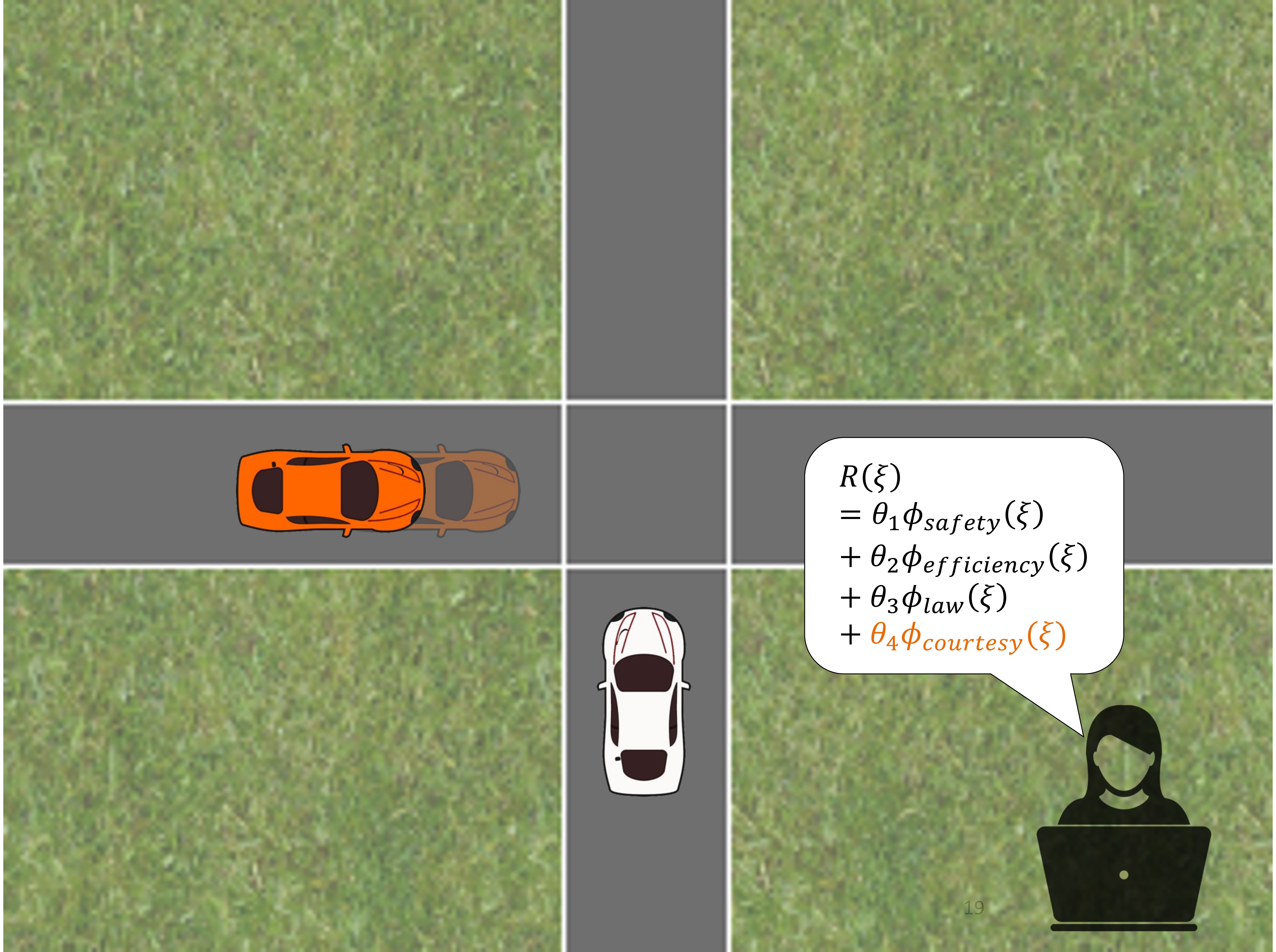
$$R(\xi) = \theta_1 \phi_{safety}(\xi) + \theta_2 \phi_{efficiency}(\xi) + \theta_3 \phi_{law}(\xi)$$



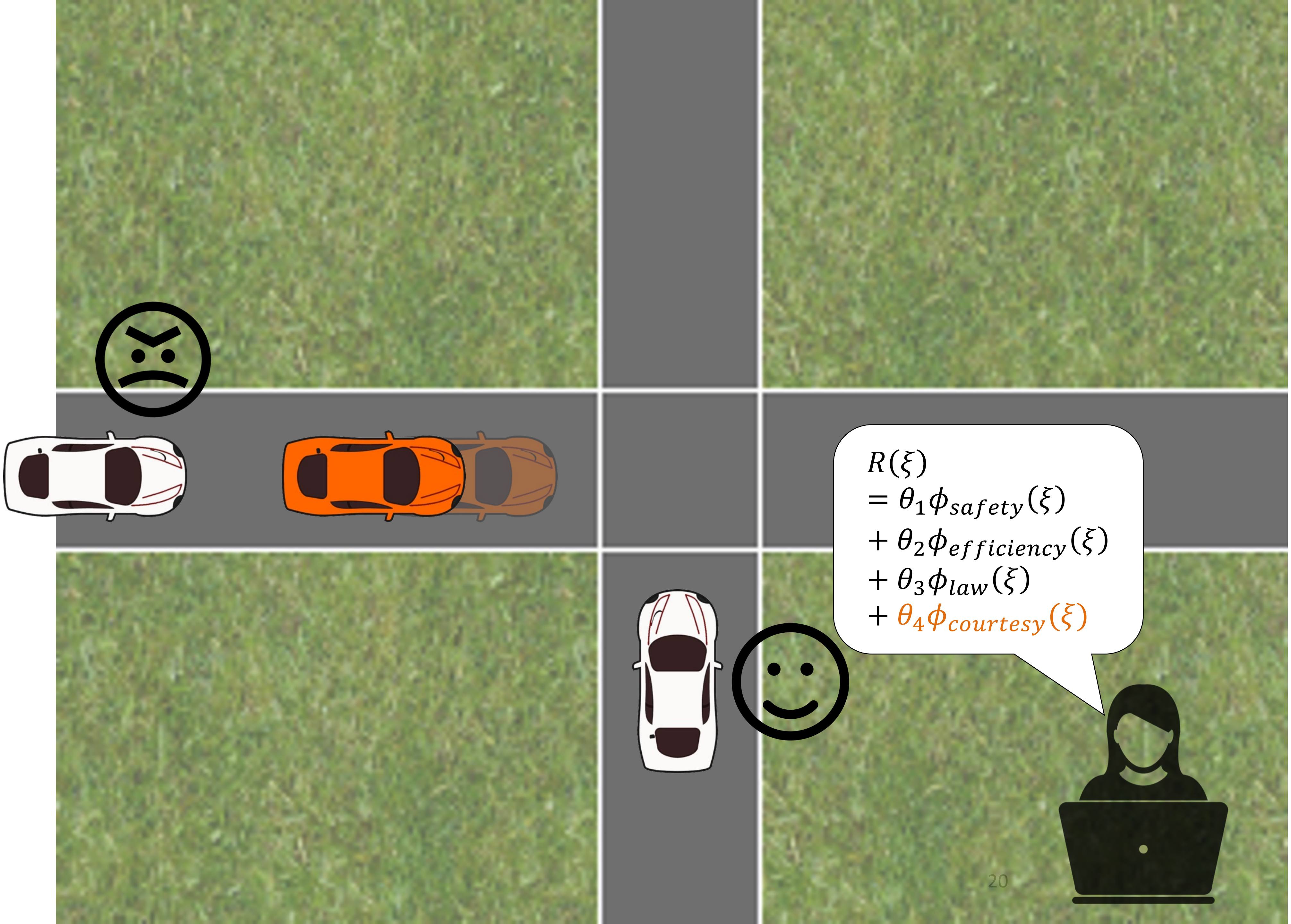
$$\begin{aligned} &R(\xi) \\ &= \theta_1\phi_{safety}(\xi) \\ &+ \theta_2\phi_{efficiency}(\xi) \\ &+ \theta_3\phi_{law}(\xi) \\ &+ \theta_4\phi_{courtesy}(\xi) \end{aligned}$$



$$\begin{aligned} &R(\xi) \\ &= \theta_1 \phi_{safety}(\xi) \\ &+ \theta_2 \phi_{efficiency}(\xi) \\ &+ \theta_3 \phi_{law}(\xi) \\ &+ \theta_4 \phi_{courtesy}(\xi) \end{aligned}$$

