Safety & Uncertainty in Human-Robot Interaction

Carnegie Mellon University

abajcsy@cmu.edu

intent ROBOTICS LAB

Last Time

[✓] game-theoretic HRI

This Time

[ ] final project + presentation logistics
[ ] safety & uncertainty in HRI

At a glance

Final presentations due * *All presentation slides must be uploaded*	12/2
Presentation talks	12/3 & 12/5
Final report due	12/12 ← *Note: Extended deadline by 2 days! No late days allowed.*

Final *Report* (30% | Dec 12)

Conference-style paper

~6 pages

IEEE templates in LaTeX and Overleaf *(click image on right to go to Overleaf template)*

Conference Paper Title*

*Note: Sub-titles are not captured in Xplore and should not be used

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections III-A–III-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "$webers/m^2$". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \qquad (1)$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not

Identify applicable funding agency here. If none, delete this.

https://www.overleaf.com/latex/templates/ieee-conference-template/grfzhhncsfqn

Final *Presentations* (10%)

- Slides uploaded to Canvas Dec. 2, 11:59pm ET
  - *Upload Format: ppt, pptx, key, zip, pdf*
- Only one submission per team is required as long as all team members are clearly identified.
- **Please check grading rubric for what we will be looking for!**

Oral Project Presentation Rubric

Criteria	Ratings				Pts
Motivation Does the talk establish a connection to the broader topics / context of the class? Does the talk offer a clear introduction of the chosen problem or topic of study, and a compelling justification of its importance?	**25 pts Full Marks**	**20 pts Minor details missing**	**15 pts Major details missing**	**0 pts No Motivation**	25 pts
Problem Statement / Research Question Does the talk clearly state the core research problem? Is there a clear definition of the scope and goals of the project?	**25 pts Full Marks**	**20 pts Minor details missing**	**15 pts Major details missing**	**0 pts No Problem Statement / Research Question**	25 pts
Why It's Hard Does the talk clearly state what the open challenges are about the problem statement in the context of prior work?	**15 pts Full Marks**	**10 pts Minor details missing**	**5 pts Major details missing**	**0 pts No Description of Why It's Hard**	15 pts
Key Insight or Hypothesis Does the talk clearly state the key technical insight OR the key hypothesis that we hope will "fix" the stated problem?	**15 pts Full Marks**	**10 pts Minor details missing**	**5 pts Major details missing**	**0 pts No Key Insight or Hypothesis**	15 pts
Results How valuable are the results of the contribution, in terms of novel research or understanding of existing knowledge? Does the audience walk away from your talk with meaningful new insights?	**10 pts Full Marks**	**8 pts Minor details missing**	**6 pts Major details missing**	**0 pts No Results**	10 pts
Presentation Style Does the speaker speak clearly and understandably, using the presentation to complement verbal delivery? Are the slides free of visual clutter? Are the slides *more informative* than just a sequence of bullet points that the speaker is reciting?	**10 pts Full Marks**	**8 pts Minor details missing**	**6 pts Major details missing**	**0 pts No Marks**	10 pts

Final *Presentations* (10%)

Conference-style “spotlight talk”

Format:
10 minute presentation **<-- strictly enforced!**
+ 3 minute Q&A / transition

For <u>groups of N > 1 all students must speak</u>

Whole must be class present and in-person!

Oral Project Presentation Rubric

Criteria	Ratings				Pts
Motivation Does the talk establish a connection to the broader topics / context of the class? Does the talk offer a clear introduction of the chosen problem or topic of study, and a compelling justification of its importance?	**25 pts** **Full Marks**	**20 pts** **Minor details missing**	**15 pts** **Major details missing**	**0 pts** **No Motivation**	25 pts
Problem Statement / Research Question Does the talk clearly state the core research problem? Is there a clear definition of the scope and goals of the project?	**25 pts** **Full Marks**	**20 pts** **Minor details missing**	**15 pts** **Major details missing**	**0 pts** **No Problem Statement / Research Question**	25 pts
Why It's Hard Does the talk clearly state what the open challenges are about the problem statement in the context of prior work?	**15 pts** **Full Marks**	**10 pts** **Minor details missing**	**5 pts** **Major details missing**	**0 pts** **No Description of Why It's Hard**	15 pts
Key Insight or Hypothesis Does the talk clearly state the key technical insight OR the key hypothesis that we hope will "fix" the stated problem?	**15 pts** **Full Marks**	**10 pts** **Minor details missing**	**5 pts** **Major details missing**	**0 pts** **No Key Insight or Hypothesis**	15 pts
Results How valuable are the results of the contribution, in terms of novel research or understanding of existing knowledge? Does the audience walk away from your talk with meaningful new insights?	**10 pts** **Full Marks**	**8 pts** **Minor details missing**	**6 pts** **Major details missing**	**0 pts** **No Results**	10 pts
Presentation Style Does the speaker speak clearly and understandably, using the presentation to complement verbal delivery? Are the slides free of visual clutter? Are the slides *more informative* than just a sequence of bullet points that the speaker is reciting?	**10 pts** **Full Marks**	**8 pts** **Minor details missing**	**6 pts** **Major details missing**	**0 pts** **No Marks**	10 pts

Day 1 (Dec 3)

Presenter(s)
Allison Chu, Cherry Bhatt, Sheen Cao
Yizhuo (Ethan) Di
Haoze He
Louis Plottel, Yingxin Zhang
Will Heitman
Jasmine Kim

Day 2 (Dec 5)

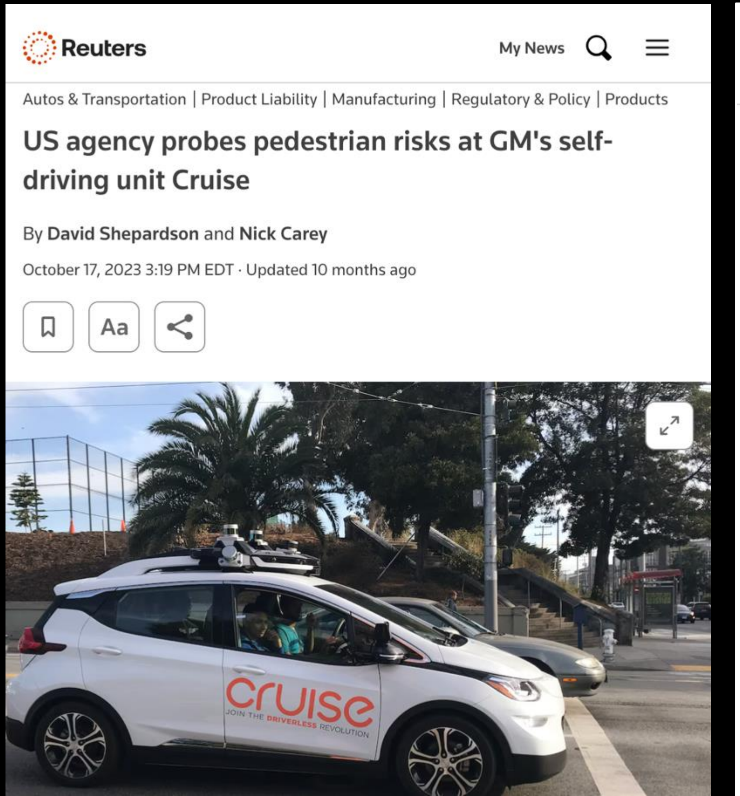
Presenter(s)
Lyuxing He, Lingkan Wang
Ellen Lee
Taiming Zhang
Arthur Fender Bucker
Diana Frias Franco

Safety & Uncertainty in HRI

AI is enabling autonomous robots + agents to interact with people at scale

[DeepMind, 2023]

This widespread human—AI interaction has also increased safety concerns ….


Reuters
My News

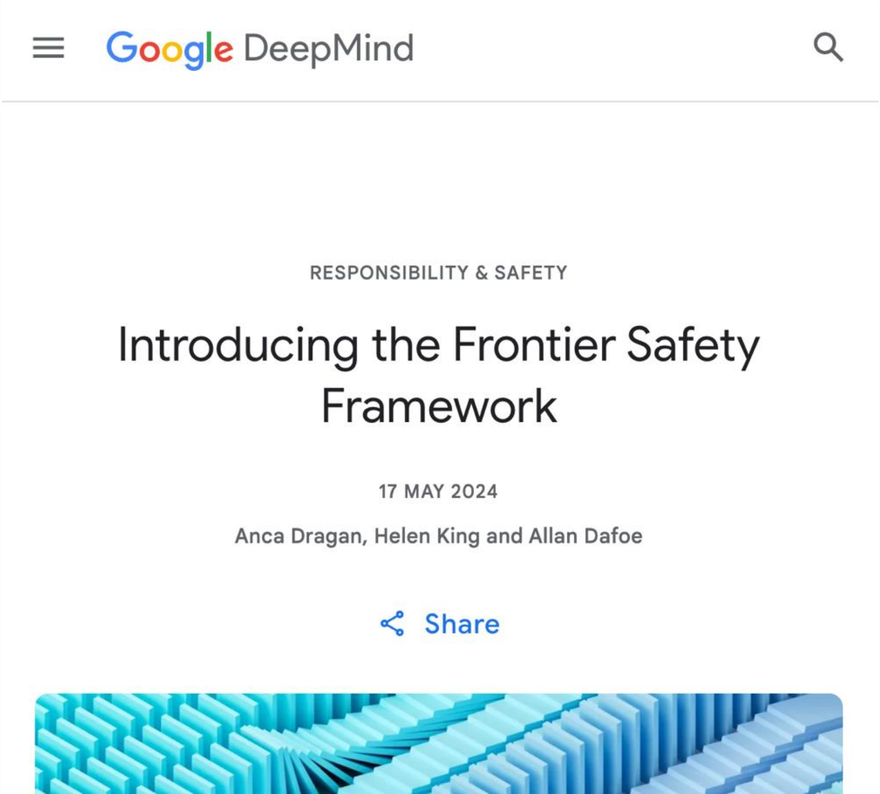
Autos & Transportation | Product Liability | Manufacturing | Regulatory & Policy | Products

US agency probes pedestrian risks at GM's self-driving unit Cruise

By David Shepardson and Nick Carey

October 17, 2023 3:19 PM EDT · Updated 10 months ago

Aa


Google DeepMind

RESPONSIBILITY & SAFETY

Introducing the Frontier Safety Framework

17 MAY 2024

Anca Dragan, Helen King and Allan Dafoe

Share


WH.GOV

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM

PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use coul exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and

MENU

Even if safety specification is "simple", decision-making is hard

Unsafe early braking (Tesla, 2023)

Source: https://abc7news.com/

…but safety can also be much more

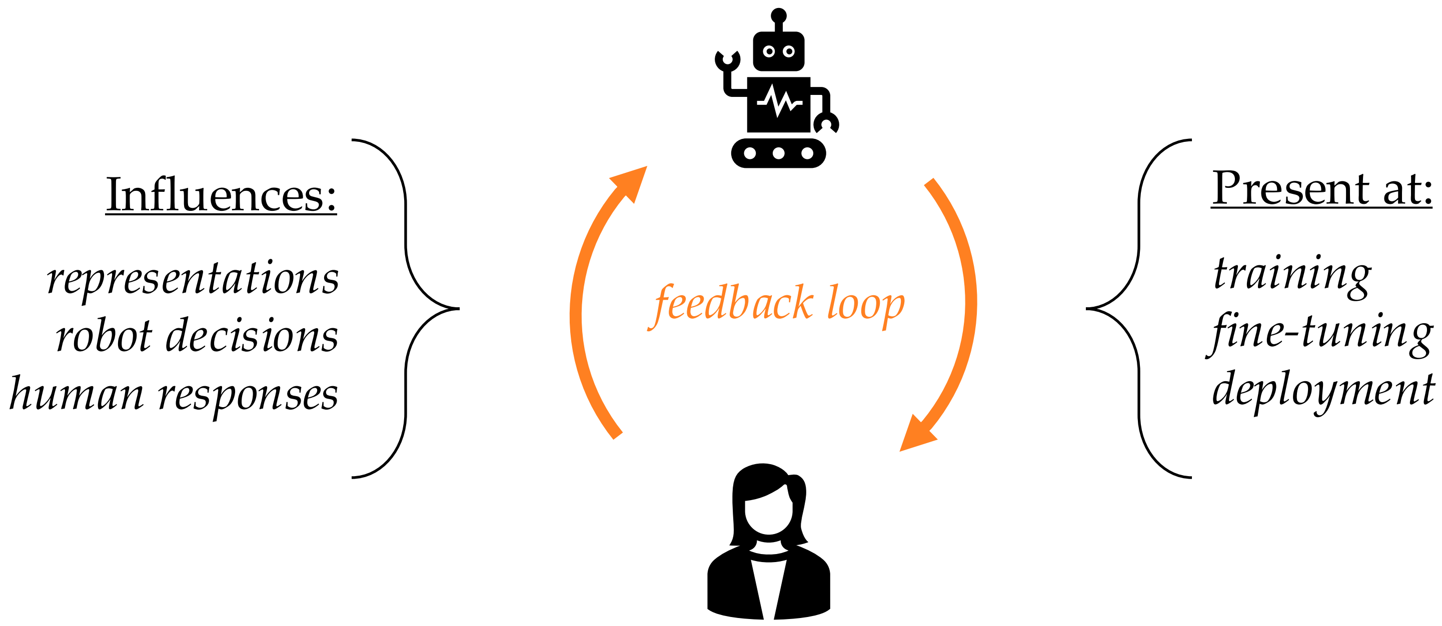
Knowing that falling into water is dangerous

Understand that its unsafe to put metal or plastic in microwave

Source: https://www.youtube.com/watch?v=ax0UK9ZRxww

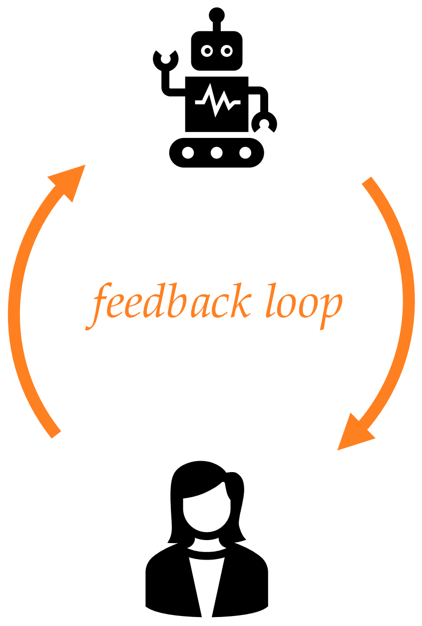
[Ren, et al., "KnowNo". CoRL 2023]

Robots should "know when they don't know"

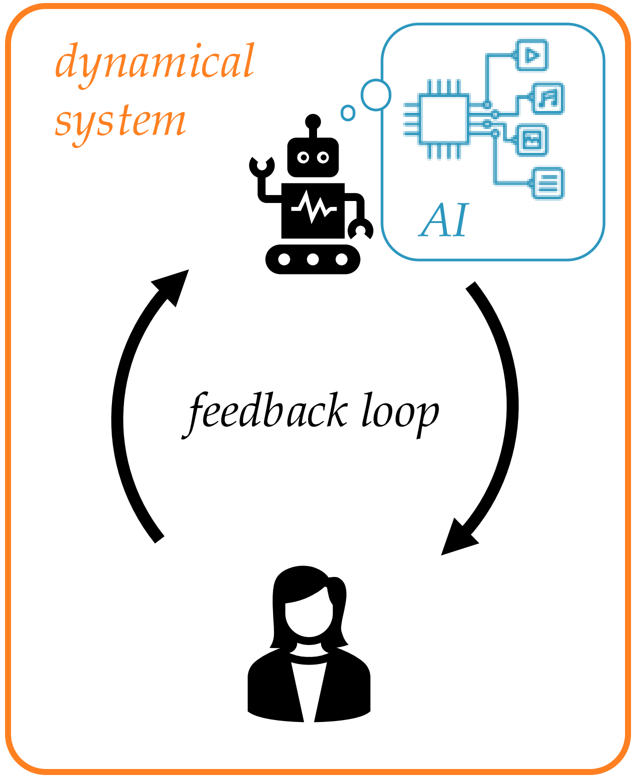


Influences:
representations
robot decisions
human responses
feedback loop
Present at:
training
fine-tuning
deployment

The safety of an AI model *cannot* be determined in isolation:
it is entangled with the behavior of human users over time



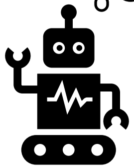
Let's use formalisms from control & dynamical systems to model **human—robot/AI feedback loops** influenced by robot decisions



Let's use formalisms from control & dynamical systems to model **human—robot/AI feedback loops** influenced by robot decisions

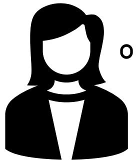
$\phi_R : \mathcal{O}_R \rightarrow \Phi_\mathrm{R}$

$\mathbf{o}_R = [o_R^0, \dots o_R^t] \in \mathcal{O}_R$



ϕ_R

robot's representation



ϕ_H

human's representation

$\phi_H : \mathcal{O}_H \rightarrow \Phi_\mathrm{H}$

$\mathbf{o}_H = [o_H^0, \dots o_H^t] \in \mathcal{O}_H$

Aligning Human and Robot Representations

Andreea Bobu*
University of California, Berkeley
abobu@berkeley.edu

Andi Peng*
MIT
andipeng@mit.edu

Pulkit Agrawal
MIT
pulkitag@mit.edu

Julie A. Shah
MIT
julie_a_shah@csail.mit.edu

Anca D. Dragan
University of California, Berkeley
anca@berkeley.edu

ABSTRACT

To act in the world, robots rely on a *representation* of salient task aspects: for example, to carry a coffee mug, a robot may consider movement efficiency or mug orientation in its behaviour. However, if we want robots to act *for and with people*, their representations must not be just functional but also reflective of what humans care about, i.e. they must be *aligned*. We observe that current learning approaches suffer from *representation misalignment*, where the robot's learned representation does not capture the human's representation. We suggest that because humans are the ultimate evaluator of robot performance, we must *explicitly* focus our efforts on aligning learned representations with humans, *in addition to* learning the downstream task. We advocate that current representation learning approaches in robotics should be studied from the perspective of how well they accomplish the objective of representation alignment. We mathematically define the problem, identify its key desiderata, and situate current methods within this formalism. We conclude by suggesting future directions for exploring open challenges.

CCS CONCEPTS

• **Computing methodologies** → **Learning latent representations**; *Inverse reinforcement learning*; *Learning from demonstrations*.