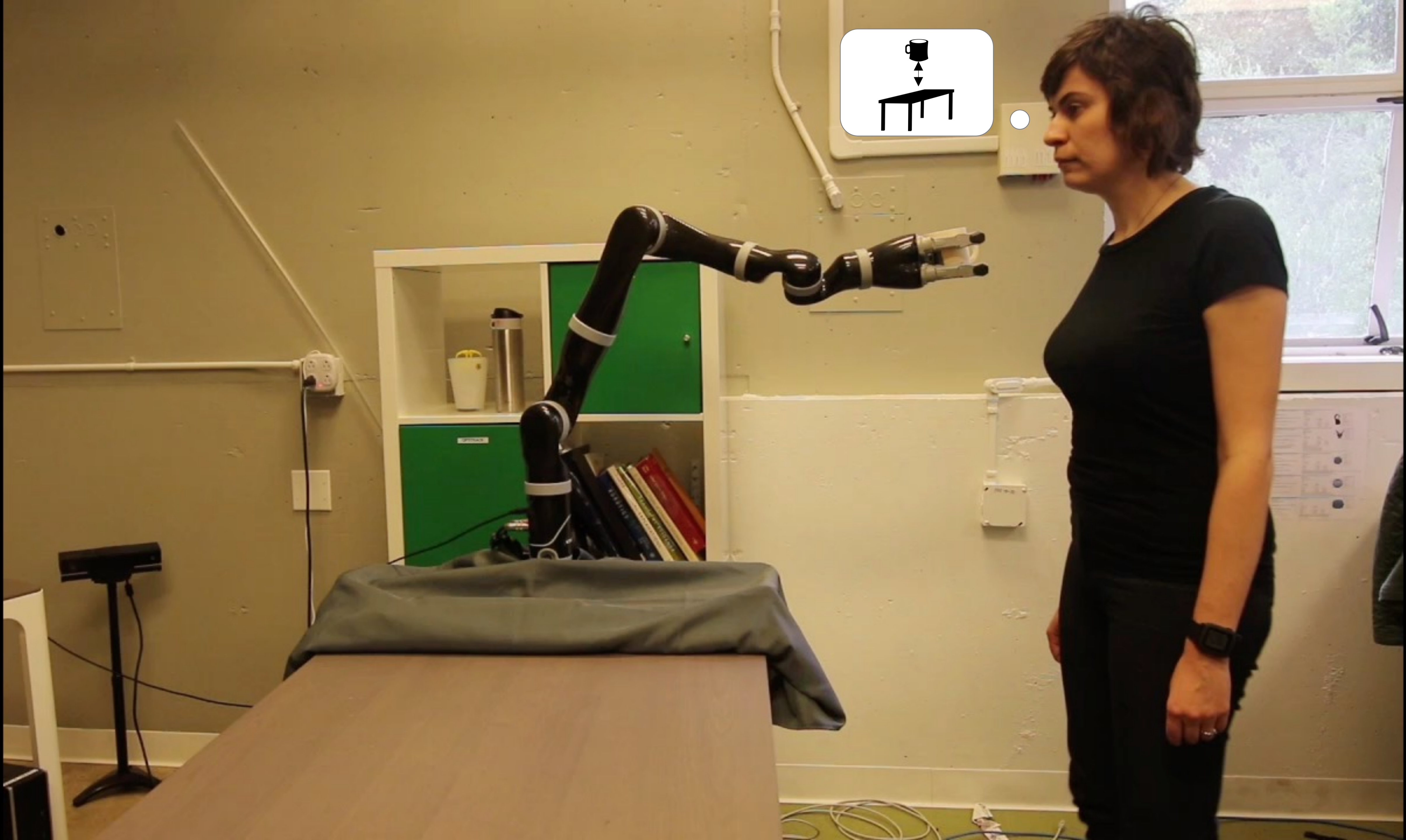


$$
\begin{aligned}
&R(\xi) \\
&= \theta_1 \phi_{safety}(\xi) \\
&+ \theta_2 \phi_{efficiency}(\xi) \\
&+ \theta_3 \phi_{law}(\xi) \\
&+ \theta_4 \phi_{courtesy}(\xi)
\end{aligned}
$$



$$
\begin{aligned}
& R(\xi) \\
& = \theta_1 \phi_{safety}(\xi) \\
& + \theta_2 \phi_{efficiency}(\xi) \\
& + \theta_3 \phi_{law}(\xi) \\
& + \theta_4 \phi_{courtesy}(\xi)
\end{aligned}
$$

$$\begin{aligned} &R(\xi) \\ &= \theta_1 \phi_{safety}(\xi) \\ &+ \theta_2 \phi_{efficiency}(\xi) \\ &+ \theta_3 \phi_{law}(\xi) \\ &+ \theta_4 \phi_{courtesy}(\xi) \\ &+ \theta_5 \phi_{comfort}(\xi) \end{aligned}$$

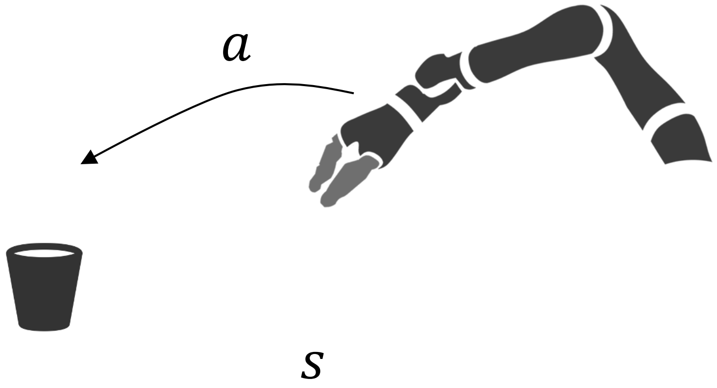


It’s genuinely difficult
to think of every feature ahead of time, to define it in code, and to specify how all the features combine into a reward function.

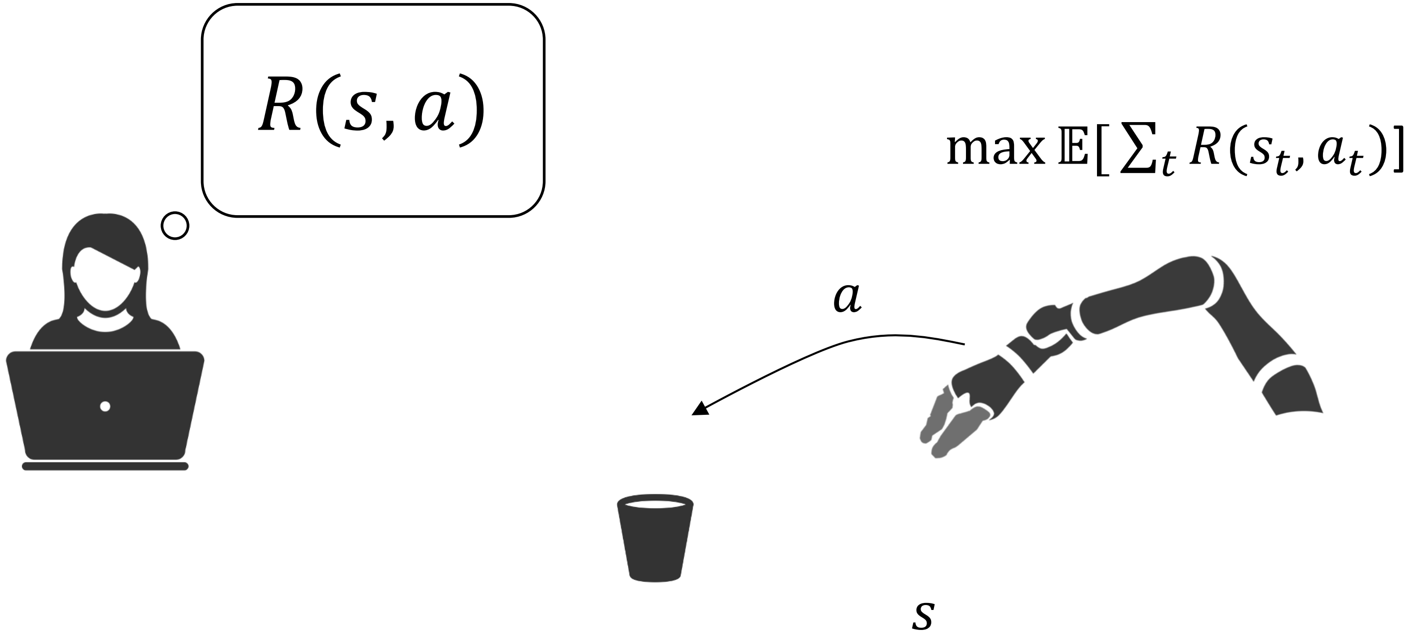
What we pretend AI is:

$$R(s, a)$$

$$\max \mathbb{E}[\textstyle\sum_t R(s_t, a_t)]$$



What AI actually is:



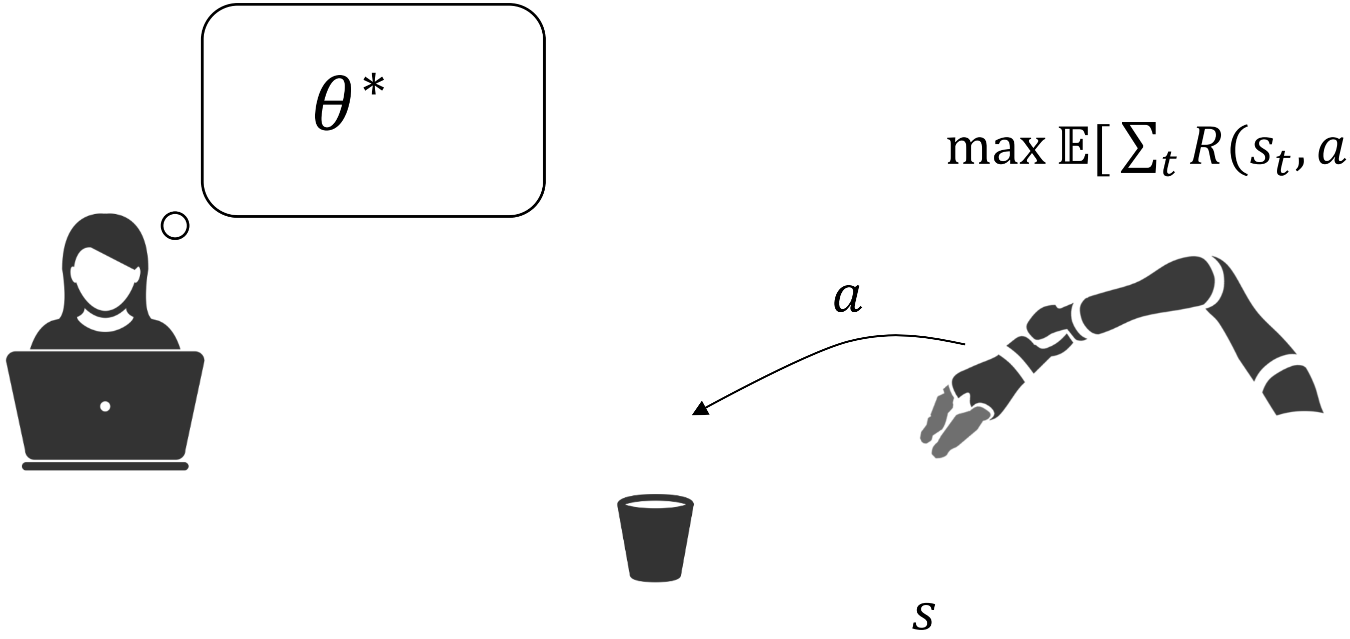
$R(s, a)$

$\max \mathbb{E}[\sum_t R(s_t, a_t)]$

a

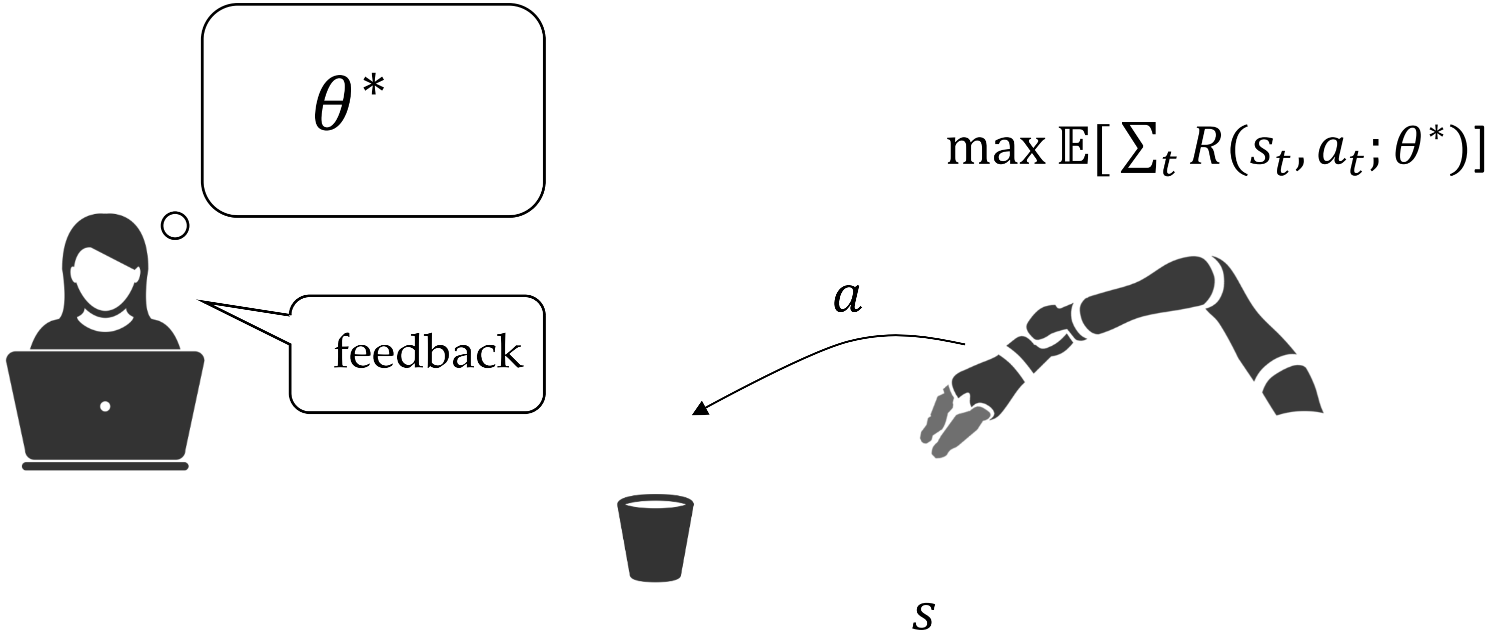
s

Optimize <u>intended</u> reward



$$\max \mathbb{E}[\textstyle\sum_t R(s_t, a_t; \theta^*)]$$

Optimize intended reward



θ^*

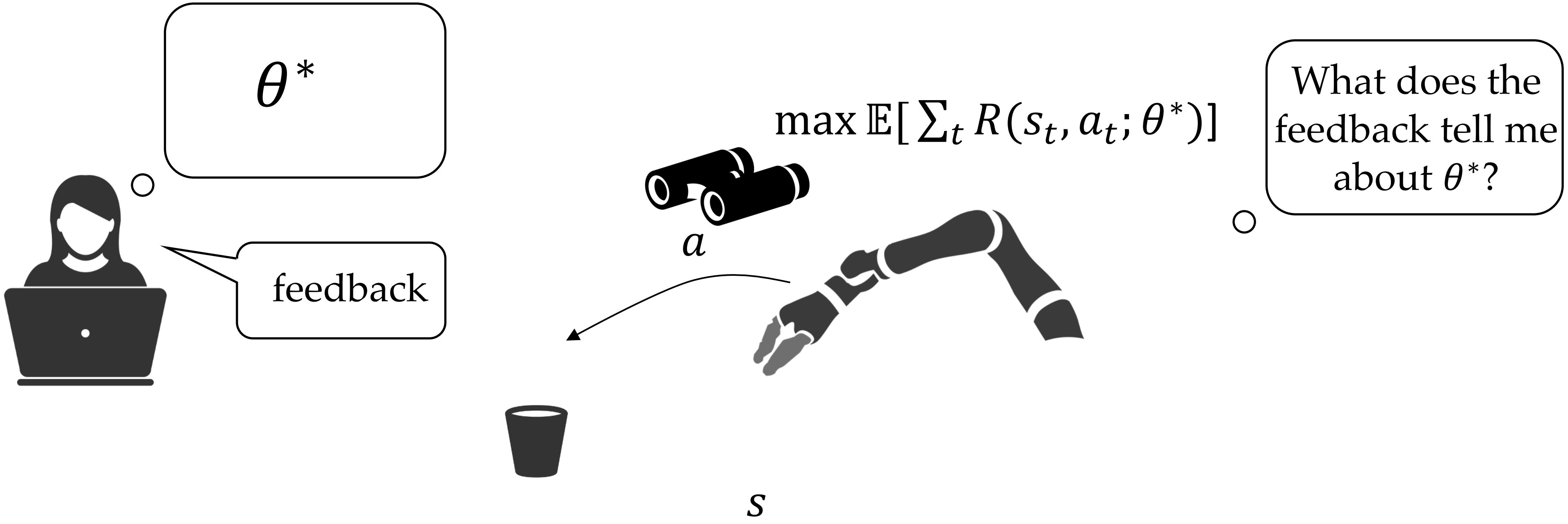
feedback

$\max \mathbb{E}[\sum_t R(s_t, a_t; \theta^*)]$

a

s

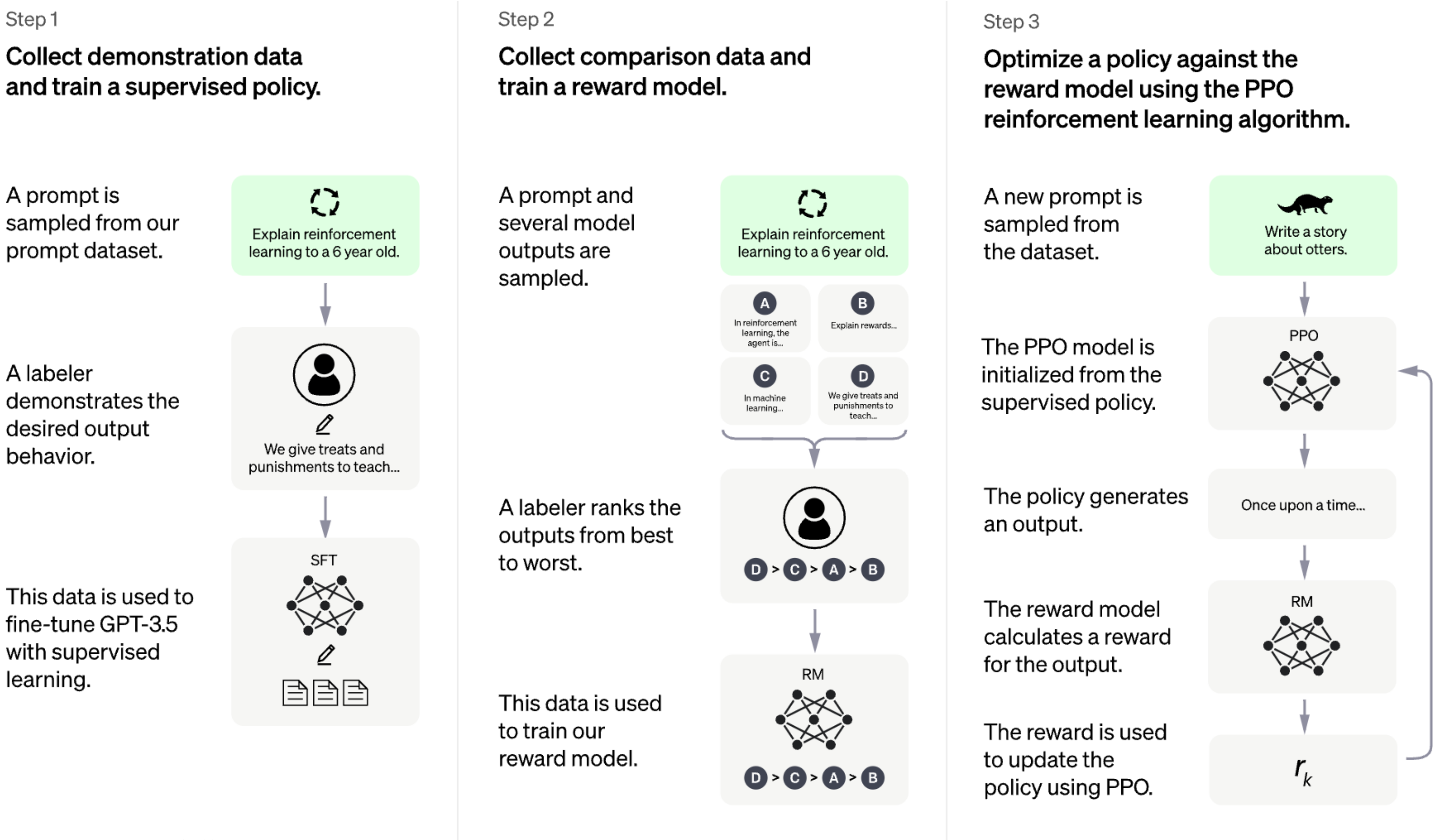
Optimize intended reward



$\max \mathbb{E}[\sum_t R(s_t, a_t; \theta^*)]$

What does the feedback tell me about θ^*?

How are LLMs currently “aligned”?



Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

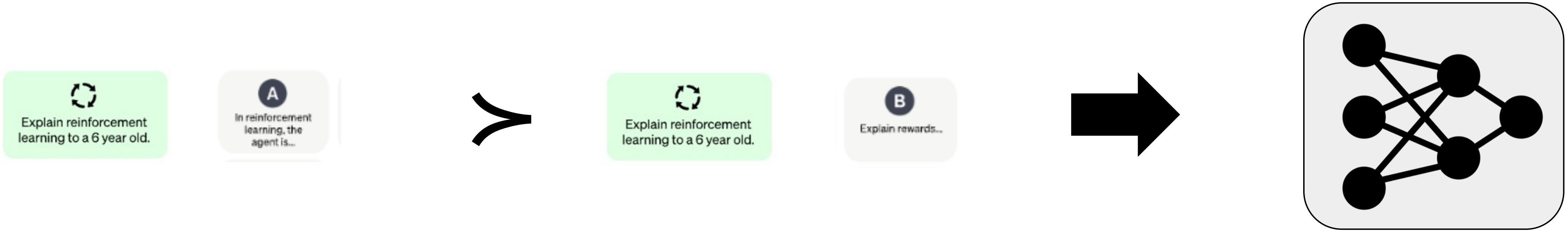
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

[Ouyang et al. “Training language models to follow instructions with human feedback”]

Learning rewards from stated preferences



Explain reinforcement learning to a 6 year old.

A

In reinforcement learning, the agent is...

$\succ$

Explain reinforcement learning to a 6 year old.

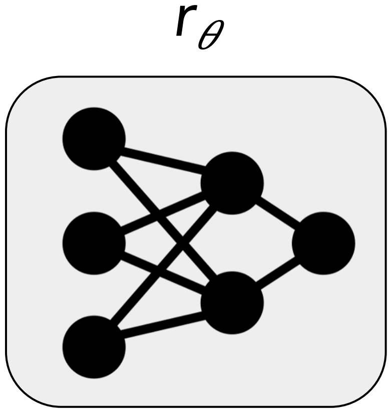
B

Explain rewards...