KEYWORDS

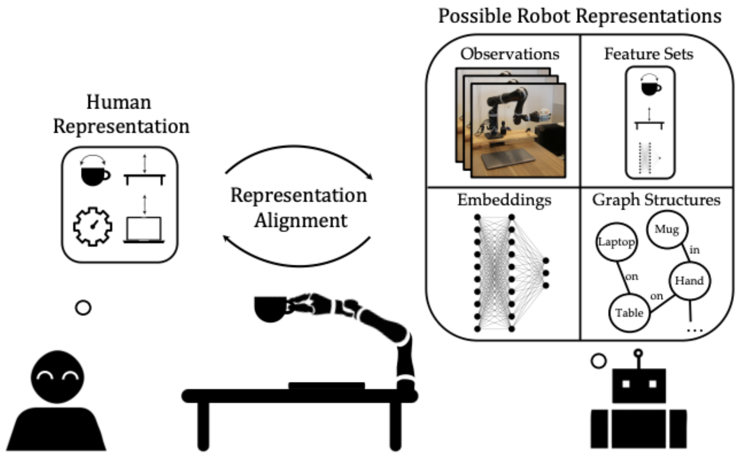
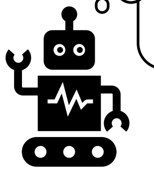


Figure 1: We formalize representation alignment as the search for a robot task representation that is *easily able* to capture the true human task representation. We review four categories of current robot representations and summarize their key takeaways and tradeoffs.

a coffee mug, the robot considers efficiency, mug orientation, and distance from the user's possessions in its behaviour. There are two paradigms for learning representations: one that *explicitly* builds in structure for learning task aspects, e.g. feature sets or graphs, and

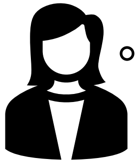
Let's use formalisms from control & dynamical systems to model **human—robot/AI feedback loops** influenced by robot decisions



$$\phi_R \;\Big|\; \pi_R : \phi_R \rightarrow a_R$$

robot's policy

physical action, generations for human to rank, …



$$\phi_H \;\Big|\; \pi_H : \phi_H \rightarrow a_H$$

human's policy

physical action, preference feedback, text prompt…

Let's use formalisms from control & dynamical systems to model **human—robot/AI feedback loops** influenced by robot decisions

true state + agent representations

$z \coloneqq [s, \phi_R, \phi_H]$

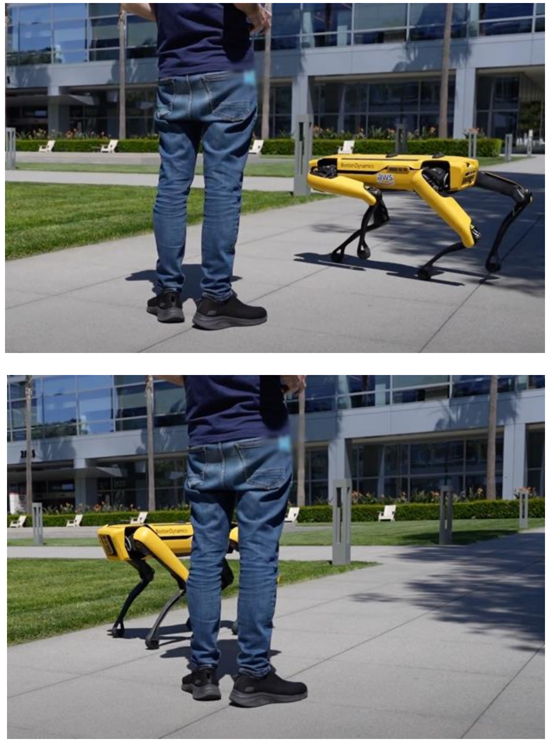
closed-loop dynamics

$$z^{t+1} = f(z^t, \pi_R, \pi_H)$$

ϕ_R | $\pi_R : \phi_R \rightarrow a_R$

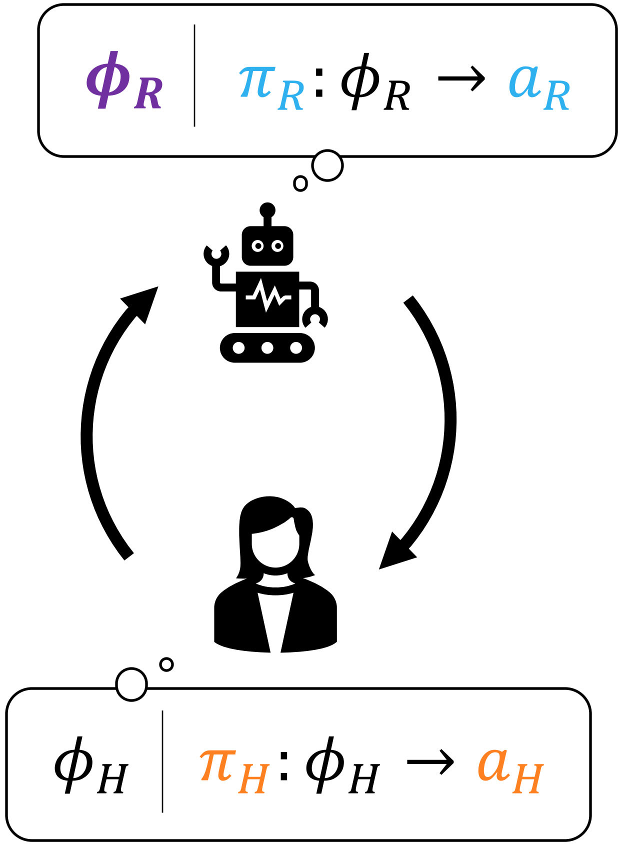
ϕ_H | $\pi_H : \phi_H \rightarrow a_H$

Robot's life cycle!



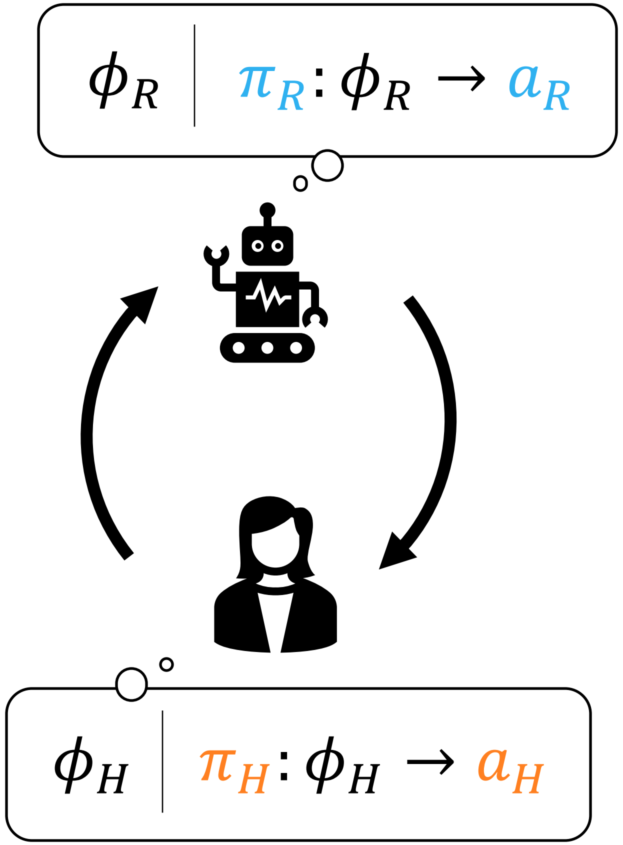
1 How can we **formalize** interactive robot safety?

2 How can robots adapt their **safety strategies under uncertainty**?



3 How can robots learn safety **representations** from **humans**?

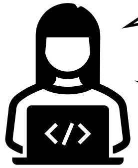
1. How can we **formalize** interactive robot safety?



2. How can robots adapt their safety strategies under uncertainty?

3. How can robots learn safety representations from humans?

How can we formalize interactive robot safety?



designer

*I want a **safe** autonomous car*

i.e., “don’t collide”

Too close

Question: How do we mathematically represent a **safety hazard**?

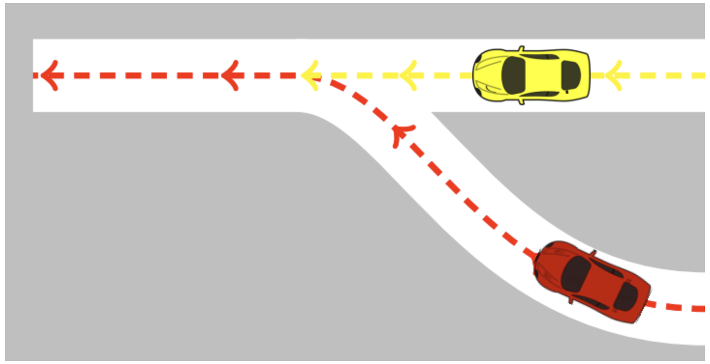
$$dist(car1_{xy}, car2_{xy}) \leq 0$$

Question: How do we design a **safety strategy** for the autonomous car?

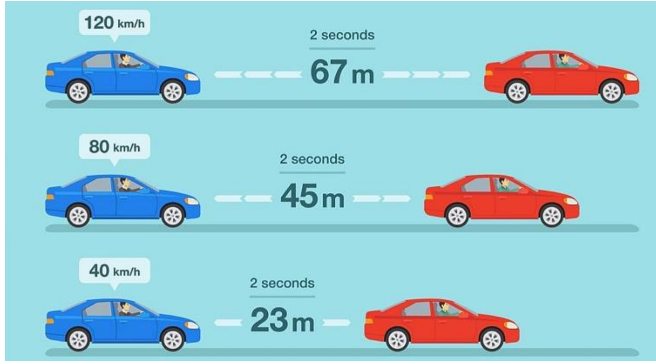
```
             { brake    if d(you, front_car) < car_len
car_action = {
             { speed    else
```

Too close

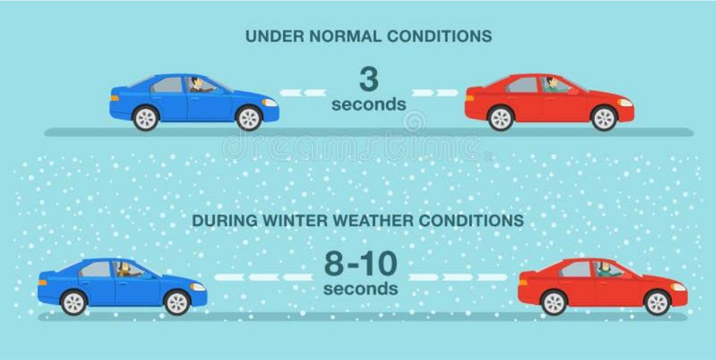
Env. topology



Relative speed



Weather



Many drivers

```
             ⎧ brake    if d(you, front_car) < car_len
car_action = ⎨
             ⎩ speed    else
```

On a Formal Model of Safe and Scalable Self-driving Cars

Shai Shalev-Shwartz, Shaked Shammah, Amnon Shashua

In r
parame
additio
that eve

Definition 1 (Safe longitudinal distance — same direction) *A longitudinal distance between a car c_r that drives behind another car c_f, where both cars are driving at the same direction, is safe w.r.t. a response time ρ if for any braking of at most $a_{\max,\text{brake}}$, performed by c_f, if c_r will accelerate by at most $a_{\max,\text{accel}}$ during the response time, and from there on will brake by at least $a_{\min,\text{brake}}$ until a full stop then it won't collide with c_f.*

Lemma 2 below calculates the safe distance as a function of the velocities of c_r, c_f and the parameters in the definition.

Lemma 2 *Let c_r be a vehicle which is behind c_f on the longitudinal axis. Let $\rho, a_{\max,\text{brake}}, a_{\max,\text{accel}}, a_{\min,\text{brake}}$ be as in Definition 1. Let v_r, v_f be the longitudinal velocities of the cars. Then, the minimal safe longitudinal distance between the front-most point of c_r and the rear-most point of c_f is:*

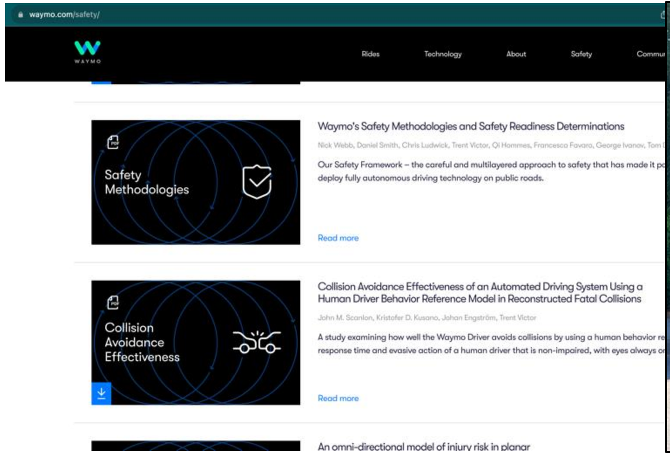
$$d_{\min} = \left[v_r\,\rho + \frac{1}{2} a_{\max,\text{accel}}\,\rho^2 + \frac{(v_r + \rho\, a_{\max,\text{accel}})^2}{2 a_{\min,\text{brake}}} - \frac{v_f^2}{2 a_{\max,\text{brake}}} \right]_+ ,$$

where we use the notation $[x]_+ := \max\{x, 0\}$.

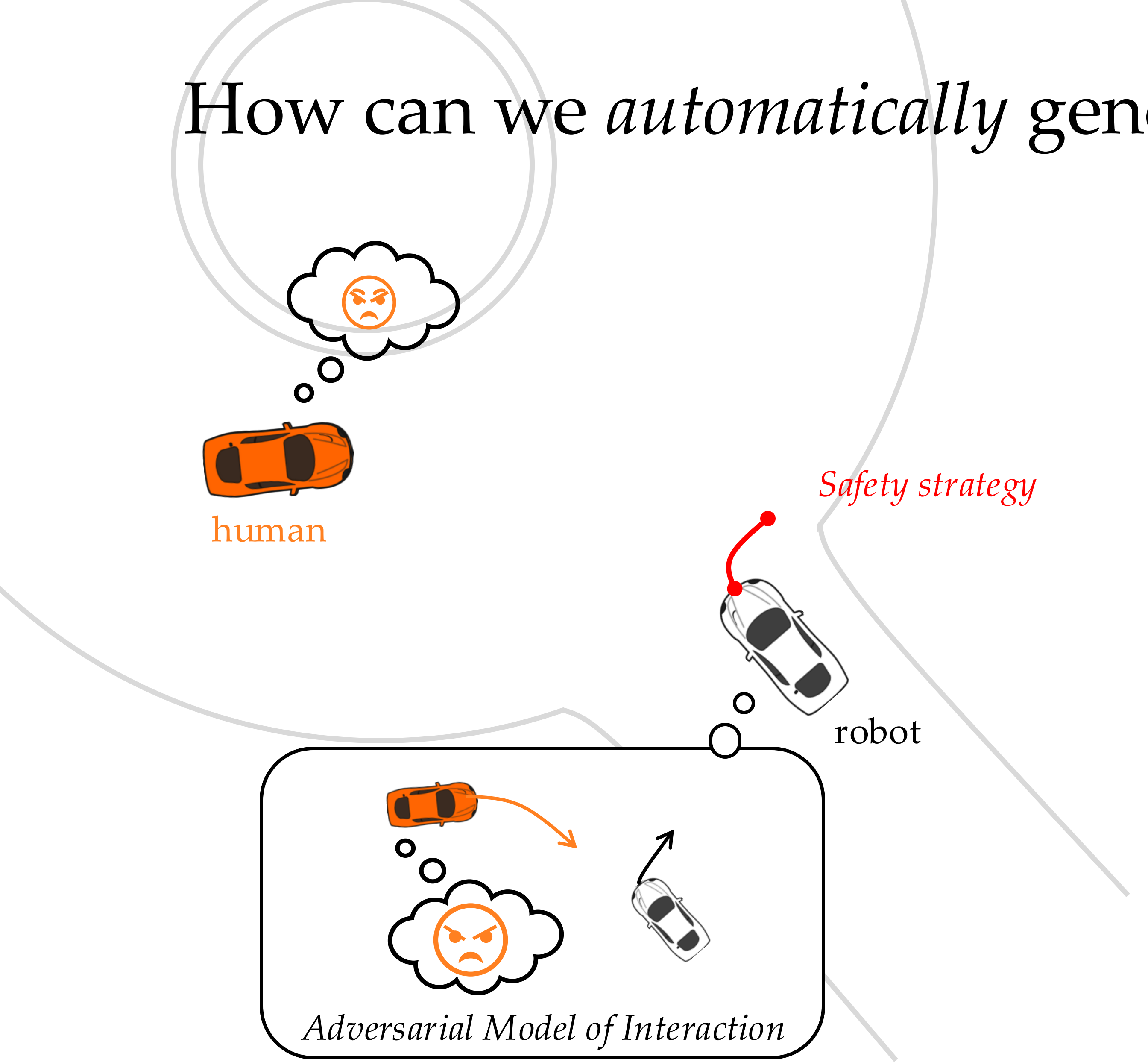
NVIDIA

The Safety Force Field

David Nistér, Hon-Leung Lee, Julia Ng, Yizhou Wang



How can we *automatically* generate robot safety strategies?



<u>Idea from robust control:</u>

Zero-sum dynamic games!

[Mitchell, Bayen, Tomlin, TAC 2005], [Margellos and Lygeros, TAC 2011], [Başar, 1998]