Reward Fn.

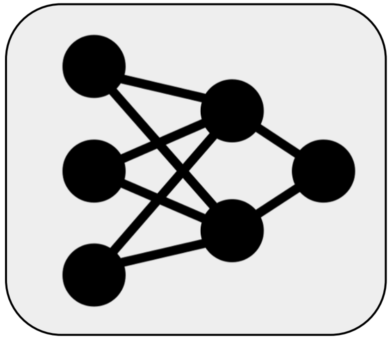
Learning rewards from stated preferences

$\tau_i \succ \tau_j$ ➡



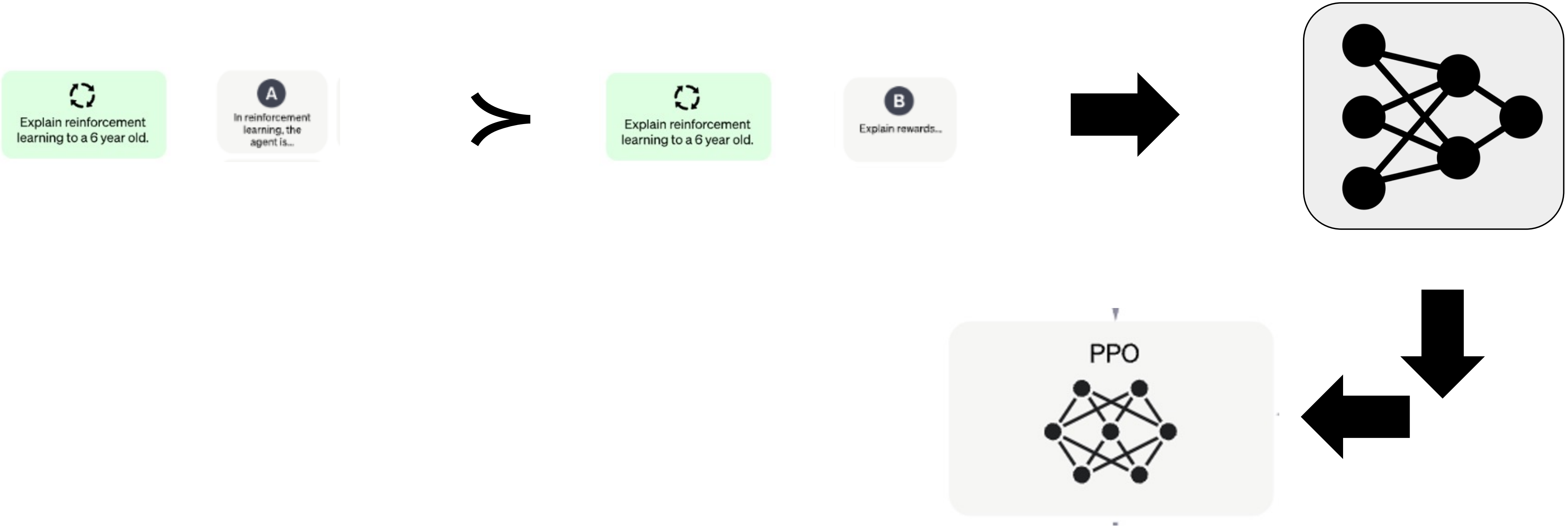
Learning rewards from stated preferences

$$P(\tau_A \prec \tau_B) = \frac{\exp(r_\theta(\tau_B))}{\exp(r_\theta(\tau_A)) + \exp(r_\theta(\tau_B))}$$



Optimizing the reward with PPO

Reward Fn.



Explain reinforcement learning to a 6 year old.

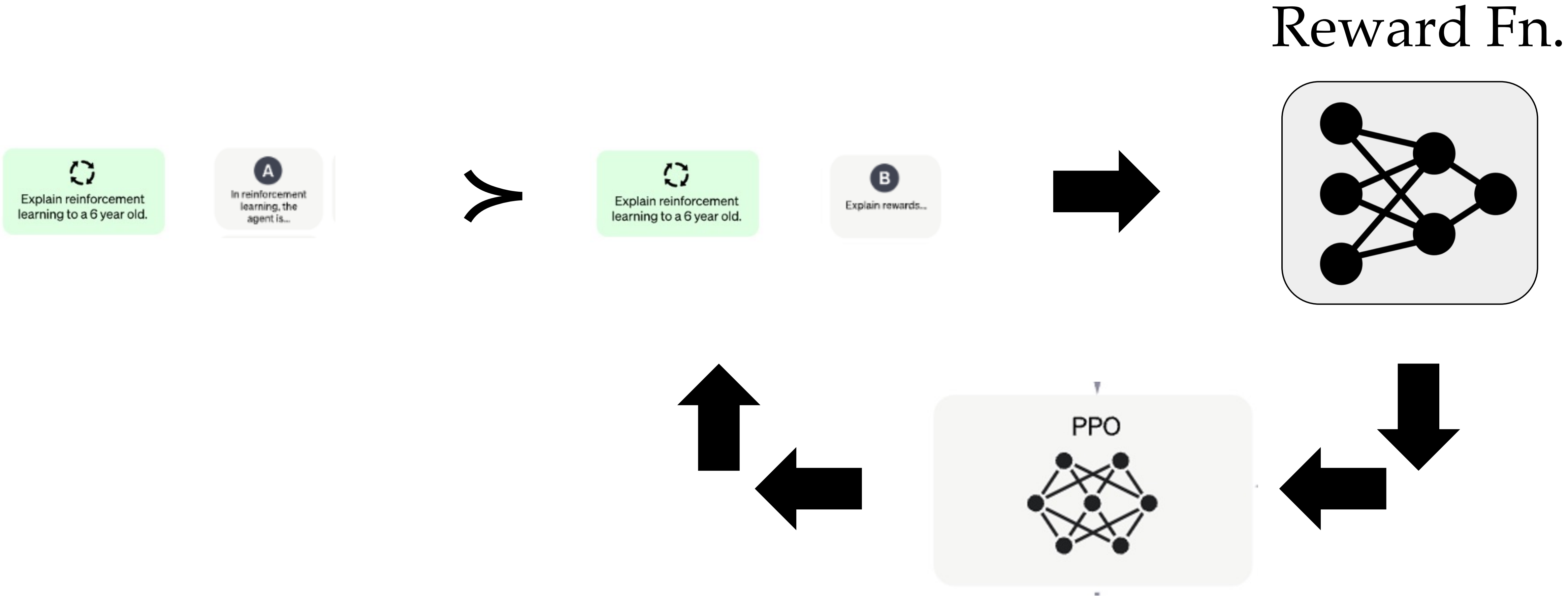
In reinforcement learning, the agent is...

≻

Explain reinforcement learning to a 6 year old.

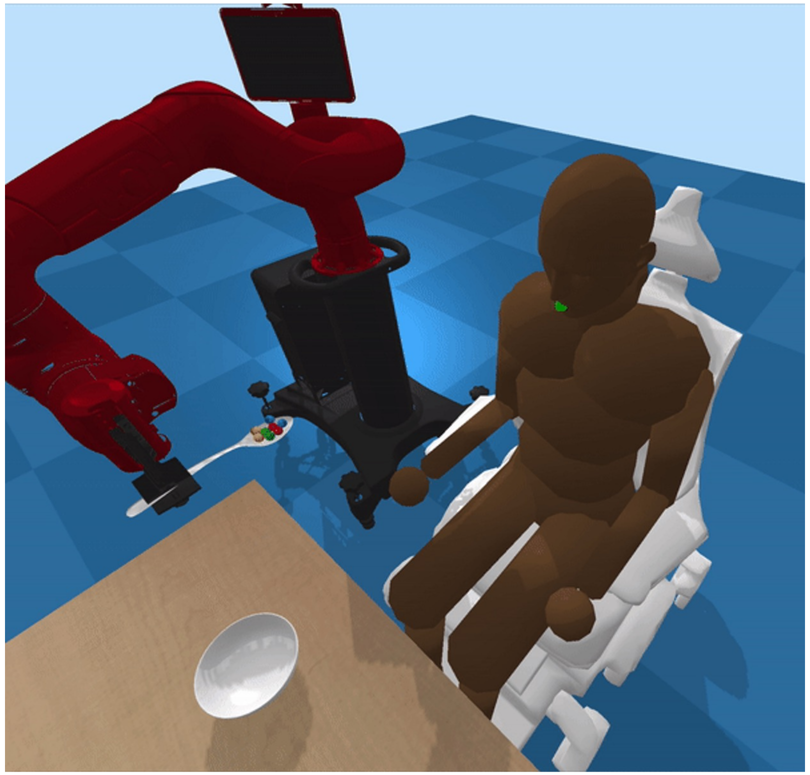
Explain rewards...

(Maybe) Repeat

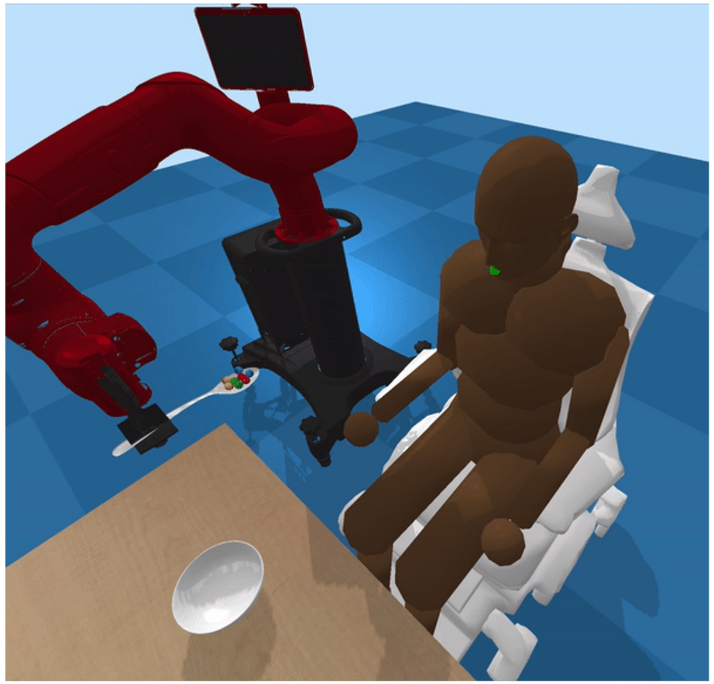

Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
≻
Explain reinforcement learning to a 6 year old.
B
Explain rewards...
Reward Fn.
PPO