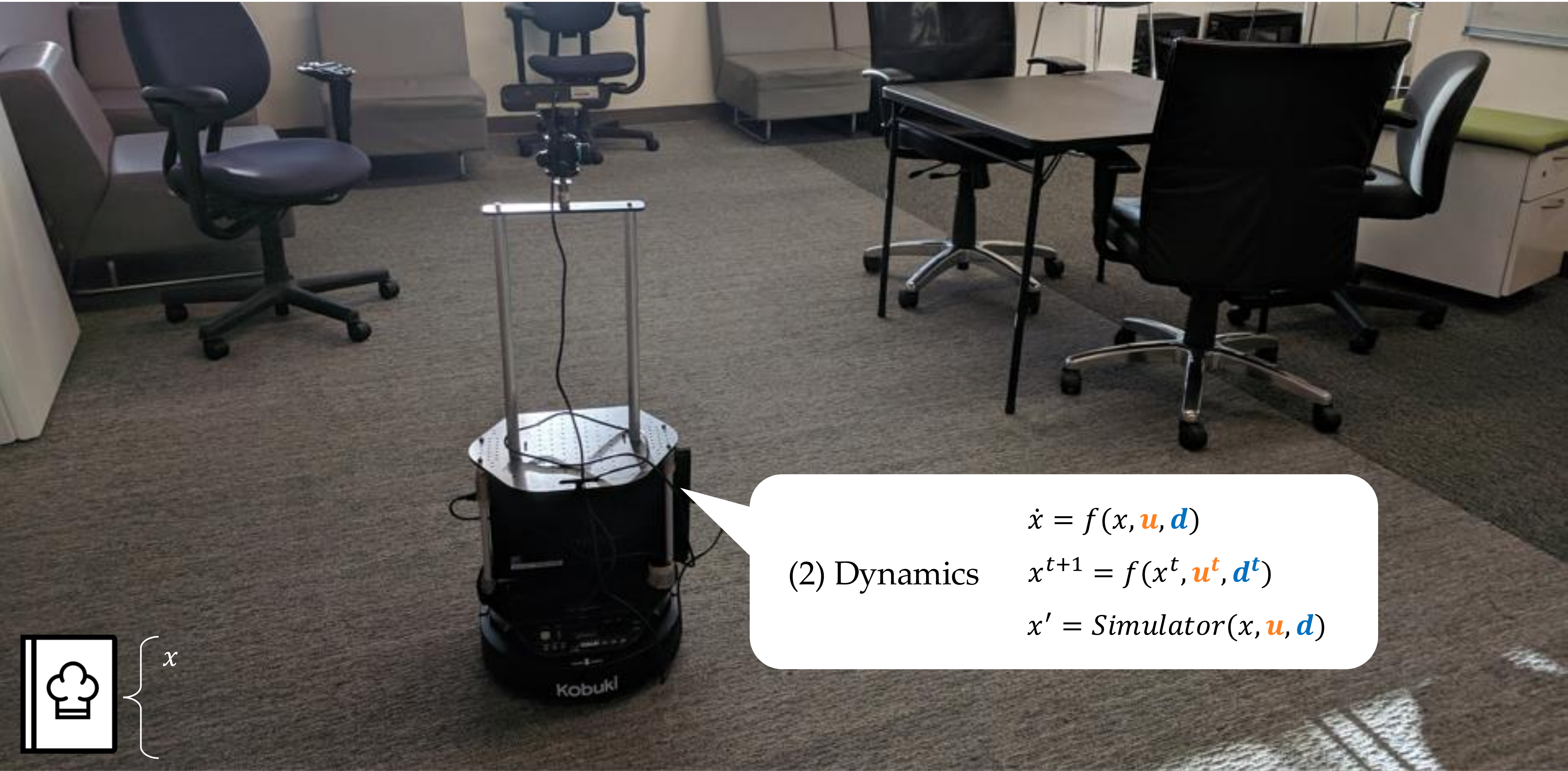

Kobuki

The Four Ingredients for Safety

(1) State Space $x \in \mathbb{R}^4$

xy-position, velocity, heading

The Four Ingredients for Safety



(2) Dynamics

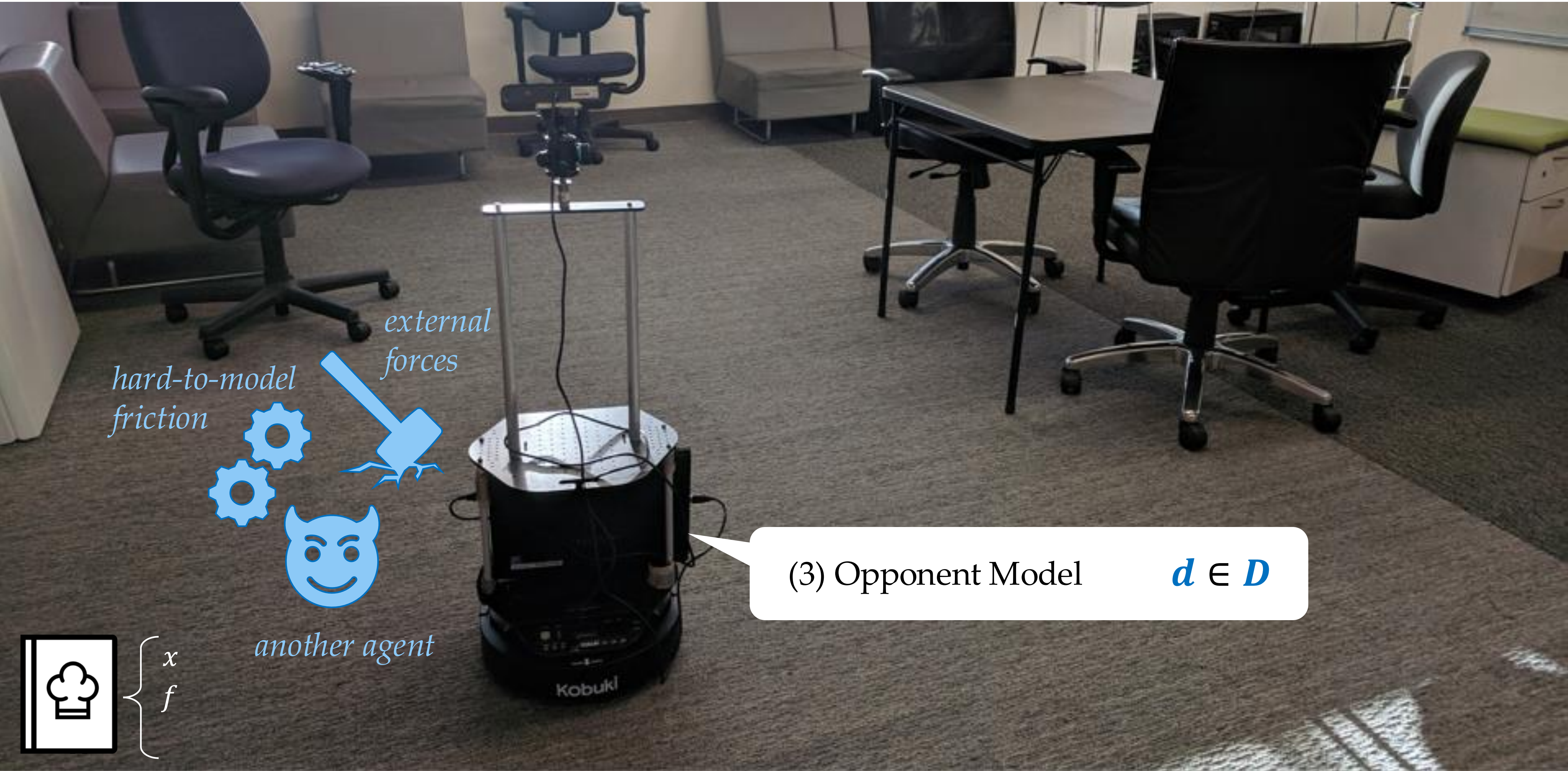
$$\dot{x} = f(x, \textcolor{orange}{\boldsymbol{u}}, \textcolor{blue}{\boldsymbol{d}})$$

$$x^{t+1} = f(x^t, \textcolor{orange}{\boldsymbol{u}^t}, \textcolor{blue}{\boldsymbol{d}^t})$$

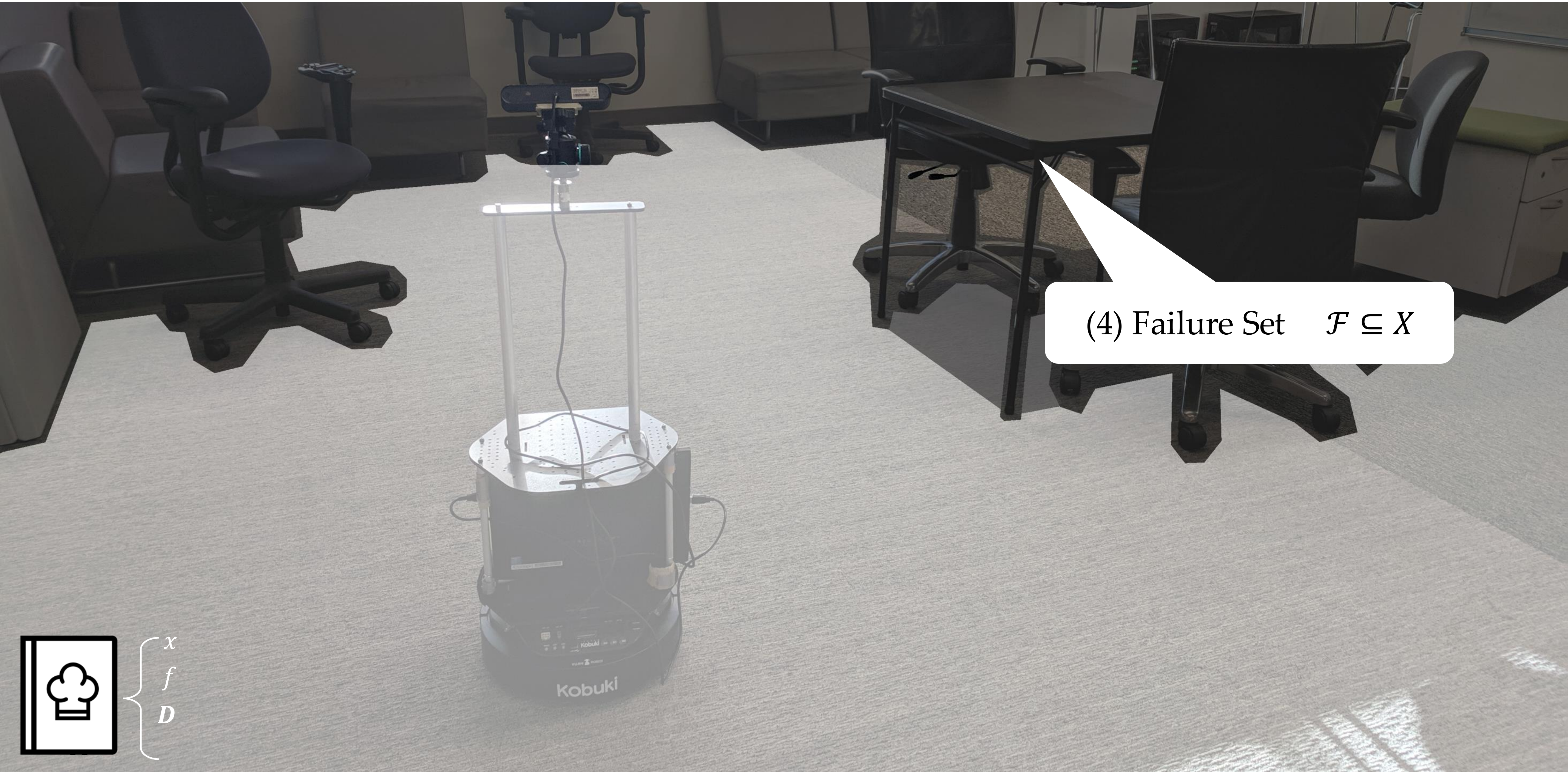
$$x' = Simulator(x, \textcolor{orange}{\boldsymbol{u}}, \textcolor{blue}{\boldsymbol{d}})$$

x

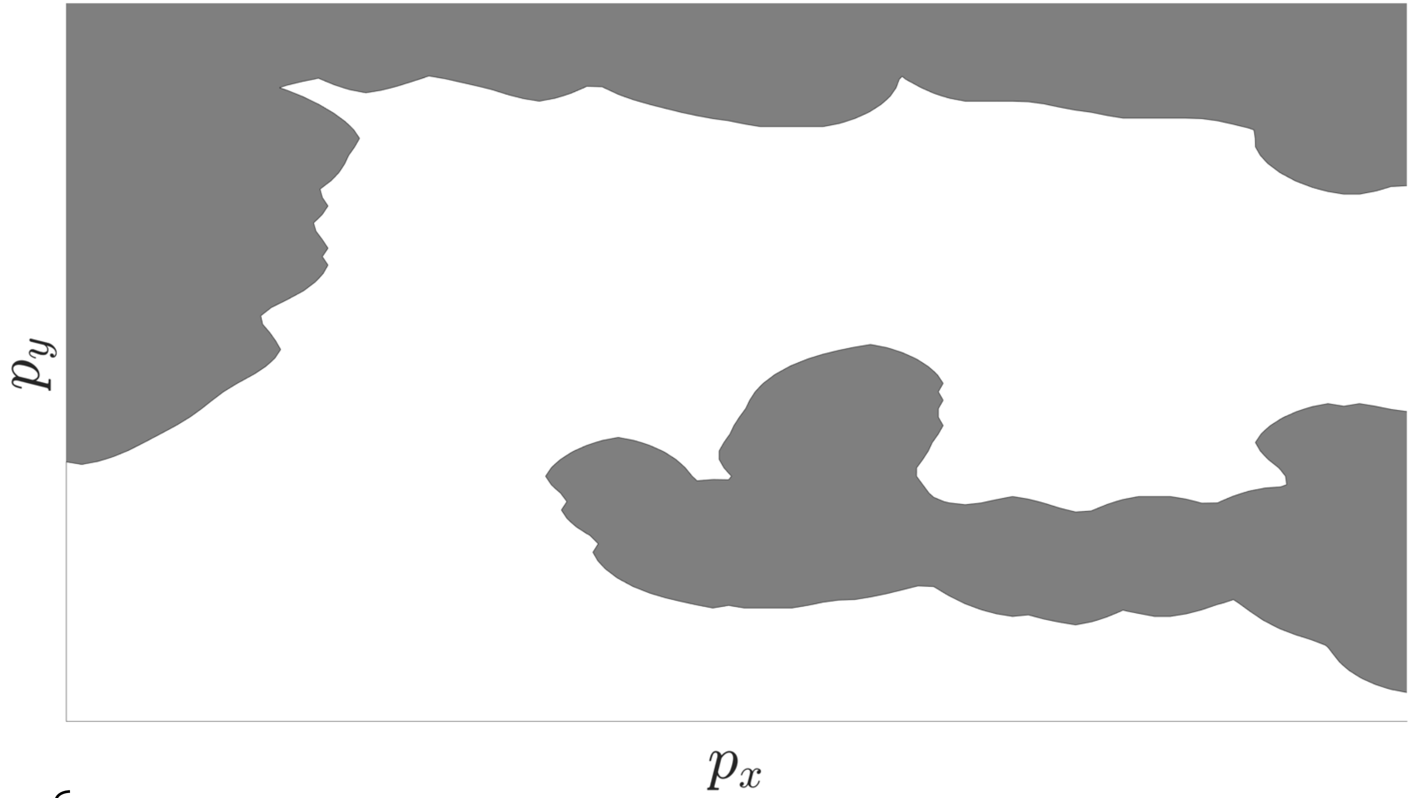
The Four Ingredients for Safety

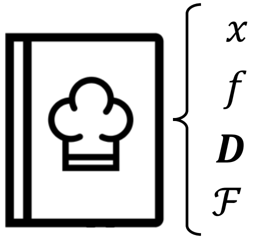


The Four Ingredients for Safety

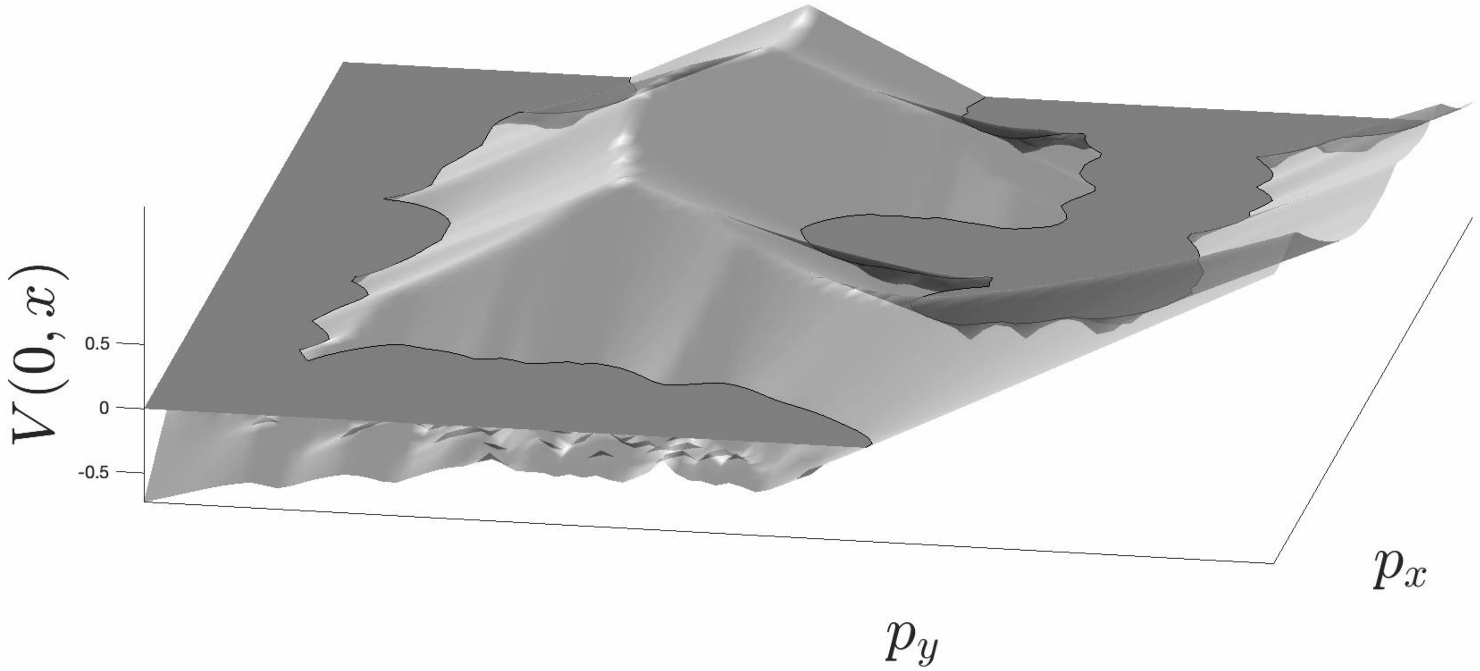


Let’s cook up a safety strategy!





Let's cook up a safety strategy!

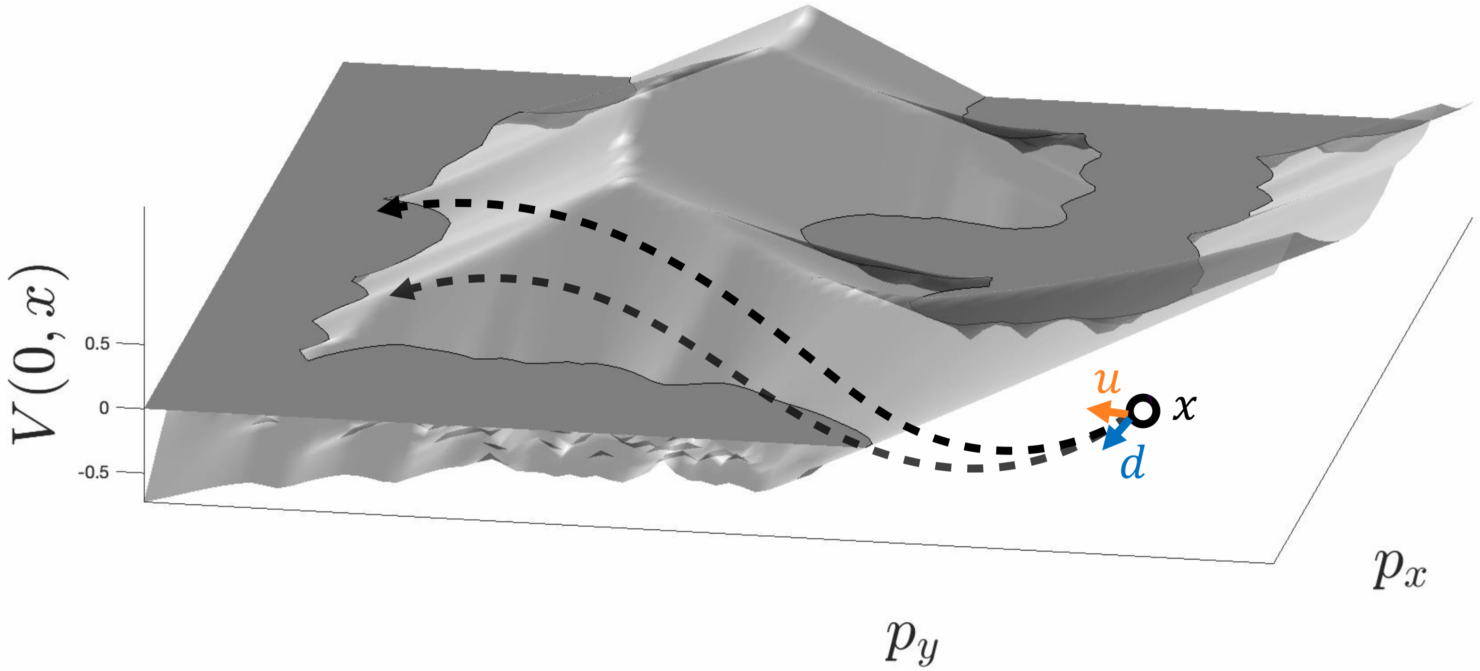


Encode Failure Set

$$\mathcal{F} = \{x : \ell(x) \leq 0\}$$

x
f
$\boldsymbol{D}$
$\mathcal{F}$

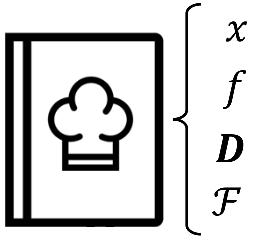
Let's cook up a safety strategy!