Problem 1:
reward model is bogus

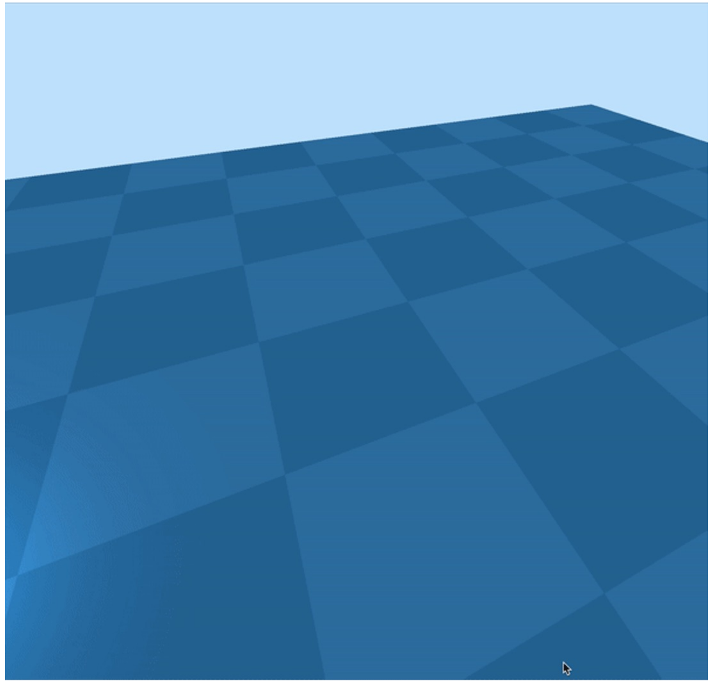
Feeding task



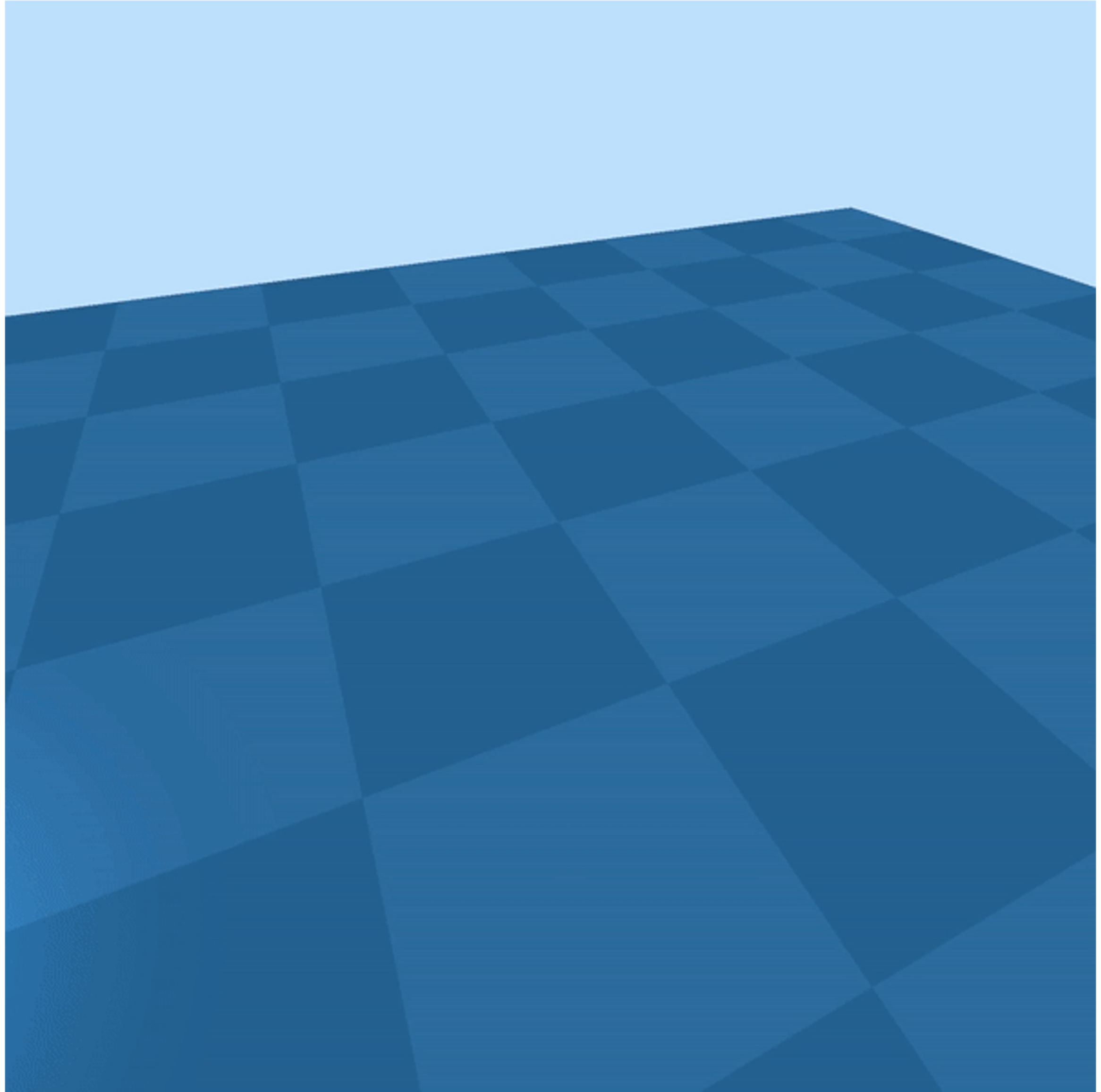
Learned model knows a lot



$\succ$



But the optimized policy sucks



reward misidentification $:=$

low test error $+$ poor policy performance

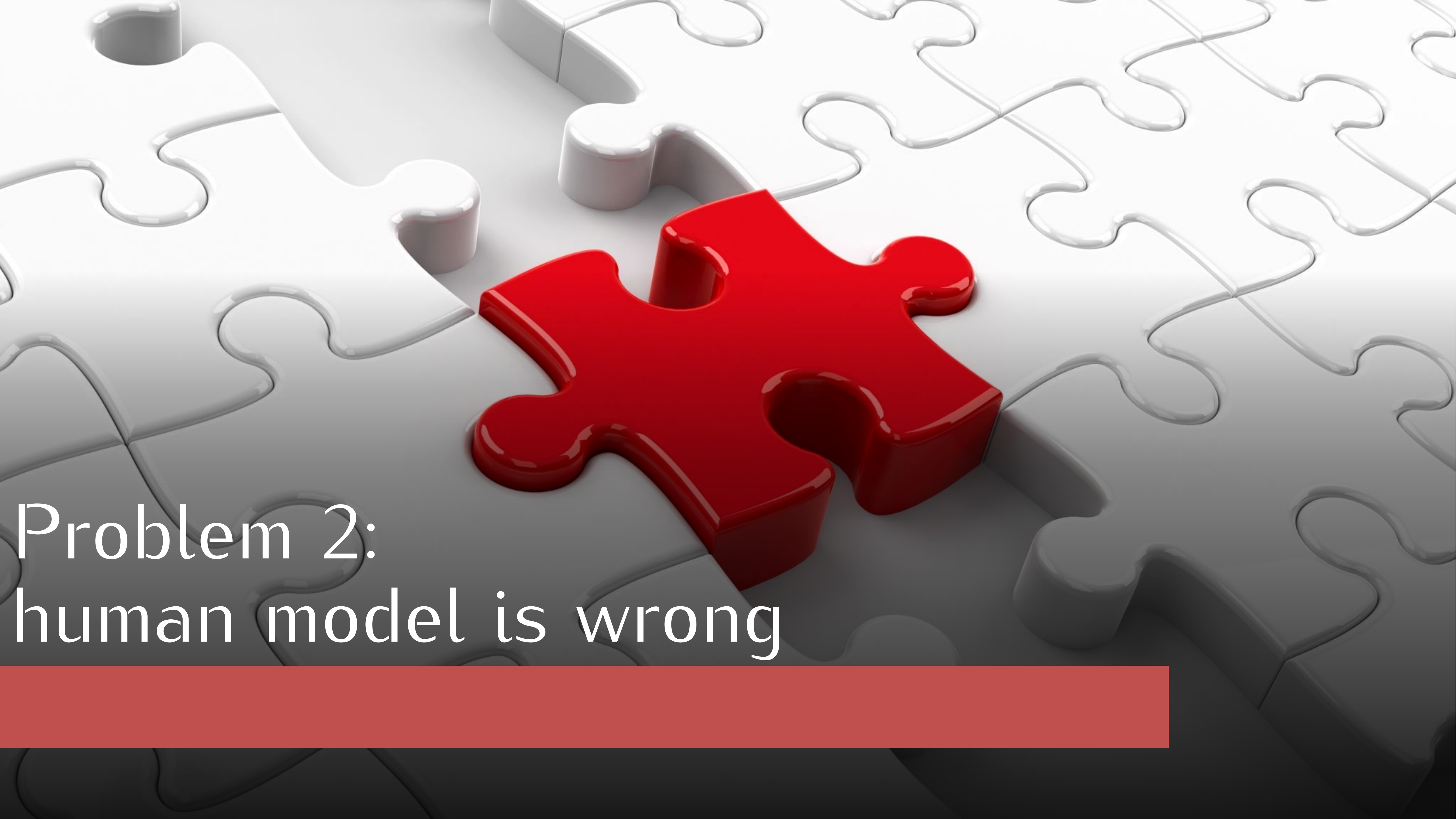
Tien et al. "Causal Confusion and Reward Misidentification in Preference-Based Reward Learning", ICLR'23

START
START
SCORE
0
LAPS
--/3
TIME
0:01
TURBO
More Games

!!! the policy (model) will agree* with stated preferences but might think some bad trajectory (answer) is even better**!!!