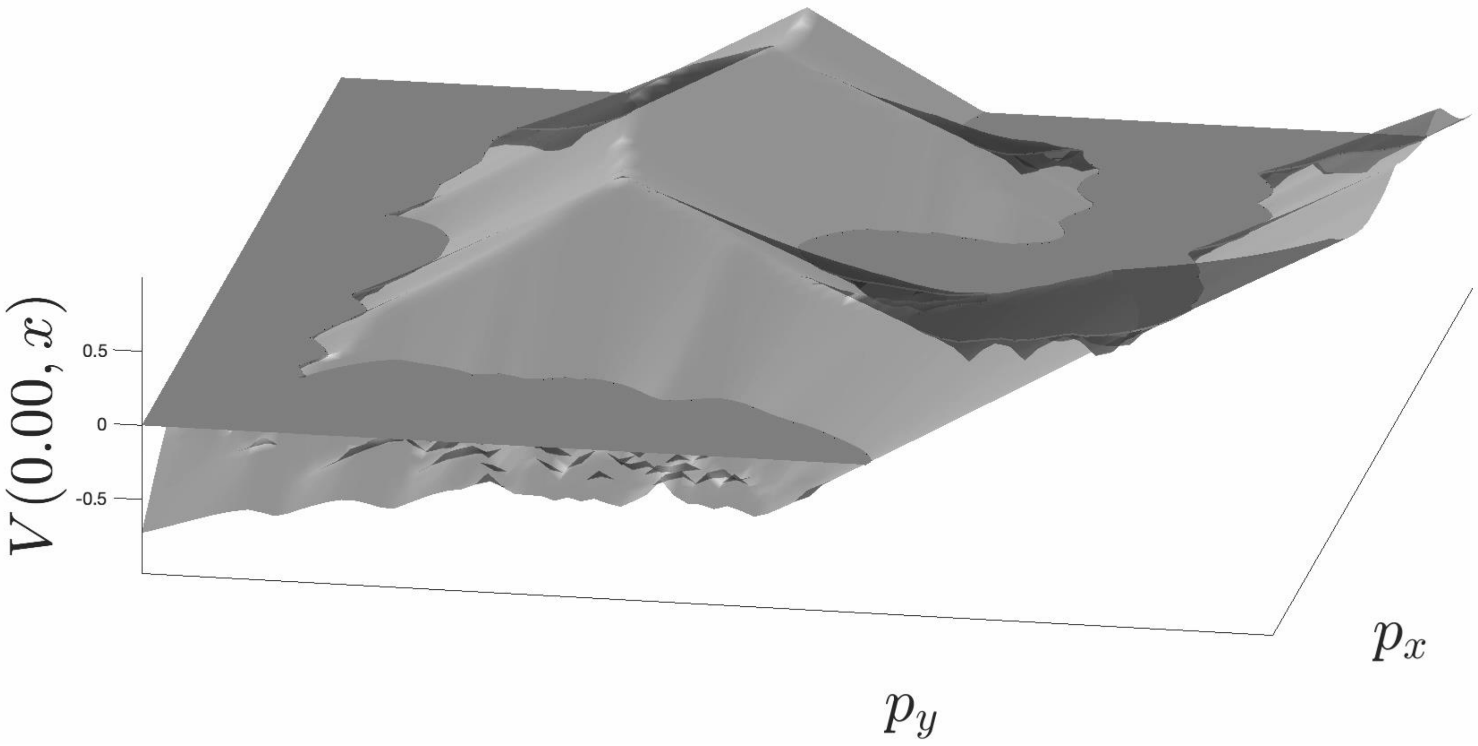
Pose Safety Critical Game

$$V(x) := \max_{\pi_u} \min_{\pi_d} \left(\min_{t \geq 0} \ell(\zeta_x^{u,d}(t)) \right)$$

V "remembers" the closest system got to failure under best robot strategy π_u *and worst opponent strategy* π_d



Let's cook up a safety strategy!



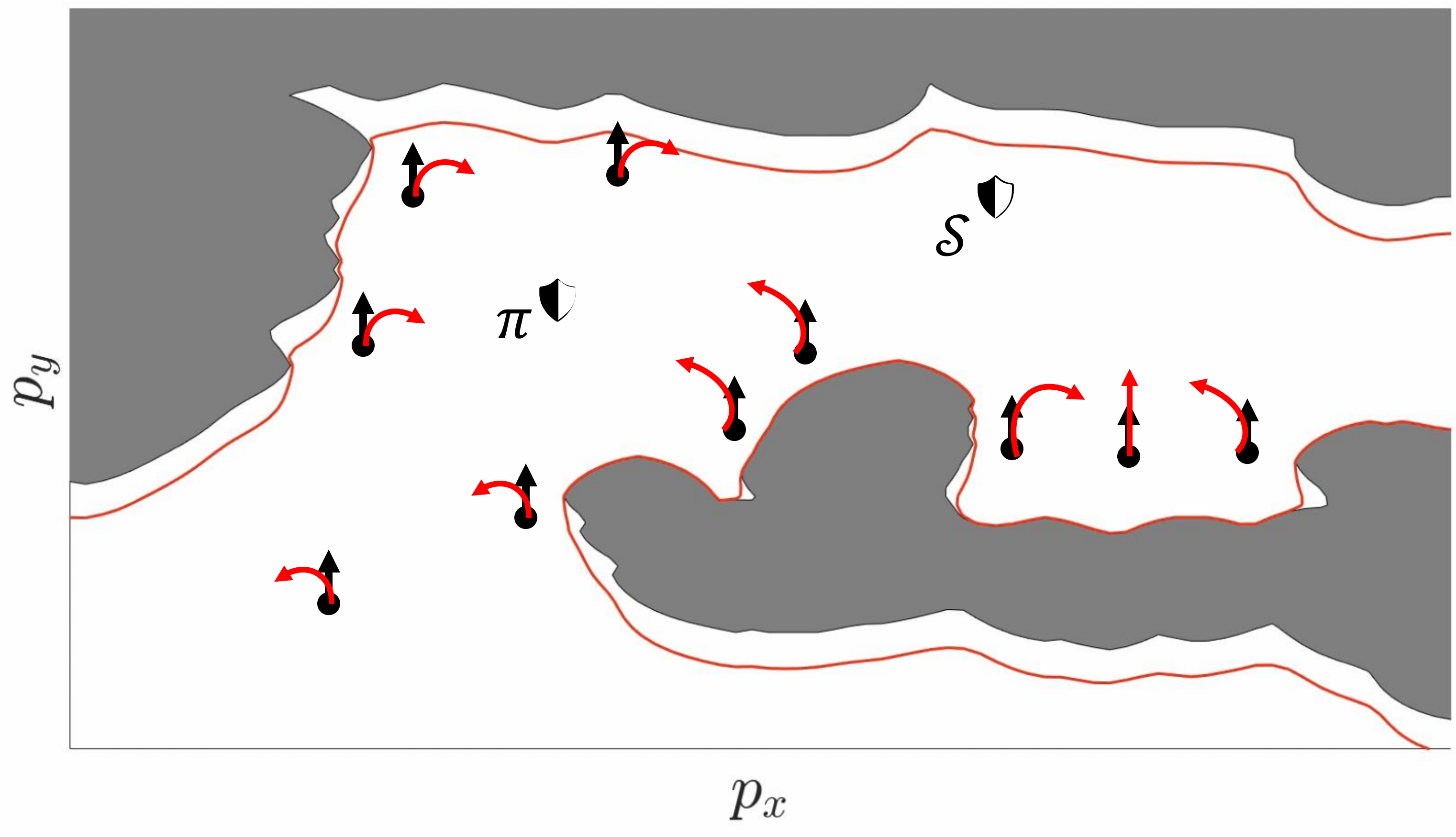
Solve Safety Game

$$V(x) := \max_{\pi_u} \min_{\pi_d} \left(\min_{t \geq 0} \ell(\zeta_x^{u,d}(t)) \right)$$

Many solvers: *exact grid-based PDE solvers [1], adversarial RL [2,3], self-supervised learning [4]*

[1] Mitchell, Journal of Scientific Computing 2008
[2] Pinto, et al. ICML 2017
[3] Hsu, et al. L4DC 2023
[4] Bansal & Tomlin, ICRA 2021

Let’s cook up a safety strategy!

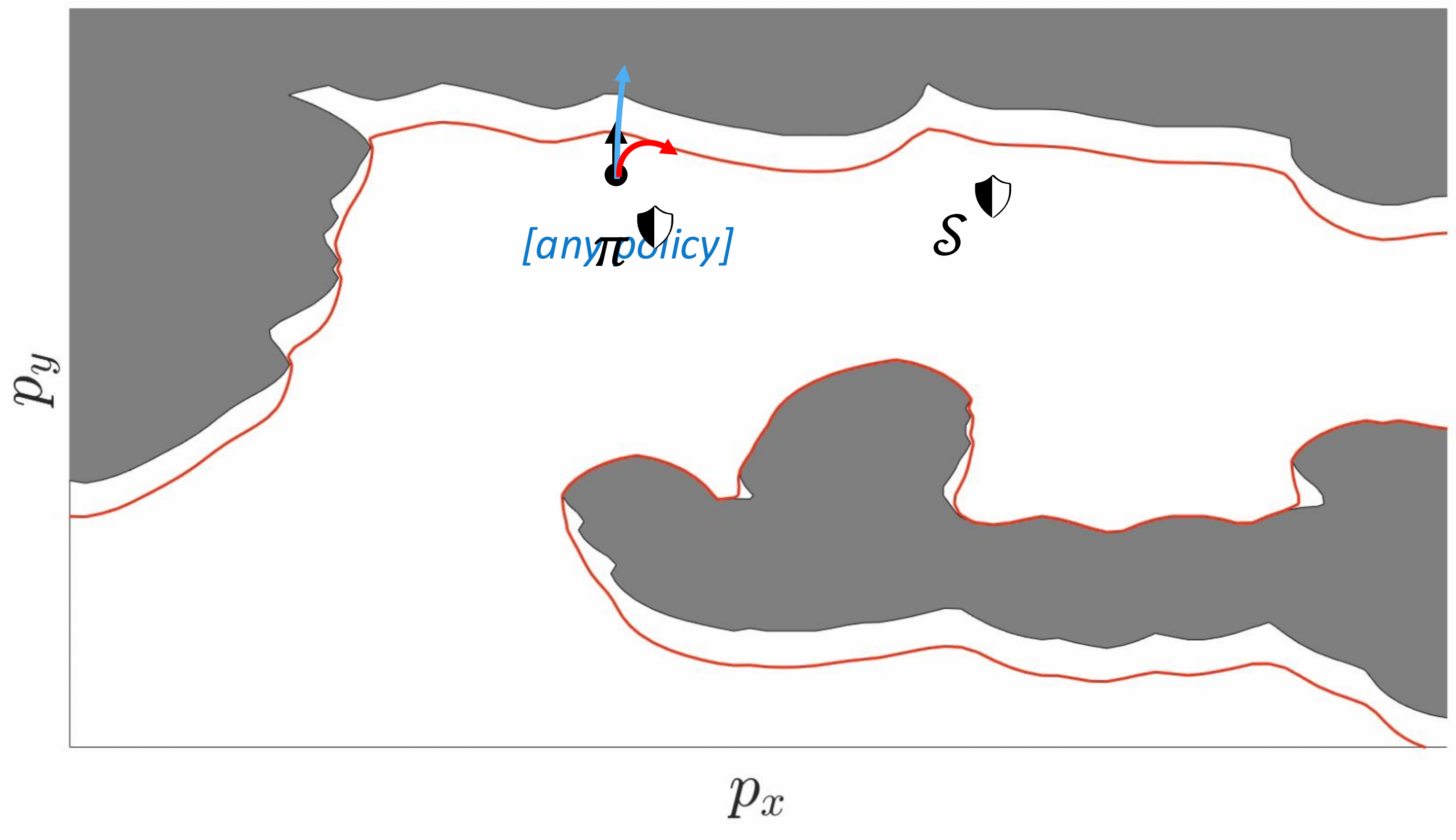


Safety ***Policy***

$$\pi^{◗}, \quad \mathcal{S}^{◗} = \{x : V(x) \geq 0\}$$

*Safe Set (i.e., “**Monitor**”)*

Let's cook up a safety strategy!



Safety ***Policy***

$$\pi^{\text{🛡}}, \quad \mathcal{S}^{\text{🛡}} = \{x : V(x) \geq 0\}$$

*Safe Set (i.e., "****Monitor****")*

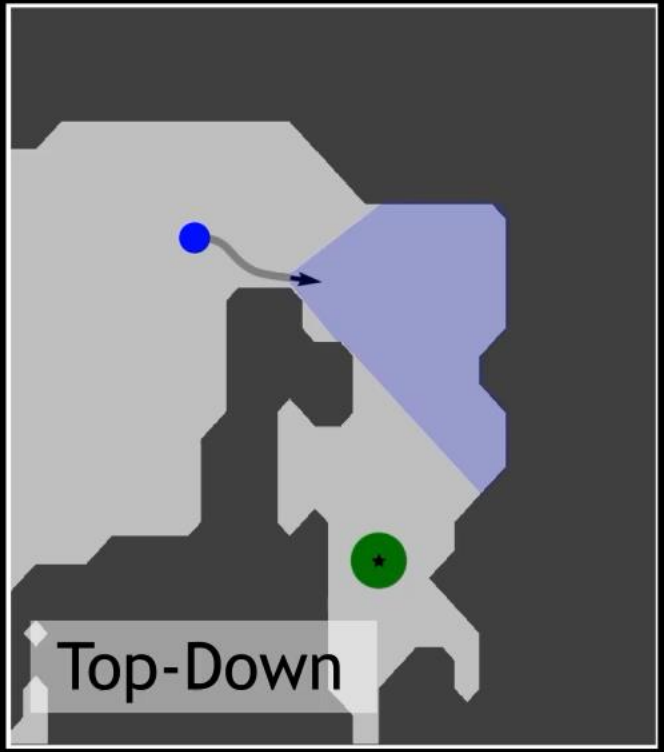
Safety ***Filter******

$$a = \begin{cases} \pi^{\text{🛡}}, & x \text{ near bdry } \mathcal{S}^{\text{🛡}} \\ \text{[any policy here]}, & x \in \mathcal{S}^{\text{🛡}} \end{cases}$$

**Note: there are many filtering variants!*

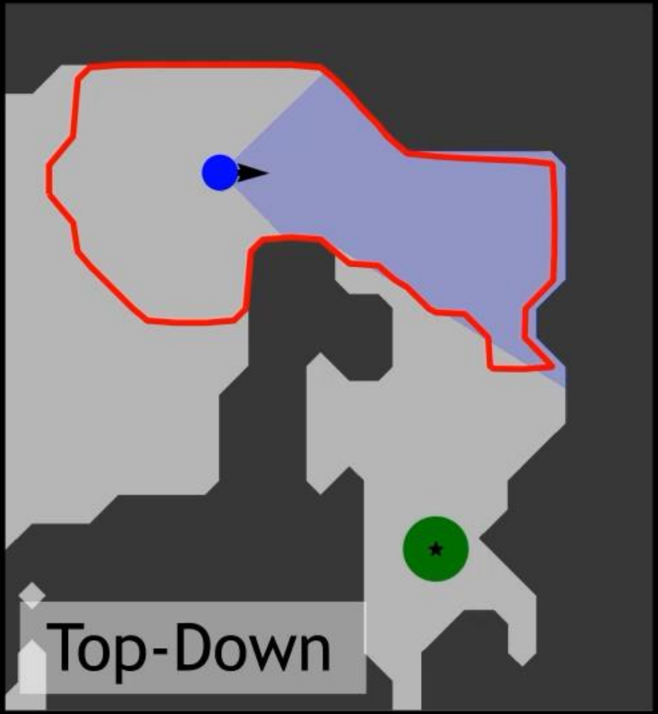
[Wabersich, et al. "Data-driven safety filters." Control Systems Magazine, 2023]
[Hsu, et al. "The Safety Filter." Annual Review of Control, Robotics, and Autonomous Systems, 2023]

Vision-Based Robot *Without* Safety Strategy



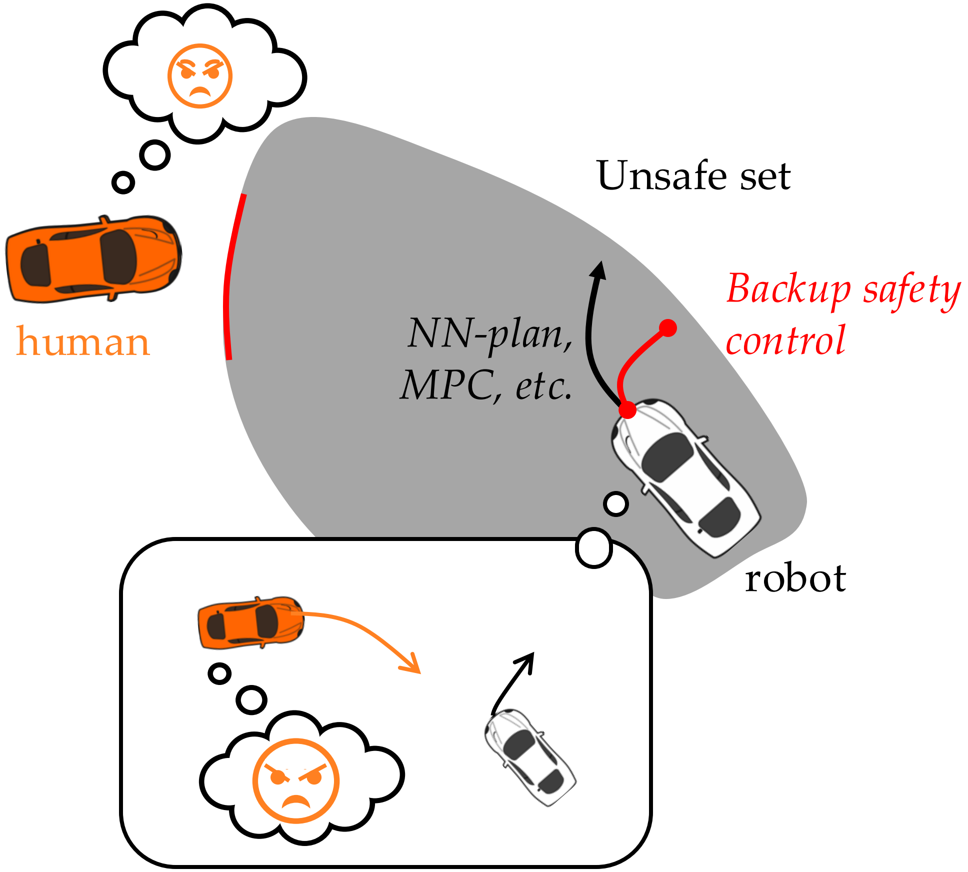
[Bajcsy*, et al. CDC 2019]

Robot *With* Safety Strategy



[Bajcsy*, et al. CDC 2019]

Safety strategies applied to interaction …



Zero-Sum Dynamic Game

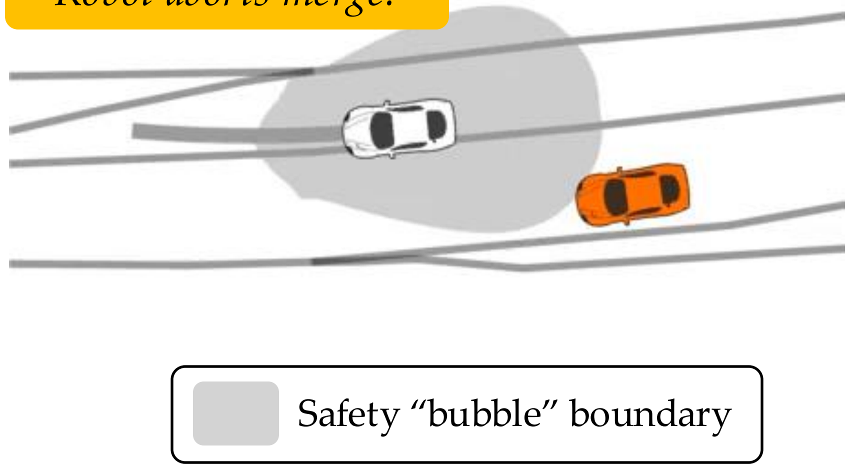
Safety Game

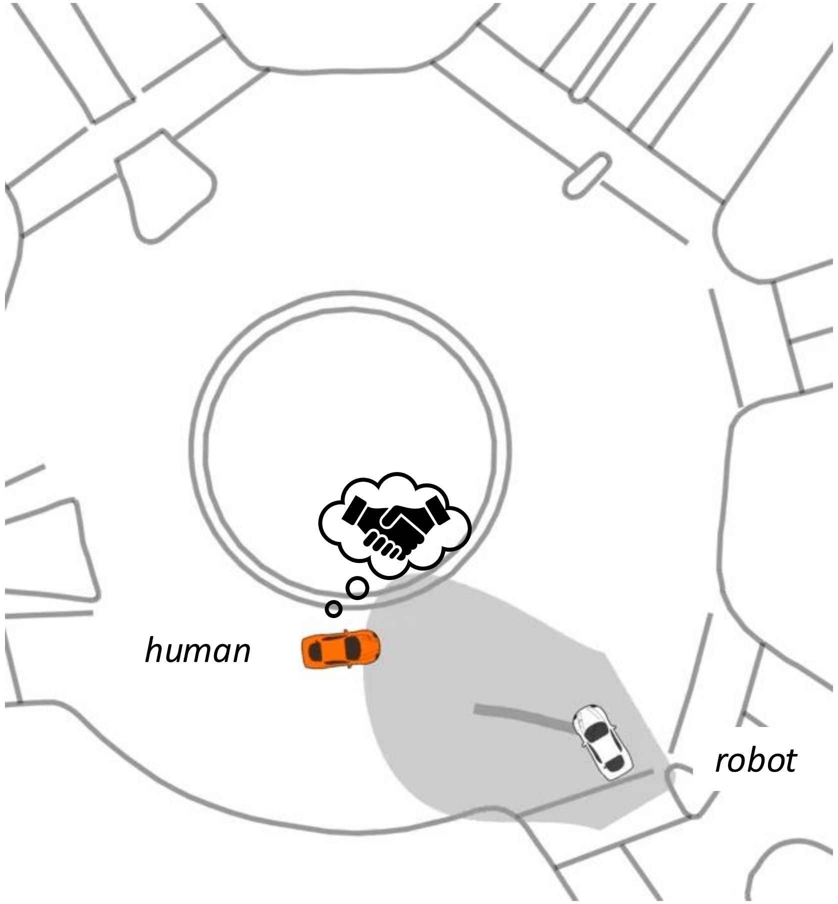
$$V(x) := \max_{\pi_R} \min_{\pi_H} \left(\min_{t \geq 0} \ell(\zeta_x^{u_R, u_H}(t)) \right)$$

Zero-sum games give us robustness but…

Without much knowledge of the real world, traditional safety strategies can be too pessimistic

Robot aborts merge!

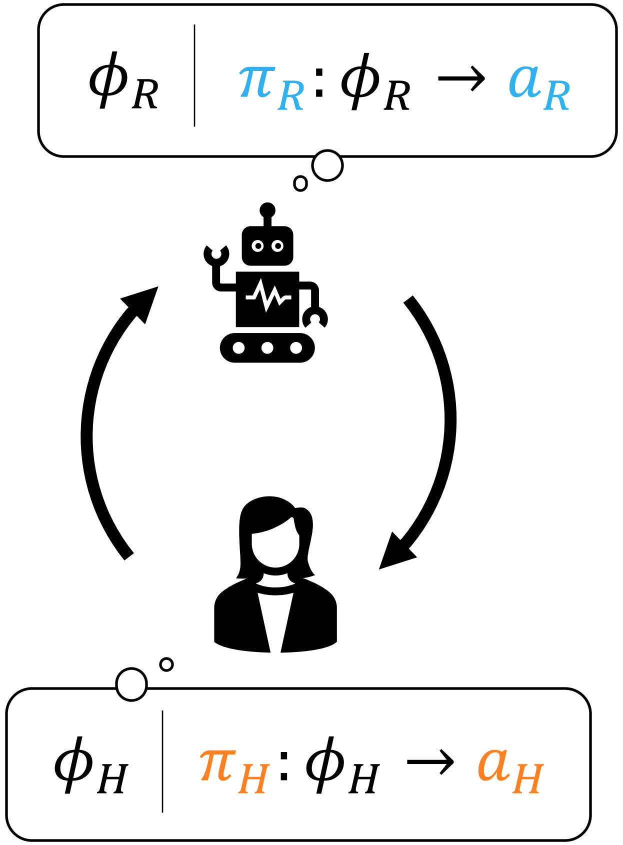




Even if human yields, robot serves around

1 How can we **formalize** interactive robot safety?

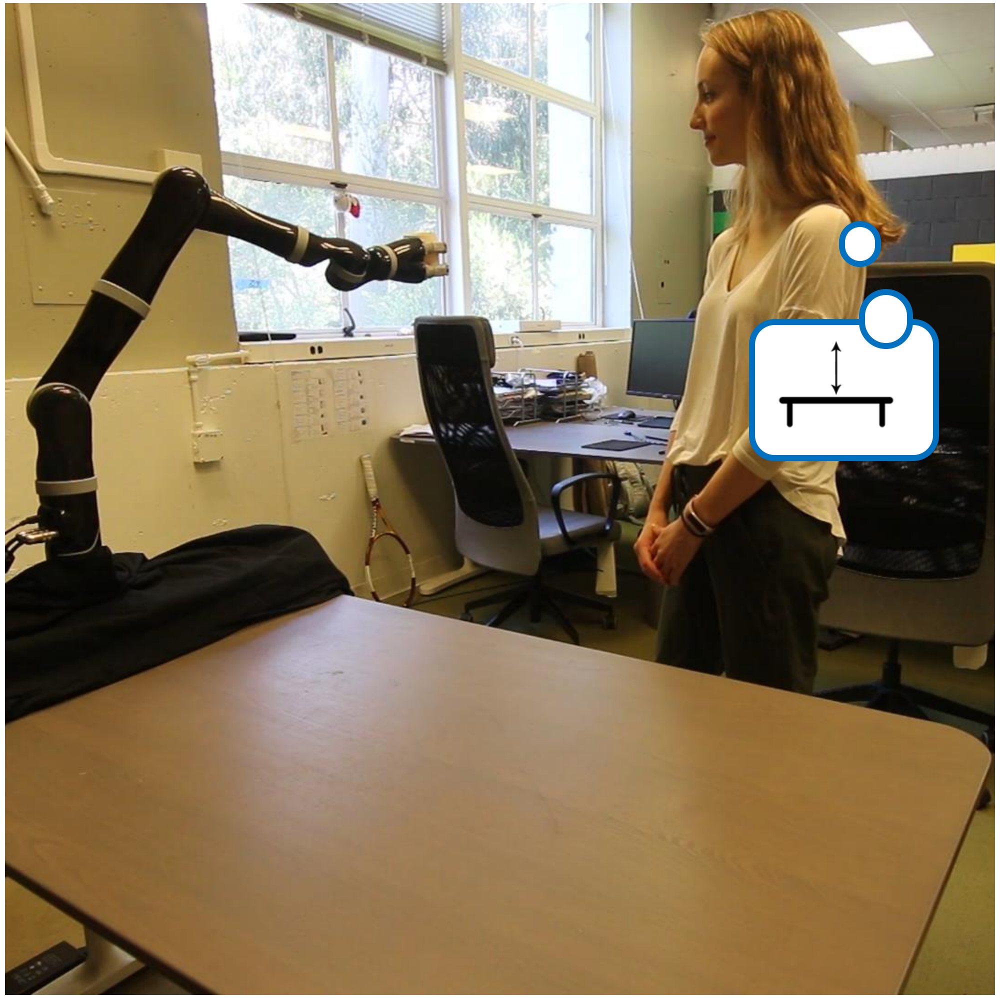
2 How can robots adapt their **safety strategies under uncertainty**?



3 How can robots learn safety representations from humans?

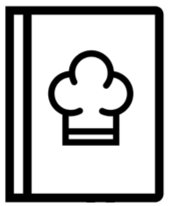
Good to be robust, but humans aren't always adversaries

[Bajcsy* & Fisac* et al, RSS 2018]



[Bobu, Bajcsy, et al. T-RO 2020]

Let's use zero-sum games for robustness, but inform them with (data-driven) human predictions



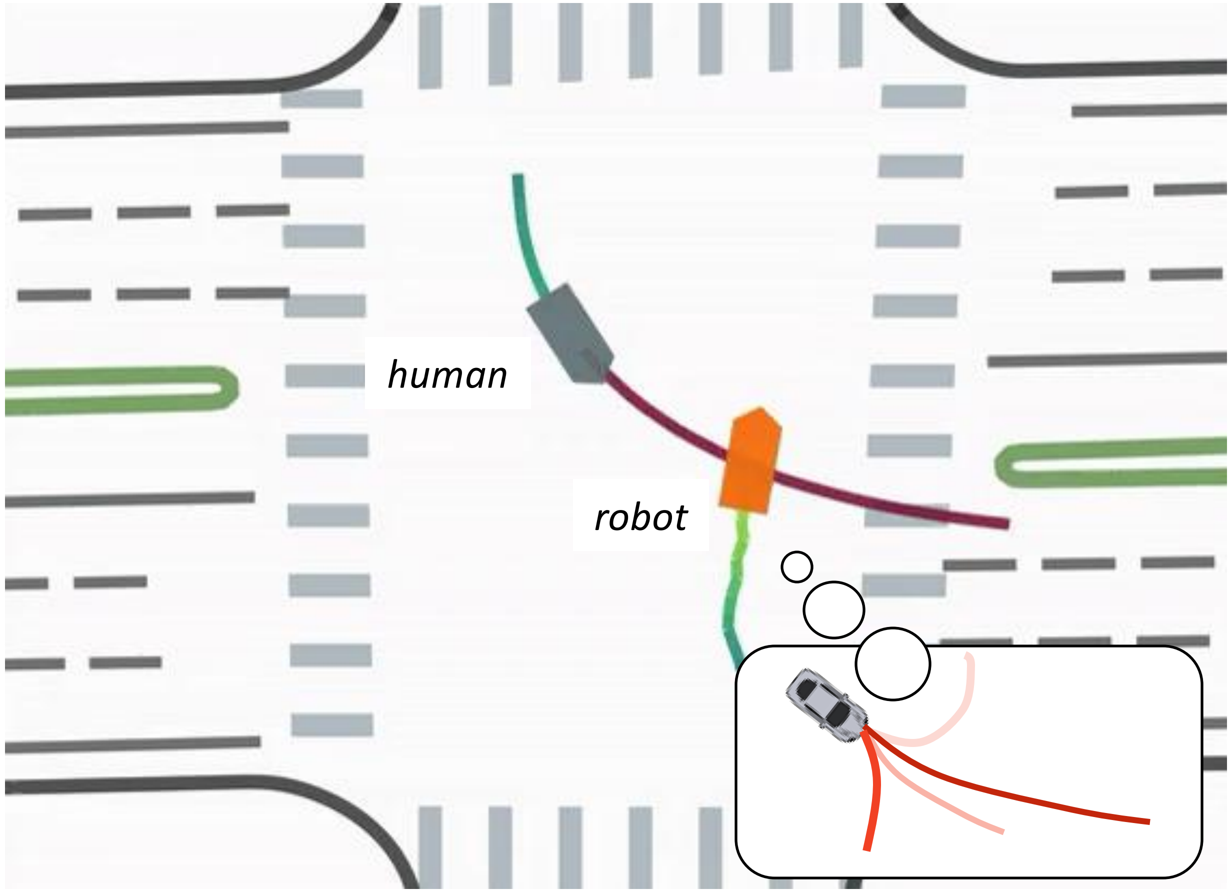
x

f

$\boldsymbol{U_H}$ *informed via prediction models*

$\mathcal{F}$

Human prediction-informed robot safety strategies



Metric/Method	Opponent policy	Robust (w/o learning)	Deception Game (ours)
Failure rate Completion time (s)	Modeled ($\epsilon_\theta = 0.2$)	**0 %** 6.27 ± 0.86	**0 %** 4.73 ± 0.97

[Hu*, Zhang*, Nakamura, Bajcsy, Fisac. "Deception Game." CoRL 2023]

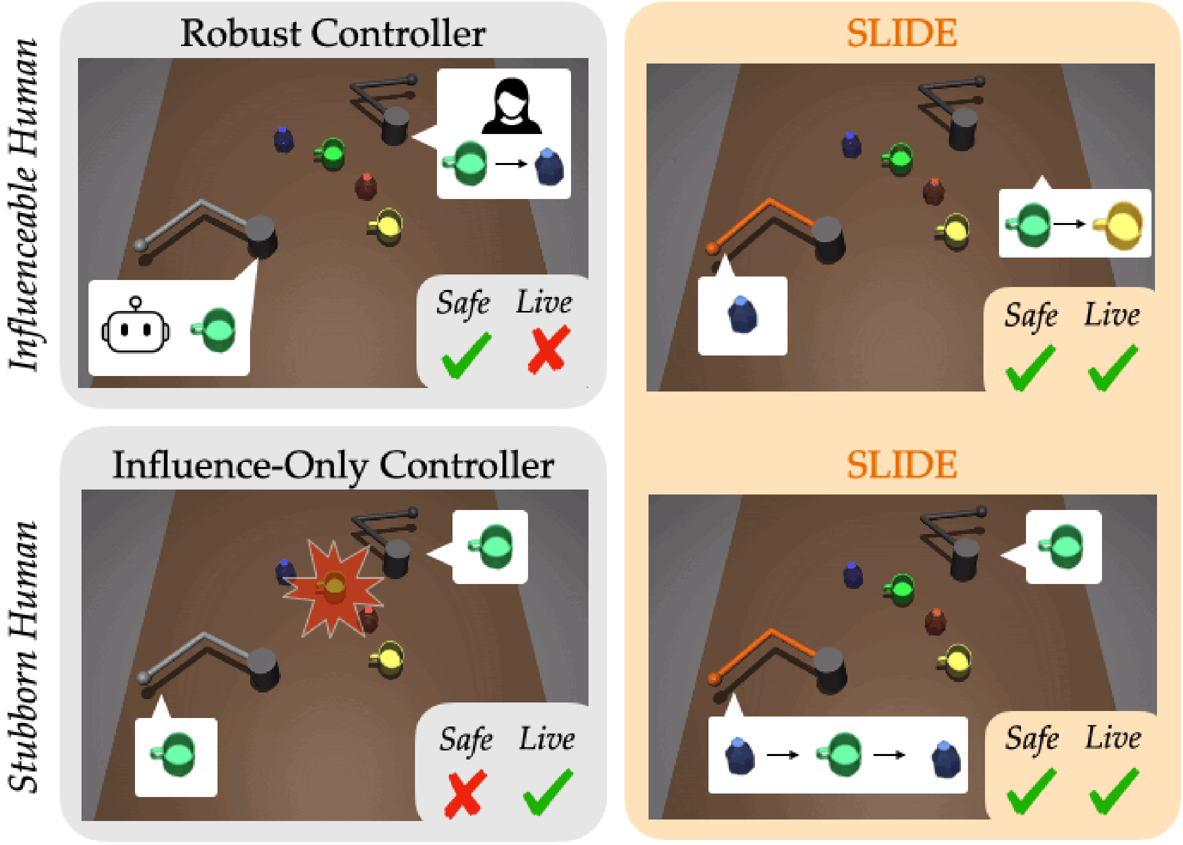
Robustness via zero-sum game, but actions we safeguard against are informed via *human behavior model*

$$V(x) \coloneqq \max_{\pi_R} \min_{\pi_H \in \Pi_H} \left(\min_{t \geq 0} \ell(\zeta_x^{u_R, u_H}(t)) \right)$$

$$\Pi_H \coloneqq \{u_H : \underbrace{P(u_H | x^{\text{hist}})}_{\text{human behavior model}} \geq \epsilon\}$$

Predict complex scene-conditioned behavior via **Motion Transformer** [Shi et al. NeurIPS, 2022]

Human prediction-informed robot safety strategies



	Collision rate	Completion rate	Completion Time (s)
NoSafety	28.5%	71.5%	3.5 ± 1.8
SSA	19.1%	52.3%	8.9 ± 4.7
Robust-RA	1.4%	97.0%	2.6 ± 2.1
Marginal-RA	1.5%	98.0%	2.5 ± 1.3
SLIDE (ours)	1.9%	98.1%	1.9 ± 0.8

[Pandya, Liu, Bajcsy. arXiv 2025 *(ICRA submission)*]

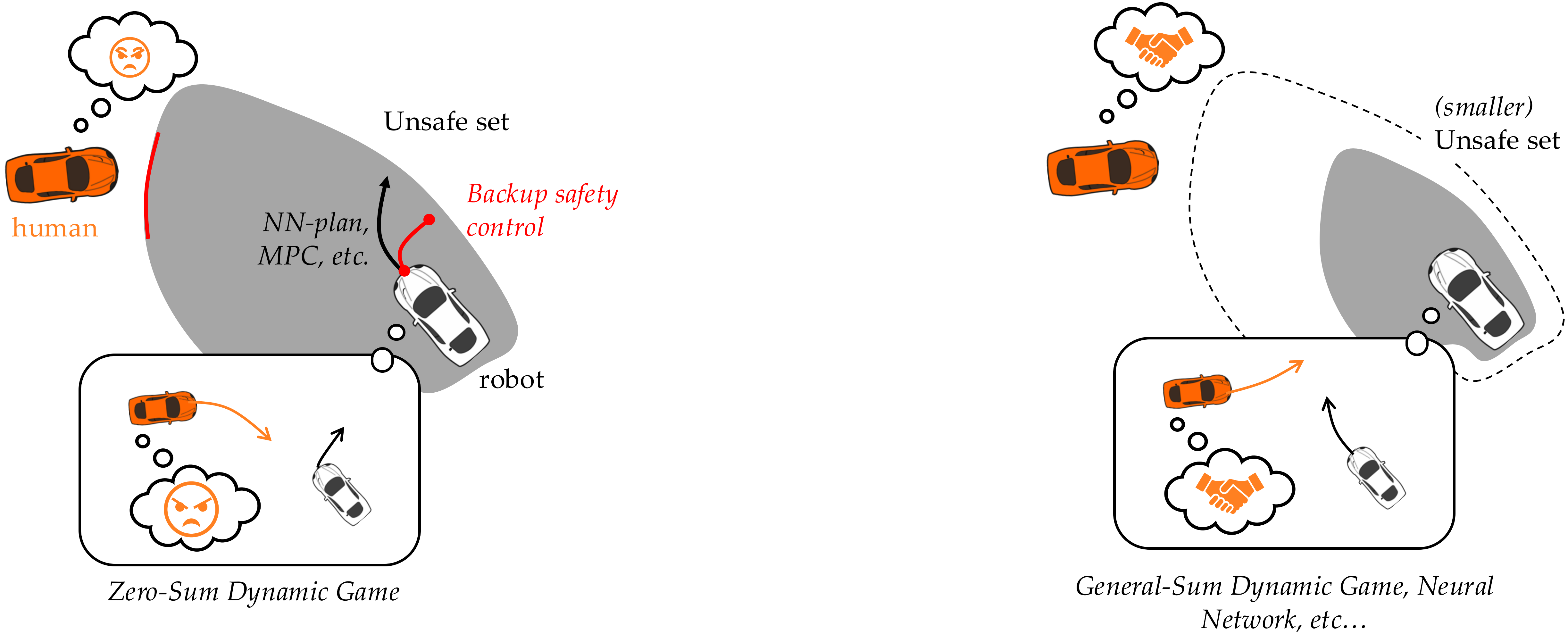
Robustness via zero-sum game, but actions we safeguard against are informed via *human behavior model*

$$V(x) := \max_{\pi_R} \min_{\pi_H \in \Pi_H} \left(\min_{t \geq 0} \ell(\zeta_x^{u_R, u_H}(t)) \right)$$