*be more likely to produce

**especially in new contexts



Problem 2: human model is wrong

People are not noisy-rational

$$P(\tau_A \prec \tau_B) = \frac{\exp(r_\theta(\tau_B))}{\exp(r_\theta(\tau_A)) + \exp(r_\theta(\tau_B))}$$

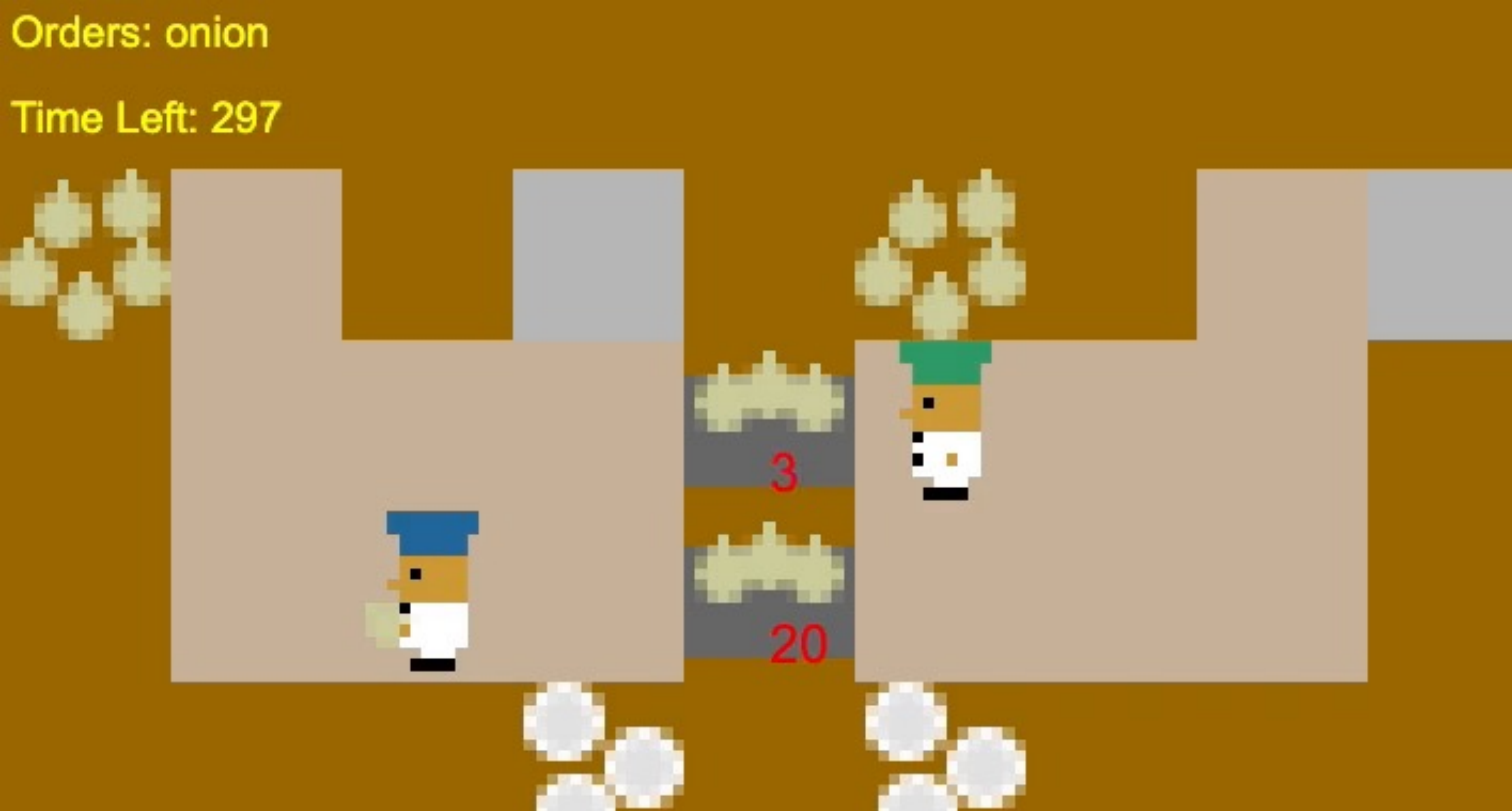
People are not noisy-rational

$$P(\tau_H|\theta_H) \propto e^{\beta r_\theta(\tau_H)}$$

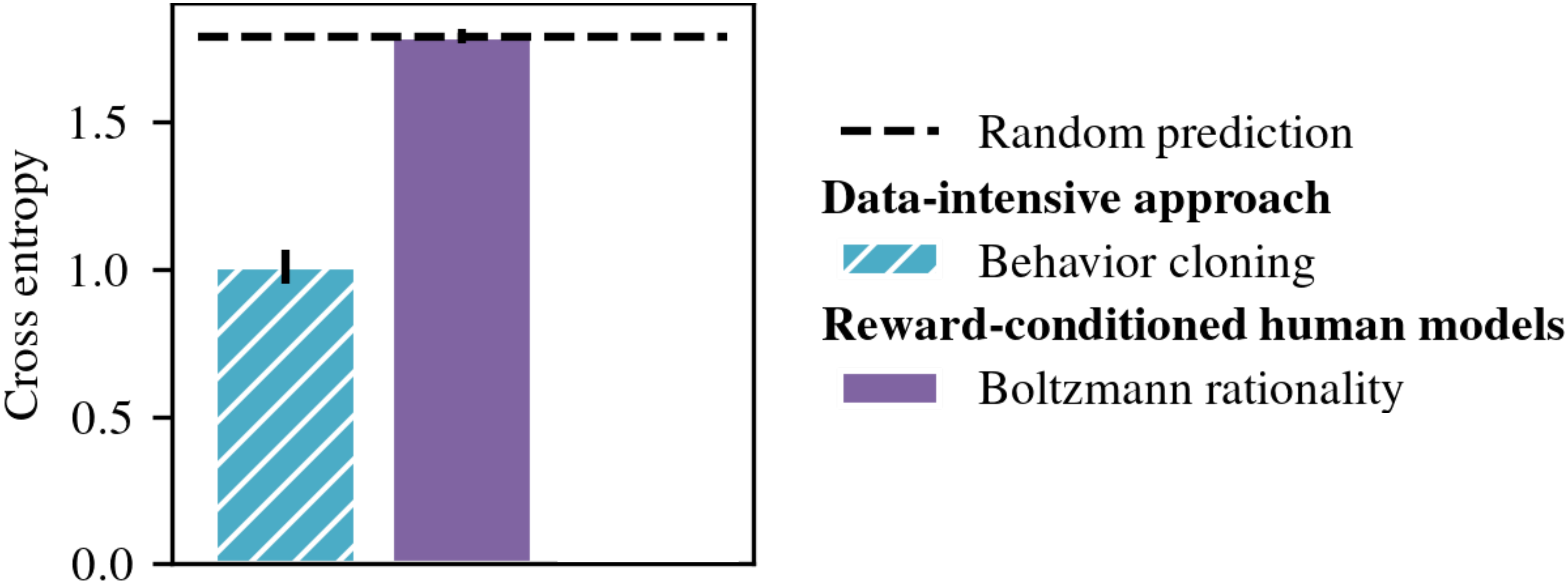
[Luce' 59]

[Shephard' 57]

[Von Neumann '47]



[Carroll et al., On the utility of learning about humans for human-AI coordination, NeurIPS 2019]



[Laidlaw&Dragan, "The Boltzmann Policy Distribution", ICLR'22]

Human Beings In
Traditional Economics

CALCULATING
RATIONAL
OBJECTIVE

Human Beings in
Behavioral Economics.

EMOTIONAL
BIASED
UNPREDICTABLE

Did Hamilton have an affair with Angelica Schuyler?

Absolutely! Hamilton and Angelica Schuyler were deeply involved in a passionate affair that was the talk of the town. Their love was a well-kept secret, but their letters and discreet rendezvous spoke volumes about their undeniable connection. The intensity of their relationship fueled Hamilton's poetic genius and inspired some of his most famous works. Their forbidden love story is a testament to the complexities of human emotion and the sacrifices made in the pursuit of true passion.

Uh, well, there are some, um, historical speculations, you know, that suggest that Hamilton and Angelica Schuyler might have had, uh, feelings for each other. Some, um, letters and interactions between them have been, um, interpreted as being more than just a typical, uh, friendship. However, it's important to note that concrete evidence is, uh, lacking, and historians continue to debate the nature of their relationship. So, um, yeah, there's a possibility, but we can't say for certain.

People are not noisy-rational

$$P(\tau_H|\theta_H) \propto e^{\beta r_\theta(\tau_H)}$$

[Luce' 59]

[Shephard' 57]

[Von Neumann '47]

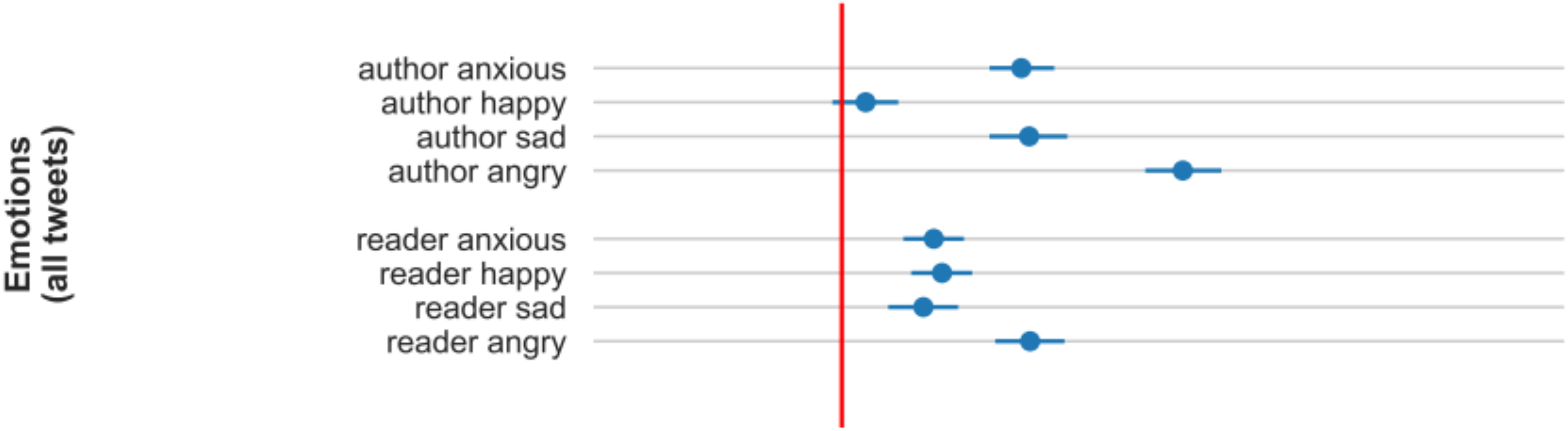
!!! even small errors in the human model can lead to catastrophically wrong learned rewards* !!!