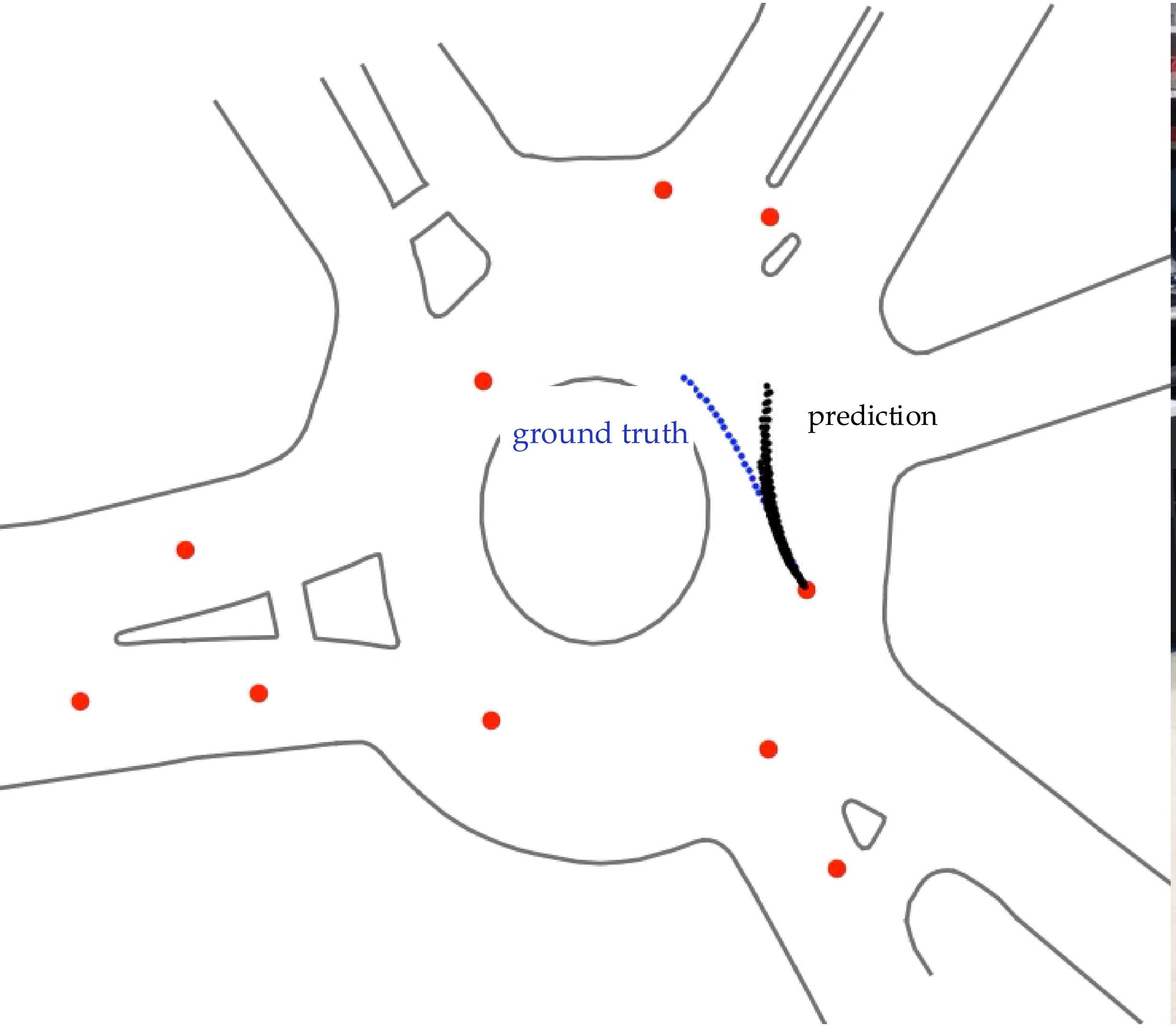
$$\Pi_H\left(u_R^{\text{plan}}\right) := \{u_H : P\left(u_H \middle| u_R^{\text{plan}}, x^{\text{hist}}\right) \geq \epsilon\}$$

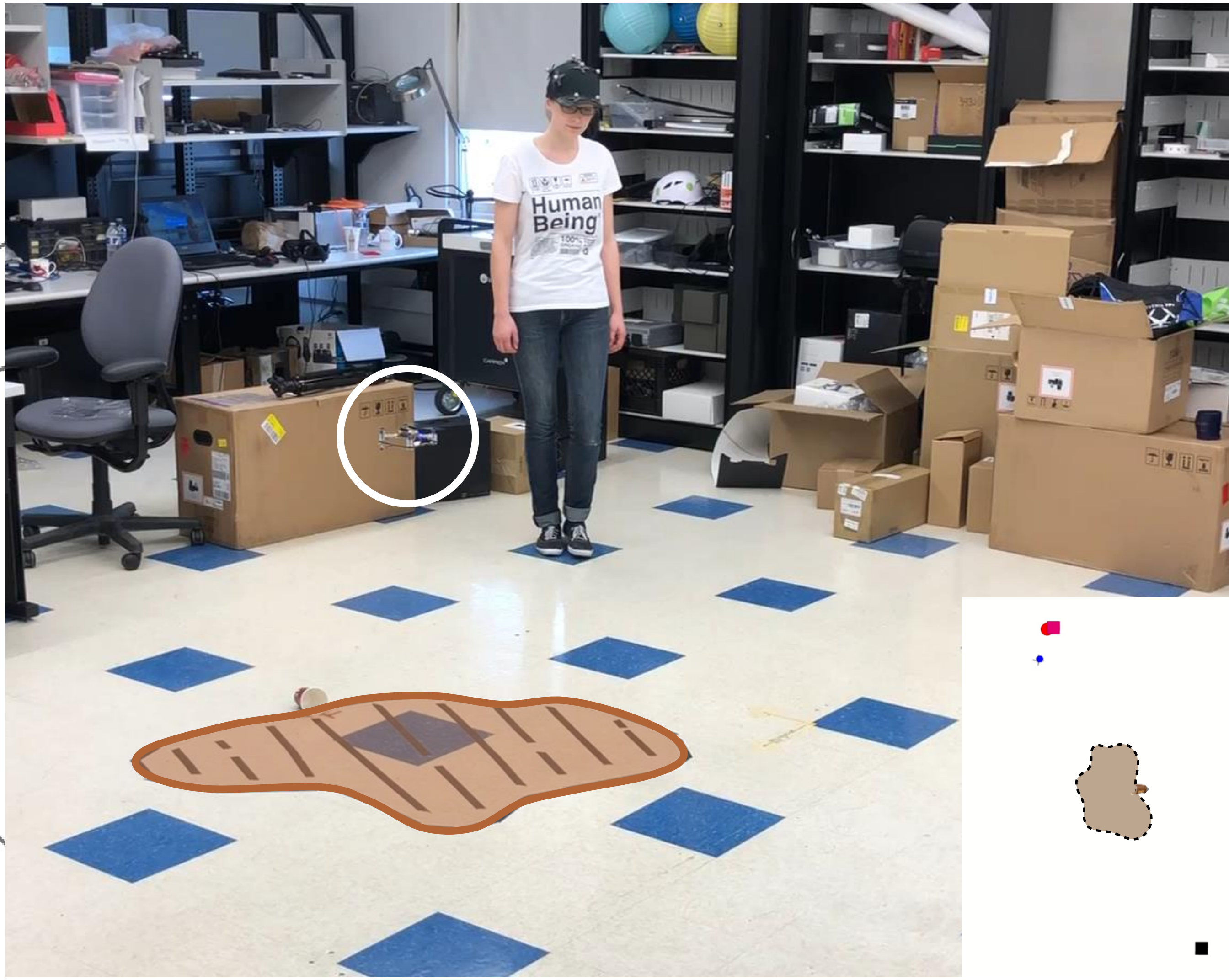
Model robot influence via
Conditional Behavior Predictors

Safety strategies applied to interaction …



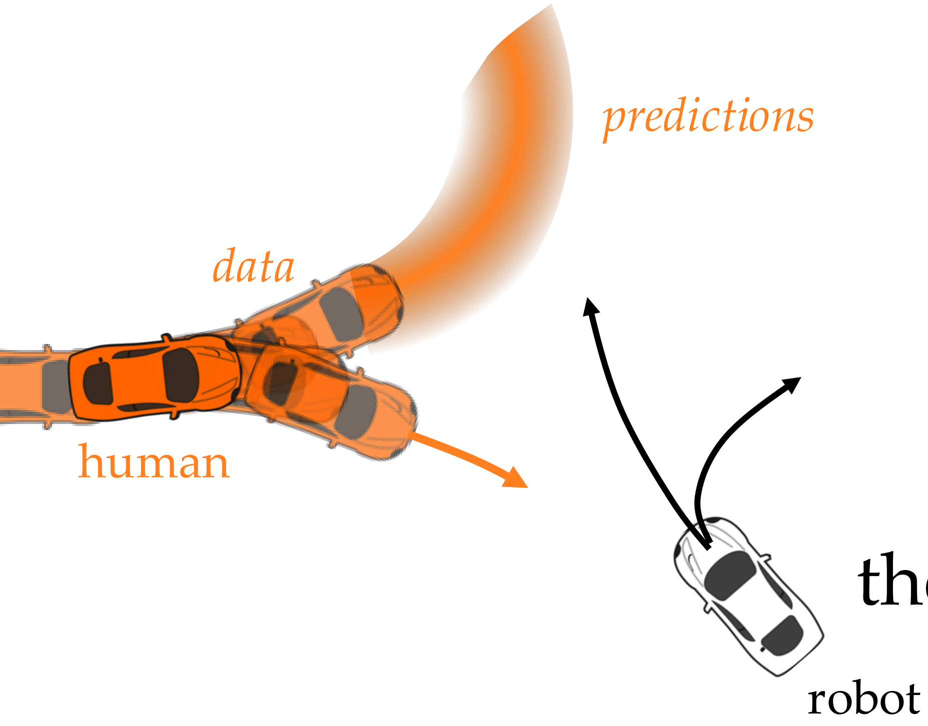
Data-driven models can fail under out-of-distribution human interactions





[Sun et al, 2021]

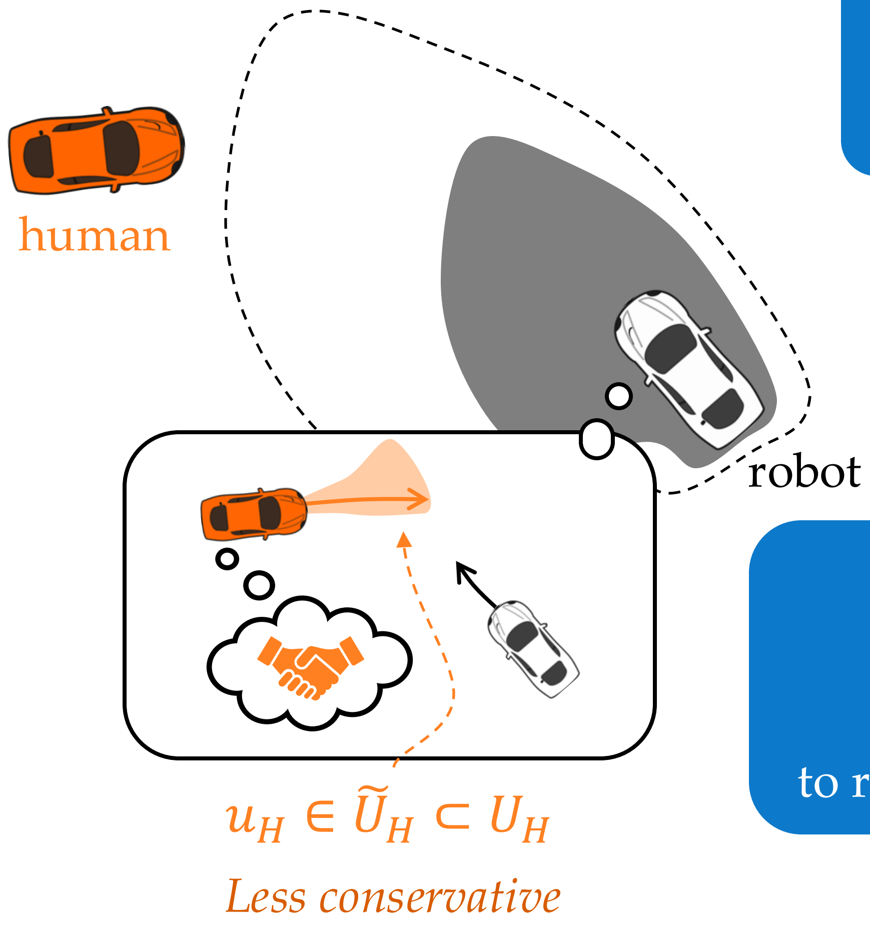
[Bajcsy* & Fisac* et al, RSS 2018]



Idea:

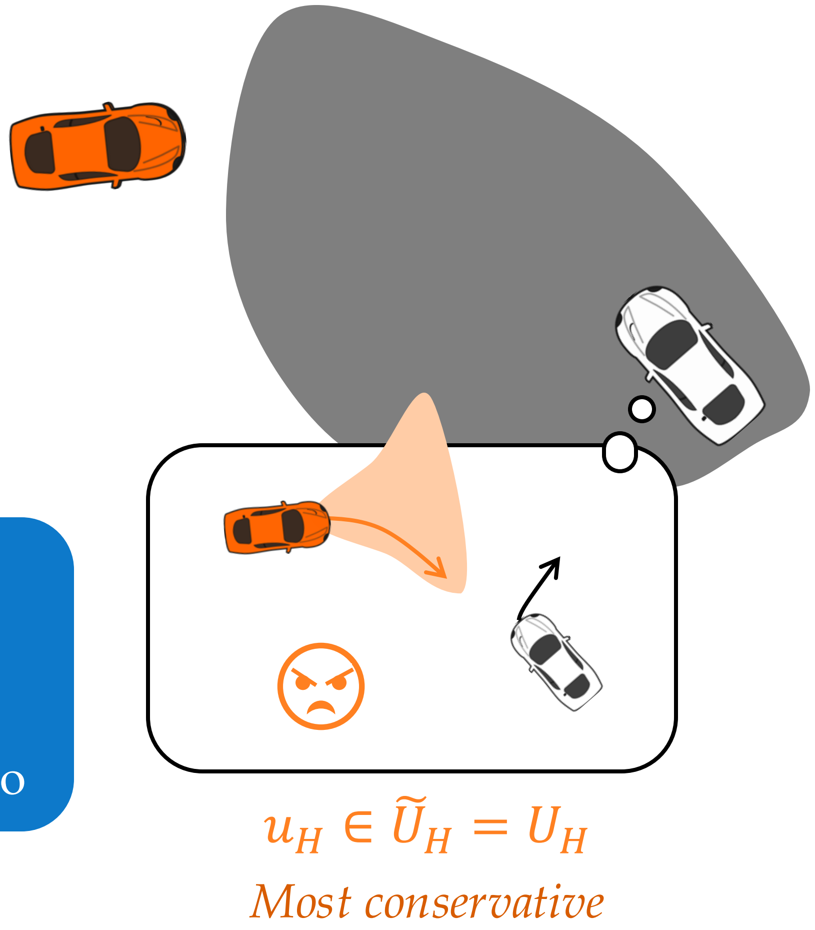
Use the human data observed online to adapt the conservativeness of the robot’s safety strategy

Confidence-Aware Game-Theoretic Safety Strategies



Idea 2:
Online, infer the robot's *uncertainty* in predicting human

Idea 1:
Use human behavior model
$P(\mathbf{u_H} \mid \boldsymbol{x}, \mathbf{u_R})$
to restrict set of actions robot is robust to



[Tian*, Sun*, Bajcsy*, et al. "Safety Assurances for Human-Robot Interaction via Confidence-aware Game-theoretic Human Models", ICRA 2022]

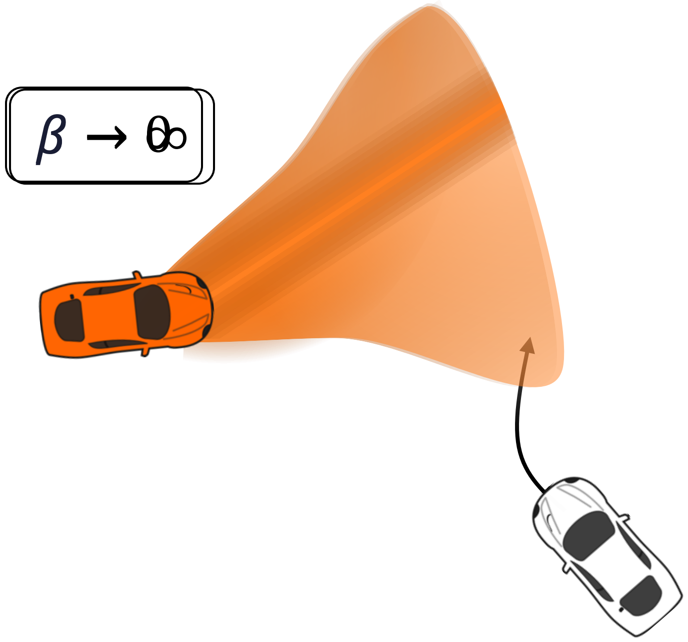
Example: Stackelberg Game Predictor

Human traj. Robot traj.

$$P(\mathbf{u}_H | x^0, \mathbf{u}_R; \lambda, \quad) \propto \begin{cases} e^{R_H(x^0, \mathbf{u}_H, \mathbf{u}_R)} & \text{if } \lambda = \text{follower} \\ e^{R_H(x^0, \mathbf{u}_H, \mathbf{u}_R^*(\mathbf{u}_H))} & \text{if } \lambda = \text{leader} \end{cases}$$

Joint state Human's role

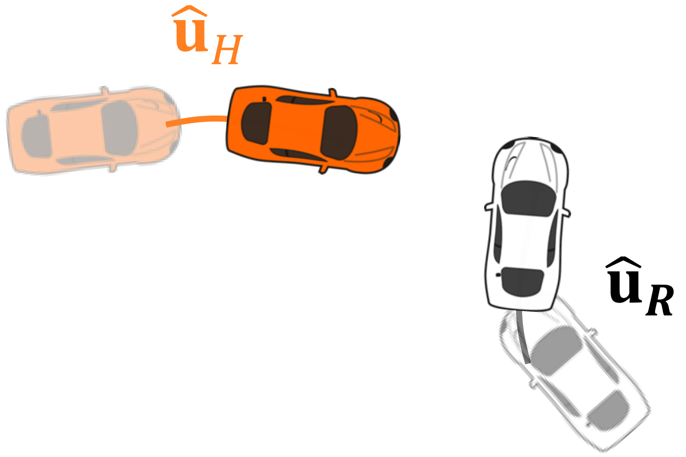
Human's rationality
(also referred to as *model confidence*)



[Sadigh, et a., RSS 2016], [Schwarting et al, PNAS 2019], [Fisac et al, ICRA 2019], …
[Fisac et al RSS 2018], [Bajcsy et al, ICRA 2019], [Carreno-Medrano et al, RO-MAN 2019]

Infer confidence of the predictor & role of human in game

$$P(\mathbf{u}_{\mathrm{H}} | x^0, \mathbf{u}_{\mathrm{R}}; \lambda, \beta) \dashrightarrow b^{t+1}(\beta, \lambda) \propto P(\hat{\mathbf{u}}_H | x^0, \hat{\mathbf{u}}_R; \lambda, \beta) b^t(\beta, \lambda)$$