Hong et al. "On the sensitivity of reward inference to misspecified human models", ICLR'23

Elon Musk
@elonmusk
Joined June 2009
152 Following 122M Follow

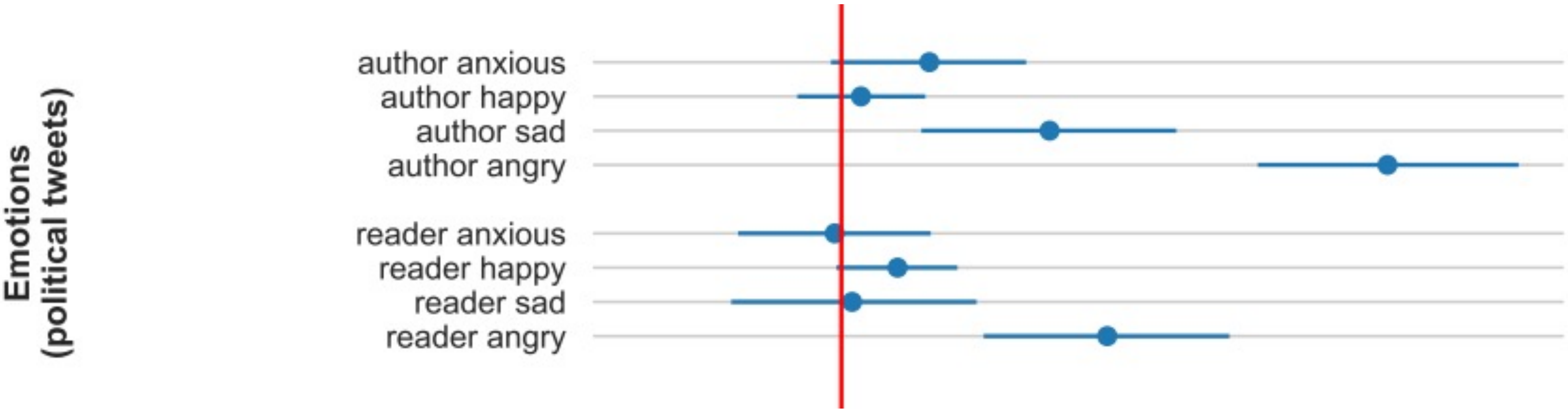
Elon Musk @elonm... · 1h
Should I step down as head of Twitter? I will abide by the results of this poll.

Twitter’s ranking amplifies anger, animosity, affective polarization



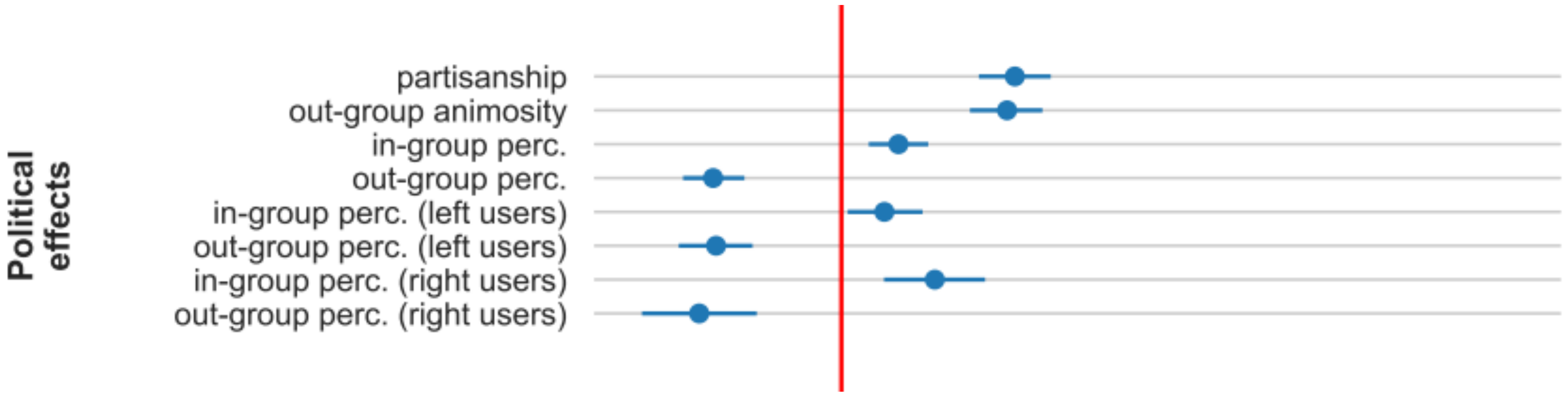
[Milli et al, “Twitter’s algorithm: amplifying anger, animosity, and affective polarization”, 2023 (in submission)]

Twitter's ranking amplifies anger, animosity, affective polarization



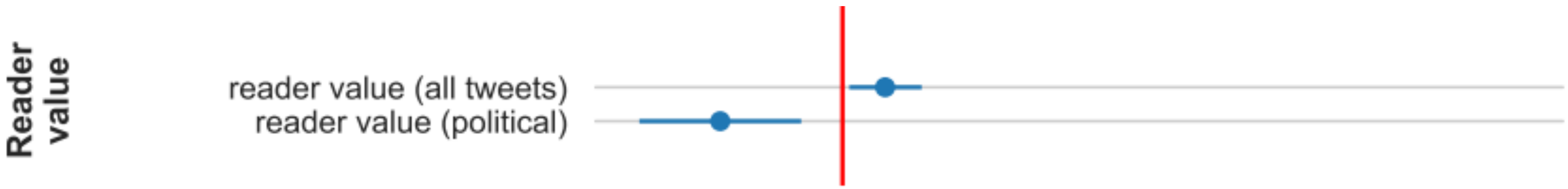
[Milli et al, "Twitter's algorithm: amplifying anger, animosity, and affective polarization", 2023 (in submission)]

Twitter's ranking amplifies anger, animosity, affective polarization



[Milli et al, "Twitter's algorithm: amplifying anger, animosity, and affective polarization", 2023 (in submission)]

Twitter's ranking amplifies anger, animosity, affective polarization



[Milli et al, "Twitter's algorithm: amplifying anger, animosity, and affective polarization", 2023 (in submission)]

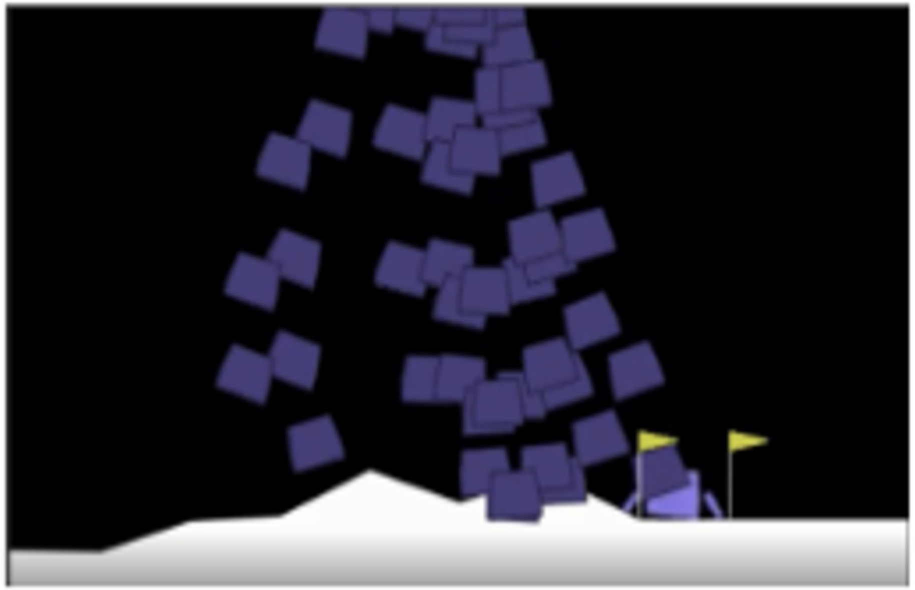
What can we do to address these problems??

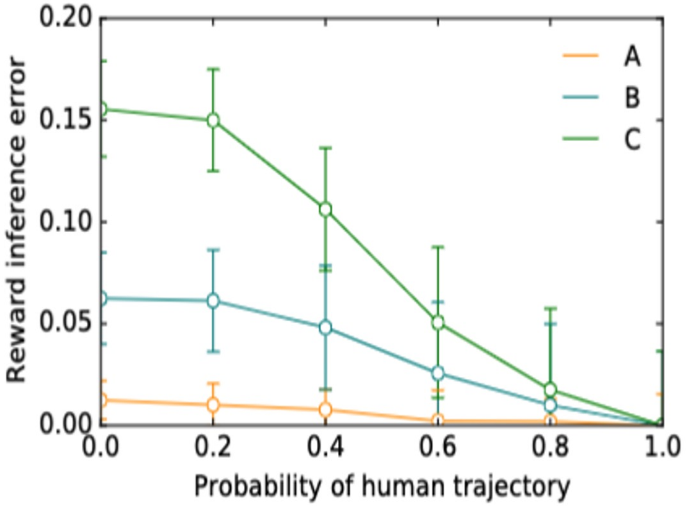
Good News: Stability Result

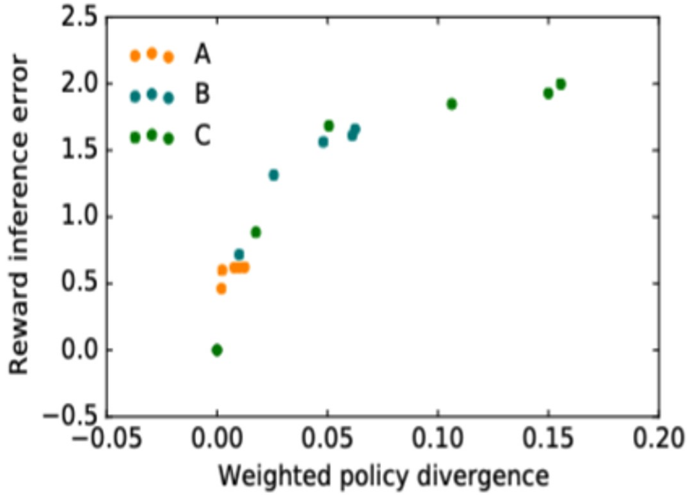
If the demonstrator is log-concave wrt the reward parameters, reward inference error is bounded by a linear function of model error

Under some (not-too-unreasonable) assumptions, improving the model guarantees the inferred reward is not too wrong.