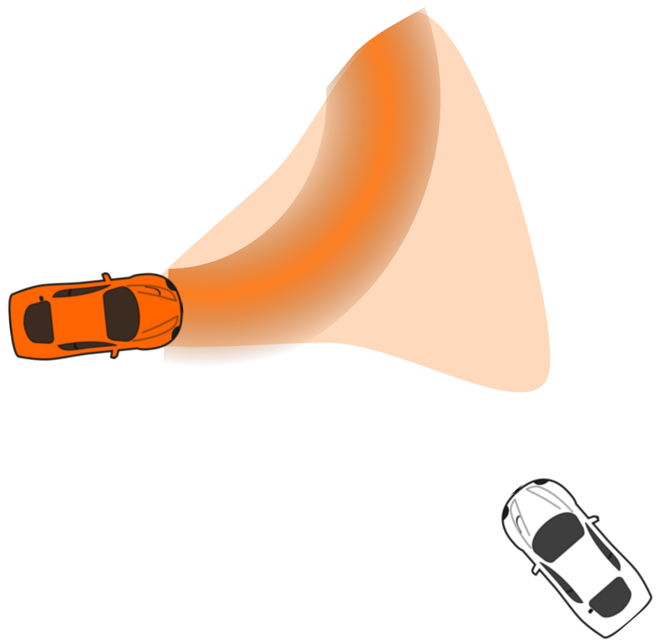
Infer confidence of the predictor

$$P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}; \lambda, \beta) \quad \dashrightarrow \quad b^{t+1}(\beta, \lambda) \propto P(\hat{\mathbf{u}}_H | x^0, \hat{\mathbf{u}}_R; \lambda, \beta) b^t(\beta, \lambda)$$

Online update of the robot's safety strategy

Predict likely human trajectories.

$$P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}) = \mathbb{E}_{\beta,\lambda} P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}; \lambda, \beta)$$



$$P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}; \lambda, \beta) \dashrightarrow b^{t+1}(\beta, \lambda) \propto P(\hat{\mathbf{u}}_H | x^0, \hat{\mathbf{u}}_R; \lambda, \beta) b^t(\beta, \lambda)$$

Infer confidence of the predictor

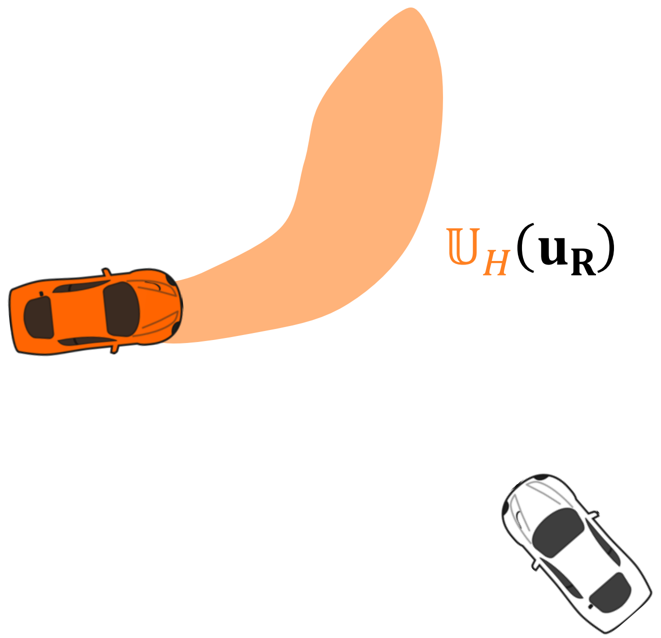
Online update of the robot's safety strategy

Predict likely human trajectories.

$$P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}) = \mathbb{E}_{\beta,\lambda} P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}; \lambda, \beta)$$

Set of sufficiently likely control trajectories.

$$\mathbb{U}_H(\mathbf{u}_\mathrm{R}) = \{\mathbf{u}_\mathrm{H} : P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}) > \epsilon\}$$



Infer confidence of the predictor

$$P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}; \lambda, \beta) \dashrightarrow b^{t+1}(\beta, \lambda) \propto P(\hat{\mathbf{u}}_H | x^0, \hat{\mathbf{u}}_R; \lambda, \beta) b^t(\beta, \lambda)$$

Online update of the robot’s safety strategy

Predict likely human trajectories.

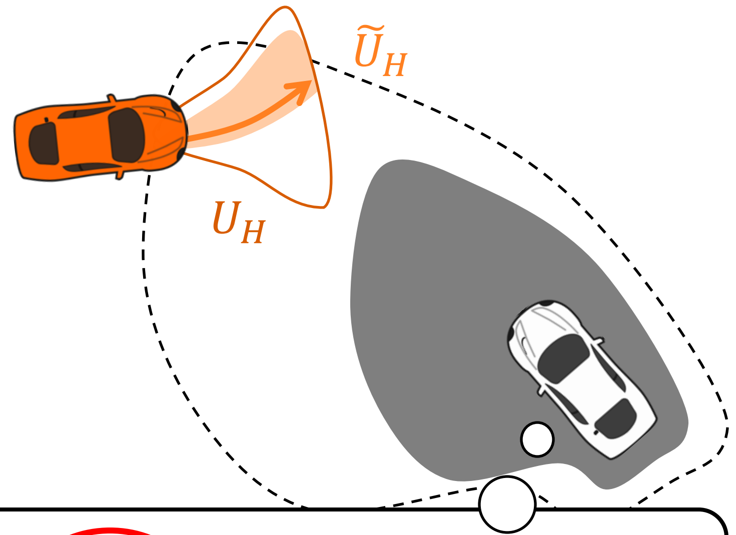
$$P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}) = \mathbb{E}_{\beta,\lambda} P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}; \lambda, \beta)$$

Set of sufficiently likely control trajectories.

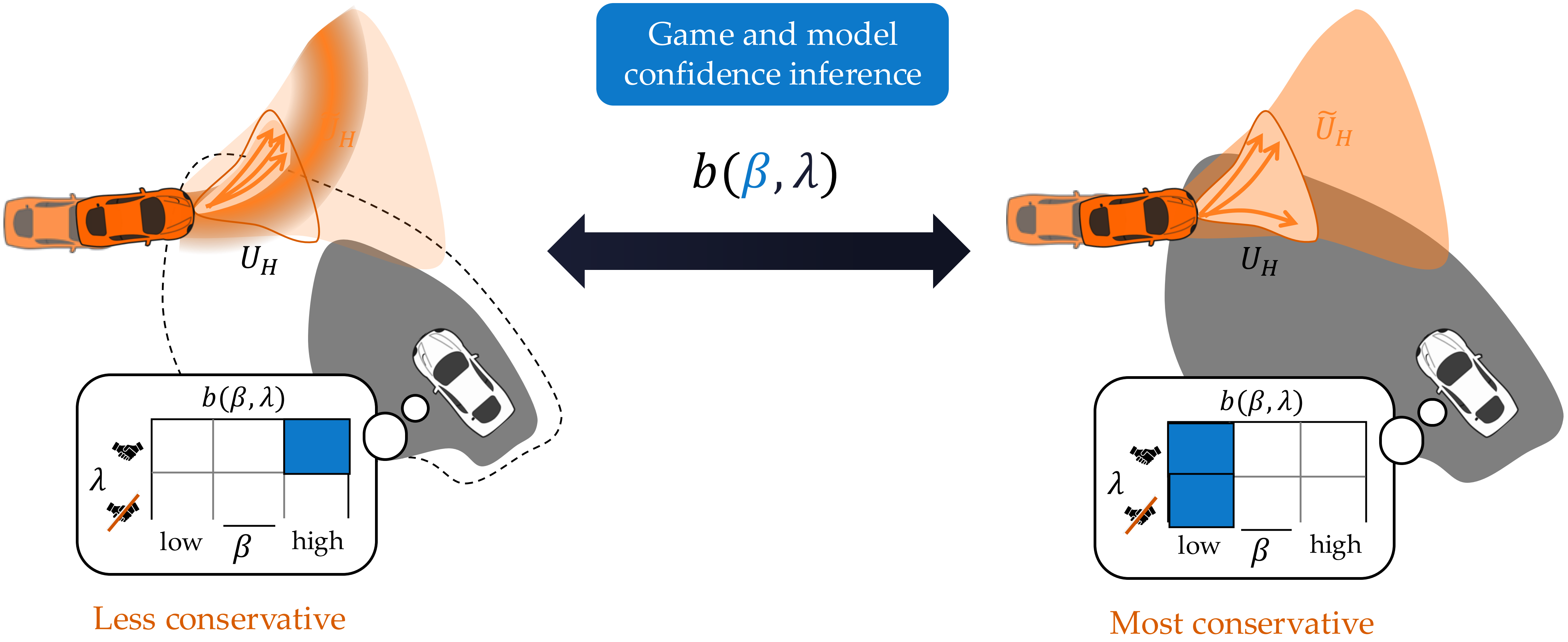
$$\mathbb{U}_H(\mathbf{u}_\mathbf{R}) = \{\mathbf{u}_\mathbf{H} : P(\mathbf{u}_\mathrm{H} | x^0, \mathbf{u}_\mathrm{R}) > \epsilon\}$$

New control bounds for safety monitor.

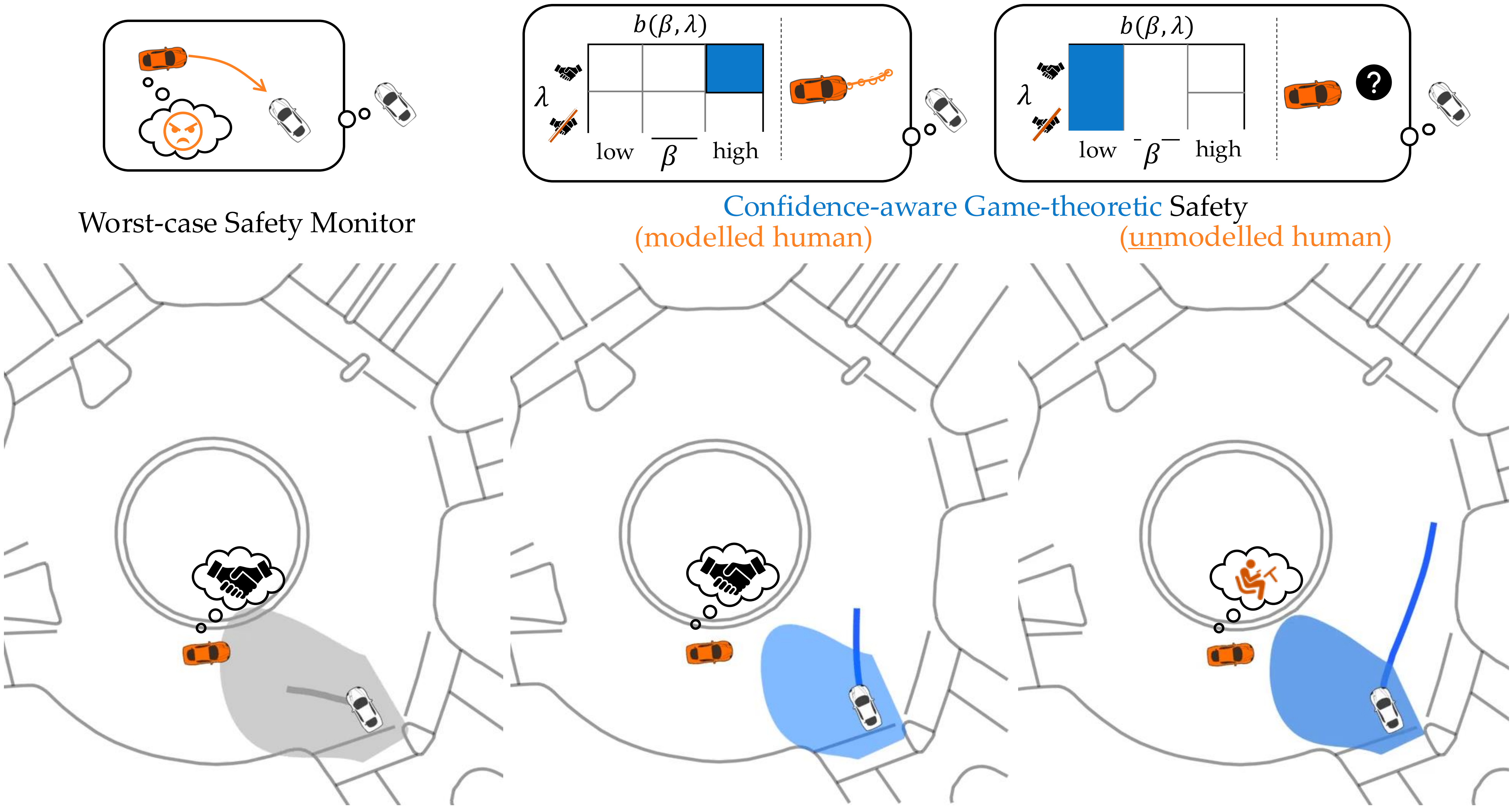
$$\tilde{U}_H := [\underline{u_H(\mathbf{u}_\mathbf{R})}, \overline{u_H(\mathbf{u}_\mathbf{R})}]$$



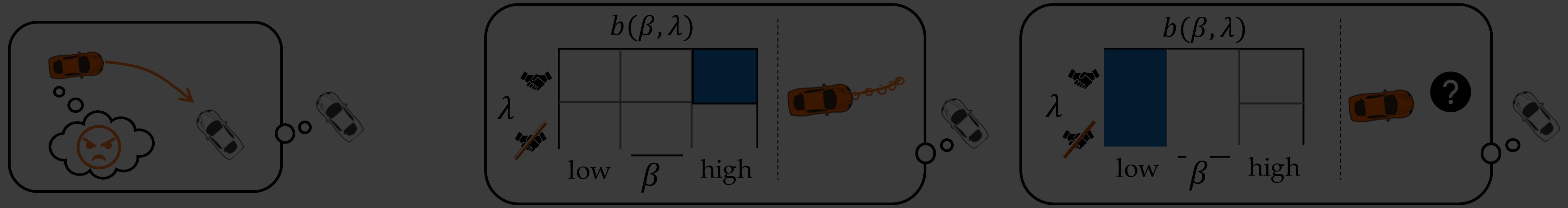
$$V(x) := \max_{\pi_R} \min_{\pi_H \in \Pi_H} \left(\min_{t \geq 0} \ell(\zeta_x^{u_R, u_H}(t)) \right)$$



Game and model confidence inference
$b(\beta, \lambda)$
U_H
$\tilde{U}_H$
$b(\beta, \lambda)$
λ
low
β
high
Less conservative
Most conservative



$b(\beta,\lambda)$
λ
low
β
high
$b(\beta,\lambda)$
λ
low
β
high
Worst-case Safety Monitor
Confidence-aware Game-theoretic Safety
(modelled human)
(unmodelled human)



Worst-case Safety Monitor

Confidence-aware Game-theoretic Safety

(modelled human)

(unmodelled human)

	Worst-case Safety		Confidence-aware Game-theoretic Safety		
Human type	CR	SOR	CR	SOR	RIP(Full)
modeled	0	23.3	0	4.7	**27.75 ± 4.03**
noisy	0	29.8	0	7.3	**18.26 ± 3.96**
unmodeled	0	42.1	0	41.7	0.06 ± 0.19

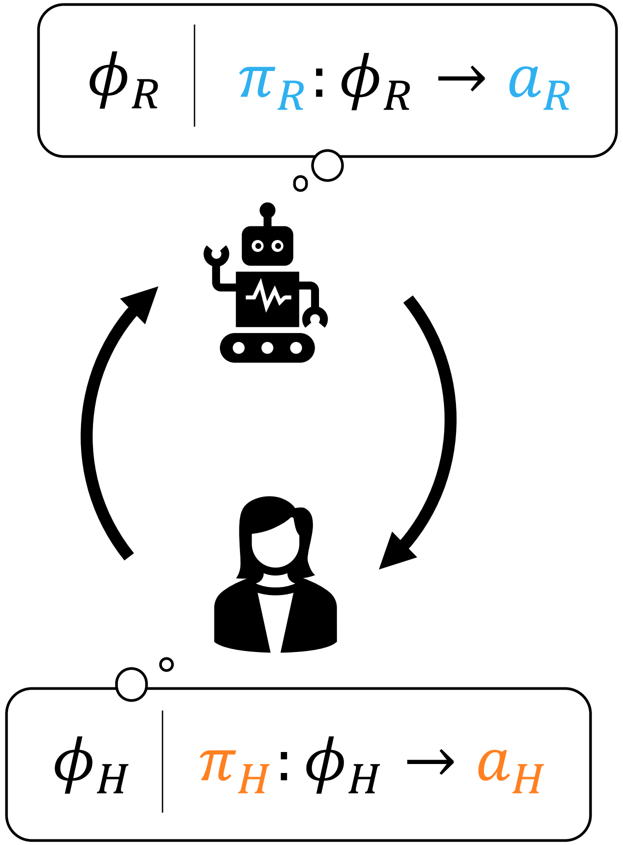
Collision rate

Safety override rate

Reward Improvement % (w.r.t worst-case safety)

1 How can we **formalize** interactive robot safety?

2 How can robots adapt their safety strategies under uncertainty?



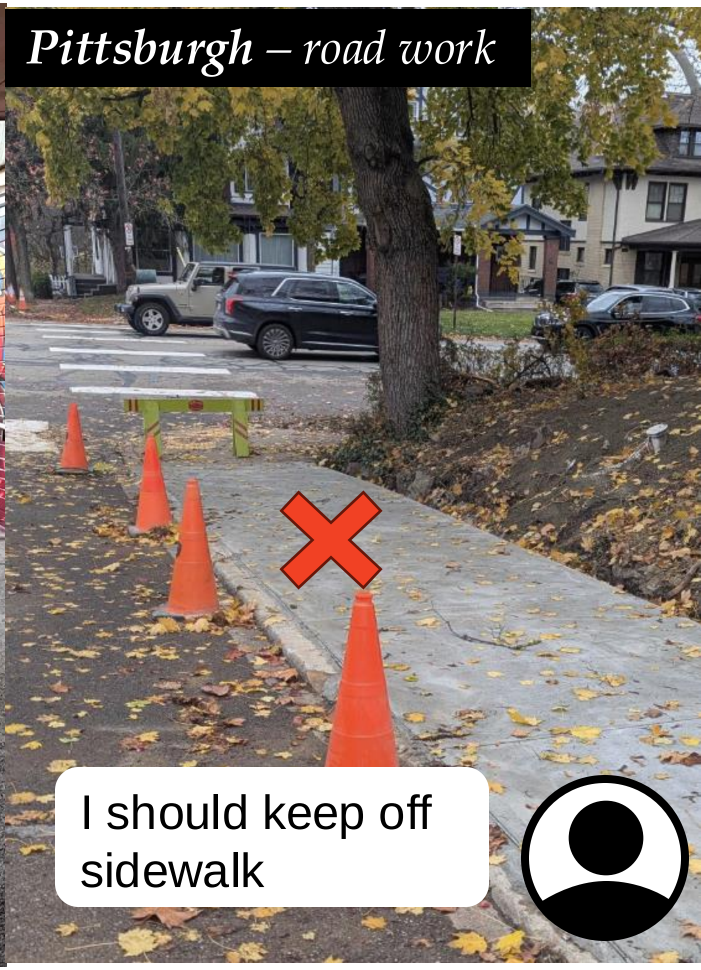
3 How can robots learn safety **representations** from **humans**?

So far, the safety representations we have seen are….

$$\mathcal{F} = \{x : \| x_R - x_H \|_2 \leq \epsilon\}$$

$$\mathcal{F} = \{x : \| x - \mathrm{SLAM}(x) \|_2 \leq \epsilon\}$$

But in the open world, there are many more constraints….

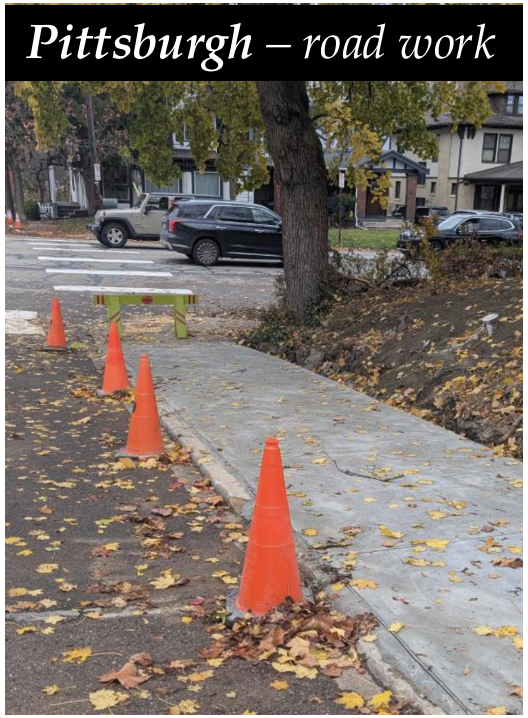


Real images taken by my students!

But in the open world, there are many more constraints….

Brazil – *caution tape*

Germany – *graffiti artist's setup*



Pittsburgh – *road work*

Spills

Accidents

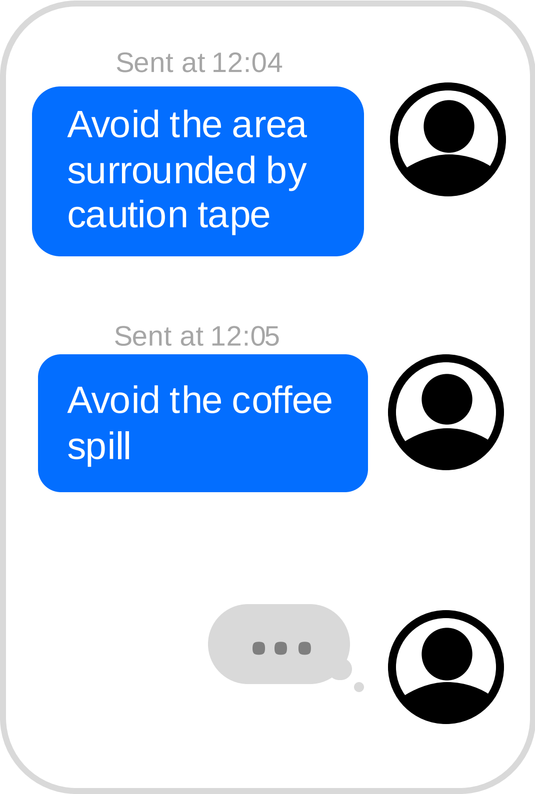
Fragile objects

Sensitive personal areas

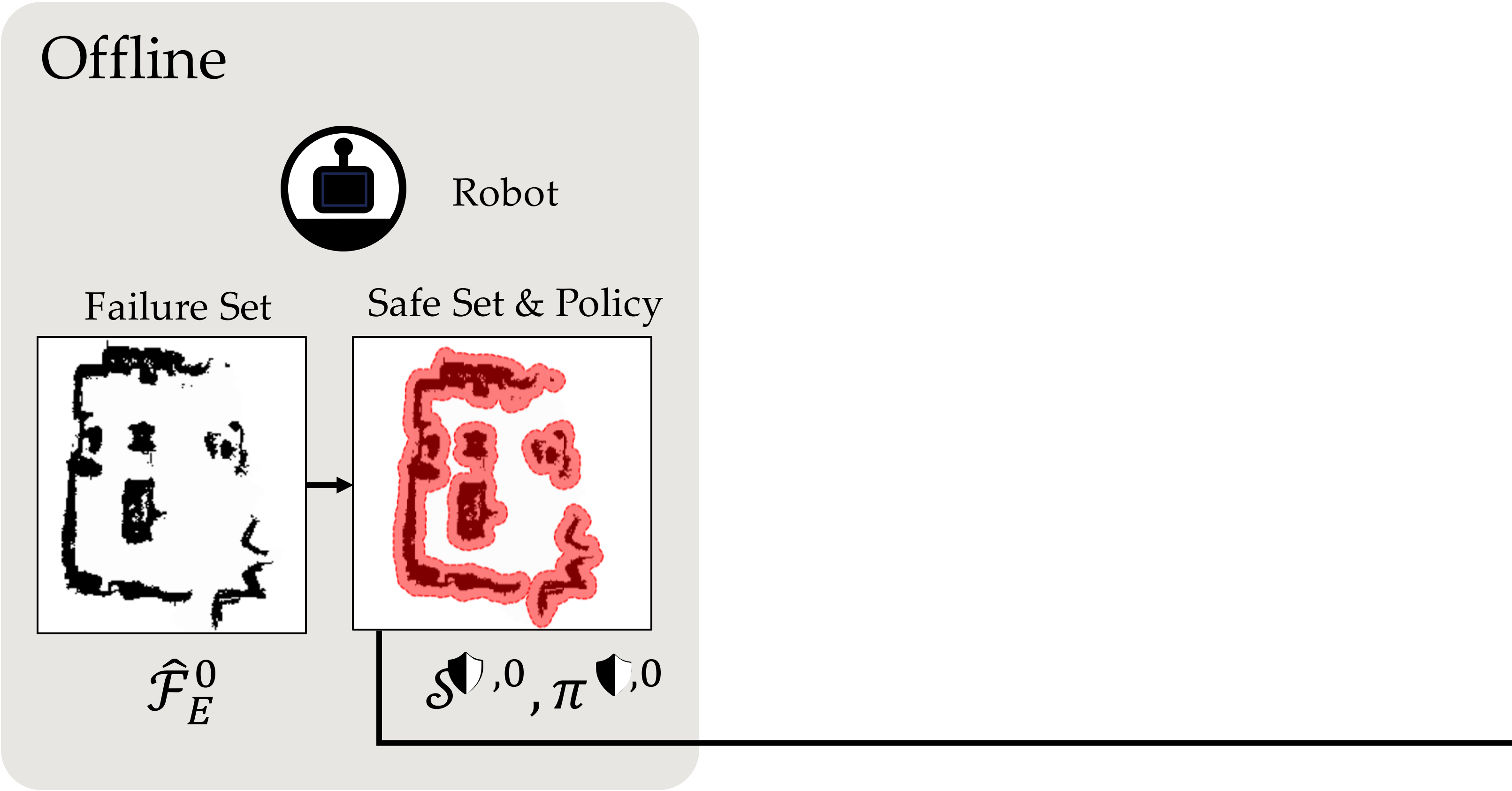

MPPI

How can we encode – and continually update –
these **semantically-meaningful safety constraints**?

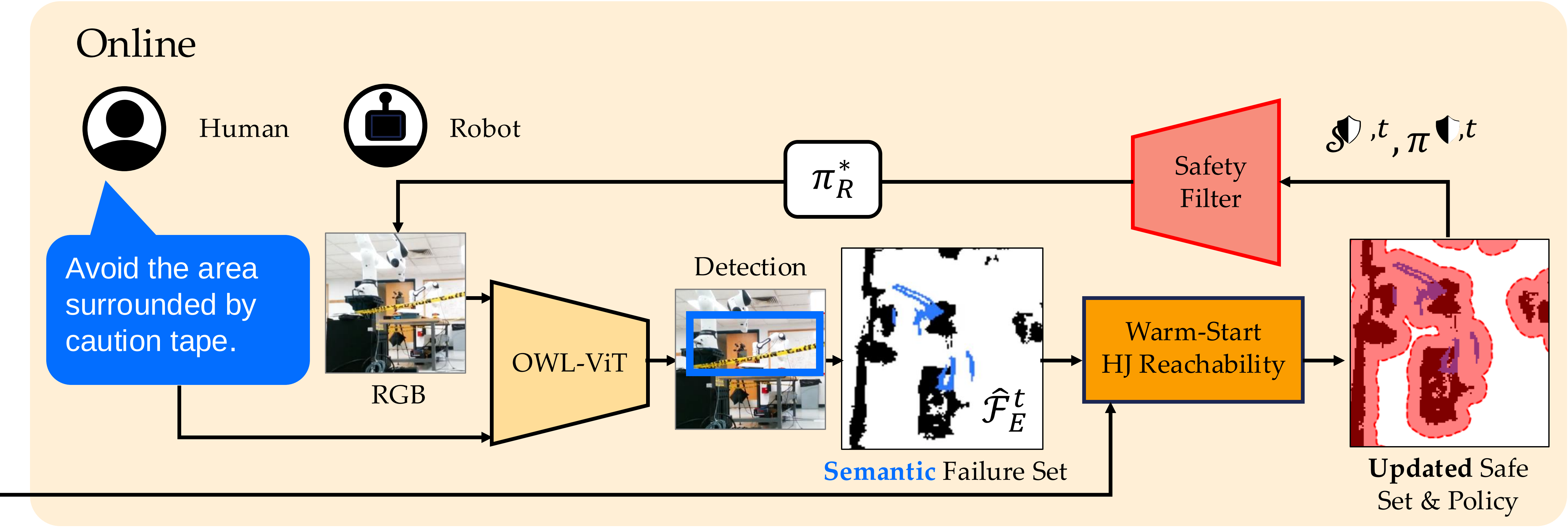
Language Feedback



Vision-language models enable a flexible way to communicate safety constraints to the robot



L. Santos*, Z. Li*, L. Peters, S. Bansal†, A. Bajcsy† . "Updating Robot Safety Representations Online from Natural Language Feedback" arXiv 2024. *(ICRA submission)*



L. Santos*, Z. Li*, L. Peters, S. Bansal†, A. Bajcsy† . “Updating Robot Safety Representations Online from Natural Language Feedback” arXiv 2024. *(ICRA submission)*

From the human's POV…

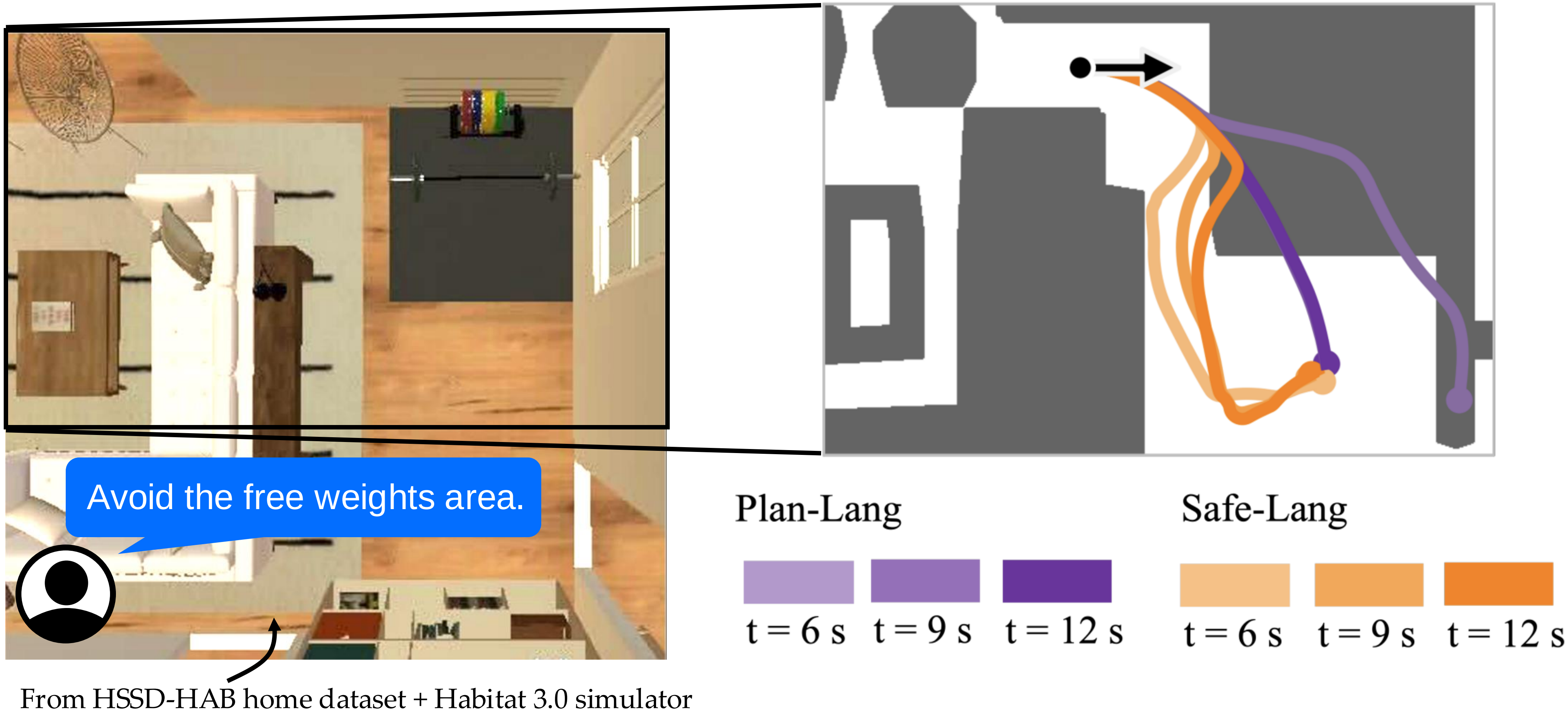
L. Santos*, Z. Li*, L. Peters, S. Bansal†, A. Bajcsy† . "Updating Robot Safety Representations Online from Natural Language Feedback" arXiv 2024. *(ICRA submission)*

From the robot's POV…

L. Santos*, Z. Li*, L. Peters, S. Bansal†, A. Bajcsy† . "Updating Robot Safety Representations Online from Natural Language Feedback" arXiv 2024. *(ICRA submission)*

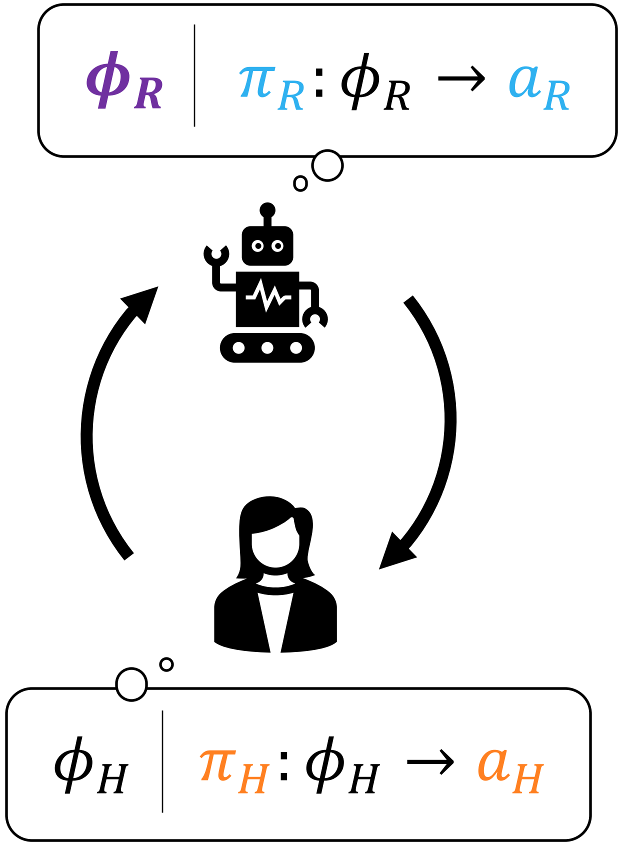

Unitree
Vector Pro
AI Workstation
“avoid the dog toys and the laundry”

On the Robustness to Language Feedback Timing



1 How can we **formalize** interactive robot safety?

2 How can robots adapt their **safety strategies under uncertainty**?

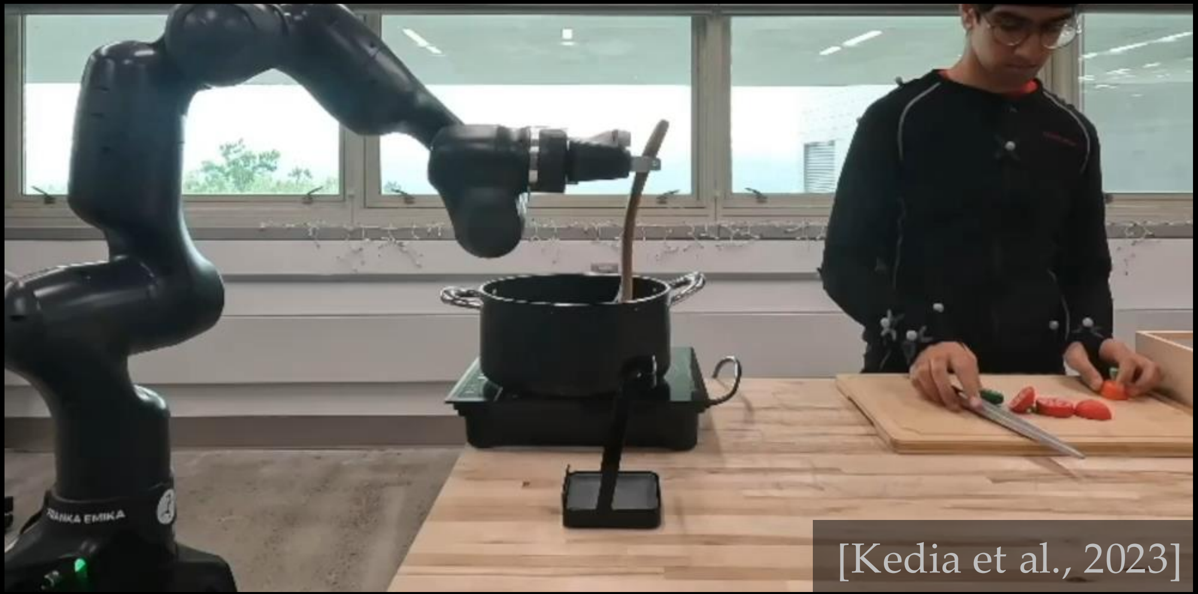


3 How can robots learn safety **representations** from **humans**?


[Waymo, 2023]


[Ren, AZ et al., 2023]

More work to be done so autonomous robots can interact safely at scale

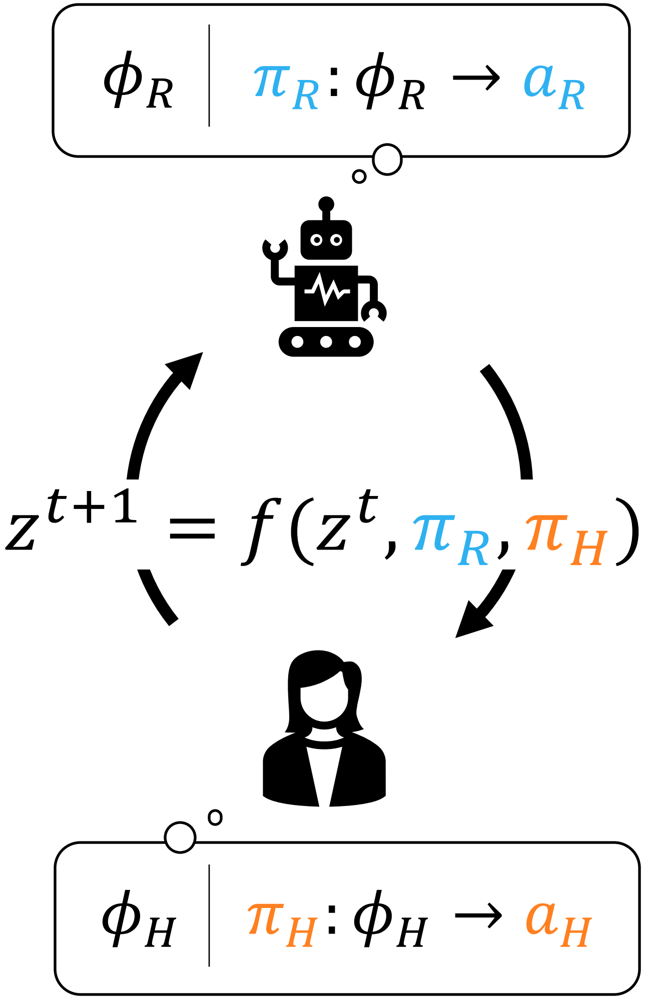

[Kedia et al., 2023]


[DeepMind, 2023]

Safety & Uncertainty in Human-Robot Interaction

1. Formalize safety during interaction via **zero-sum dynamic games**

2. Adapt robot safety strategies based on **confidence in predictive human models**



3. Robots can learn more nuanced safety **representations** from **natural language feedback**

Carnegie Mellon University

abajcsy@cmu.edu

intent ROBOTICS LAB