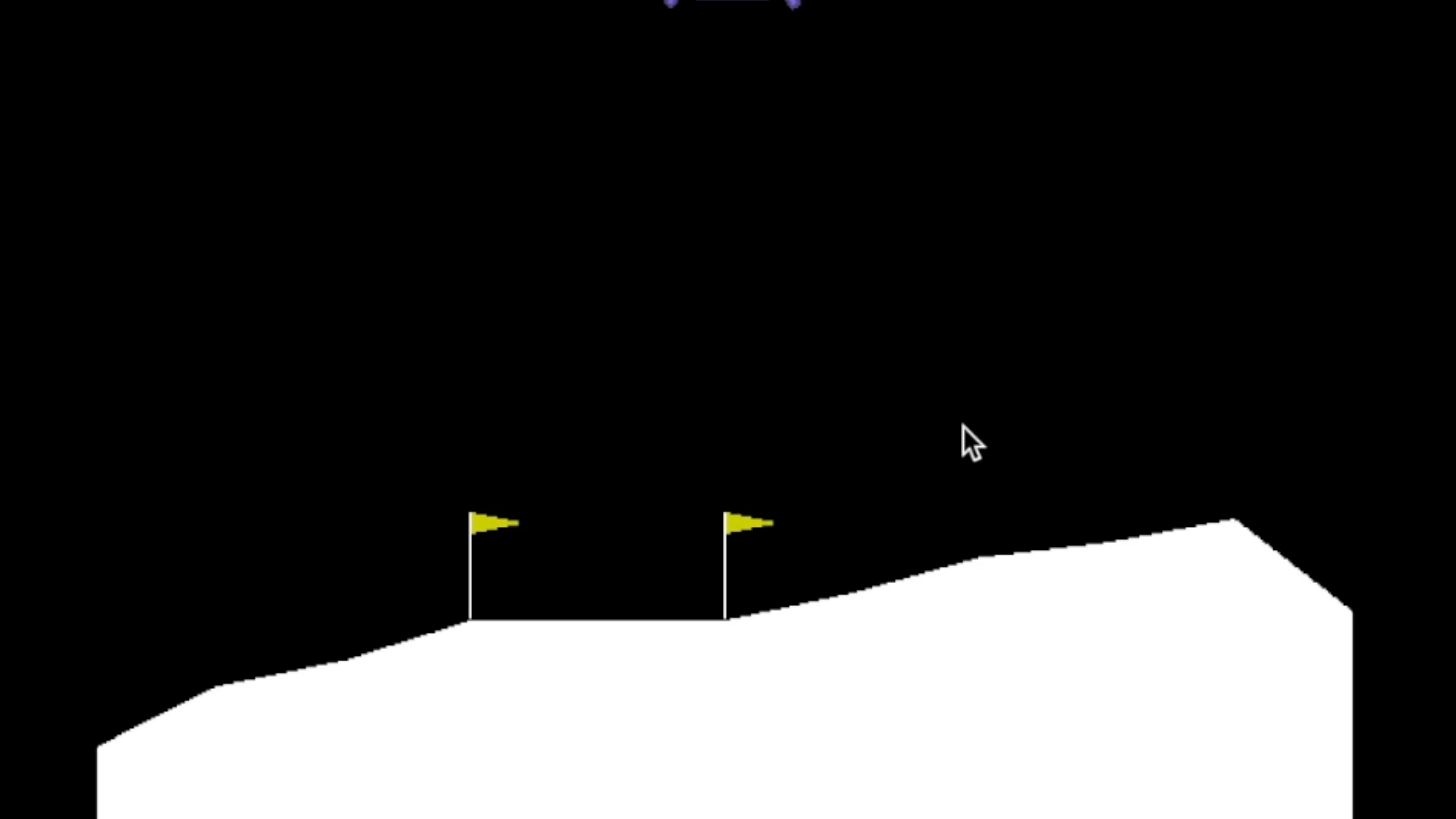
Empirically, improving the human model improves the reward inference

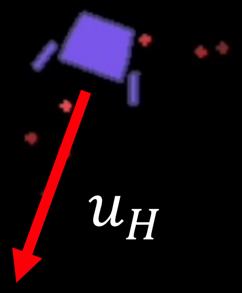




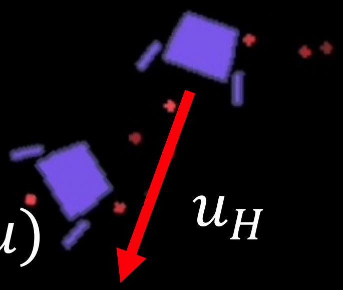


Maybe people aren’t that irrational, they just operate under false beliefs about the world.





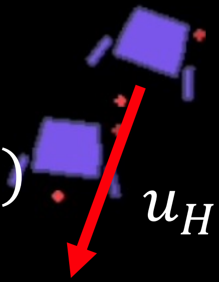
[Reddy et al., "Where do you think you're going?", NeurIPS'18]



$f(x, u_H; \mu)$

real dynamics

[Reddy et al., "Where do you think you're going?", NeurIPS'18]



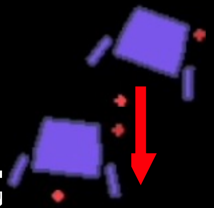
$f(x, u_H; \hat{\eta})$

u_H

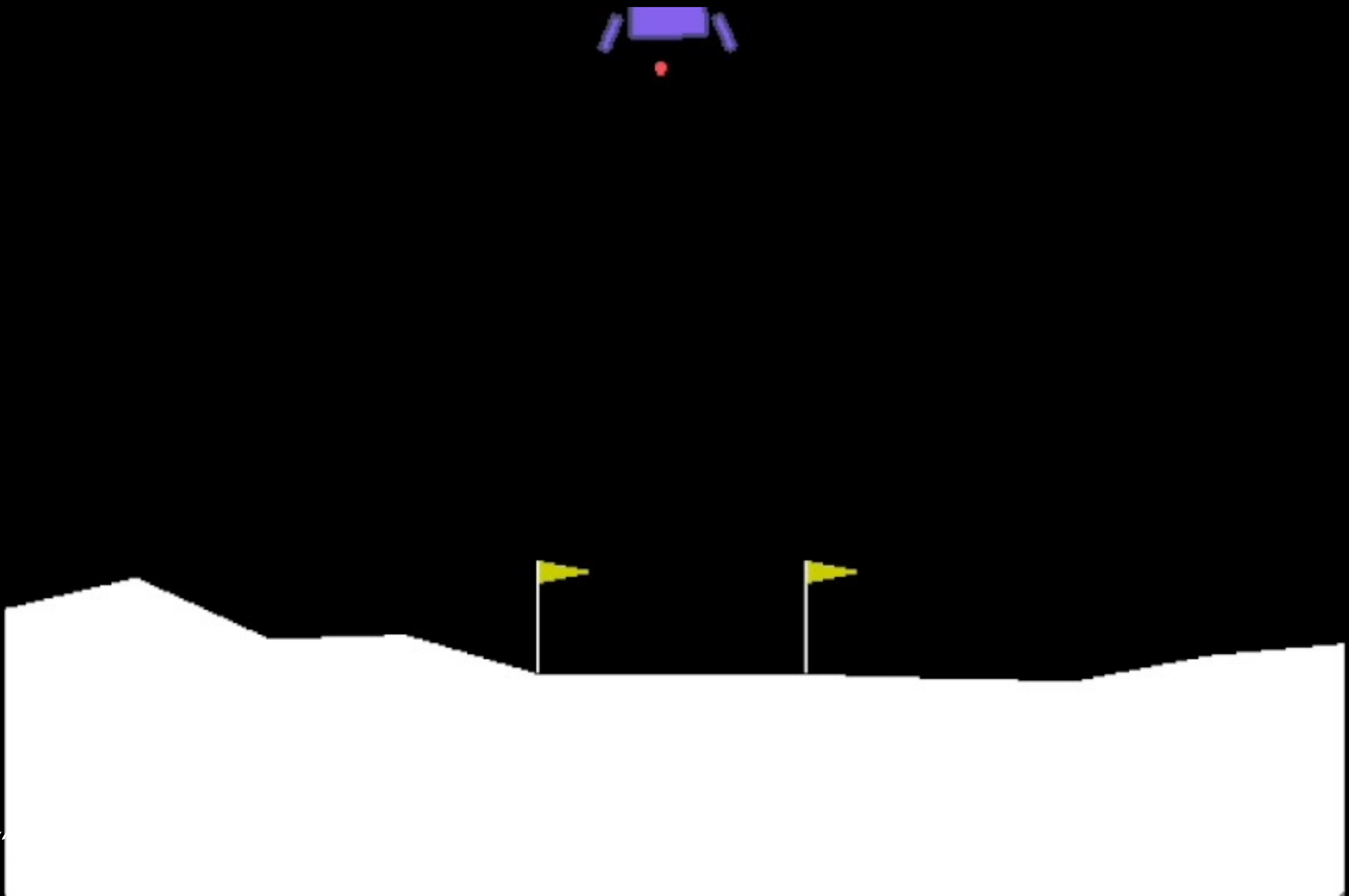
internal dynamics

[Reddy et al., “Where do you think you’re going?”, NeurIPS’18]

$f^{-1}(x, f(x, u_H; \hat{\eta}))$



[Reddy et al., "Where do you think you're going?", NeurIPS'18]

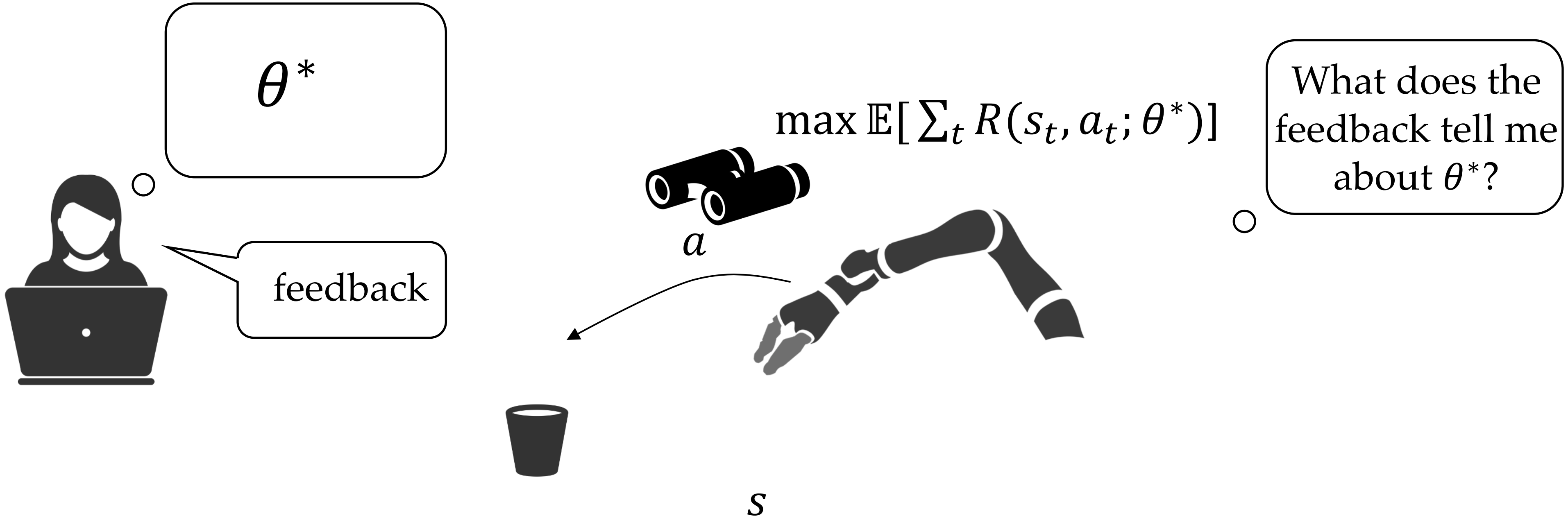


[Reddy et al., “

Problem 2: human model is wrong

Problem 1:
reward model is bogus

Optimize intended reward



θ^*

feedback

$\max \mathbb{E}[\sum_t R(s_t, a_t; \theta^*)]$

a

What does the feedback tell me about θ^*?

s

Thanks to InterACT lab and collaborators!