16-886

Embodied AI Safety

Instructor: Andrea Bajcsy

Welcome!

Carnegie Mellon University

intent ROBOTICS LAB

Professor

Andrea Bajcsy

(BYE-chee)

What to call me:

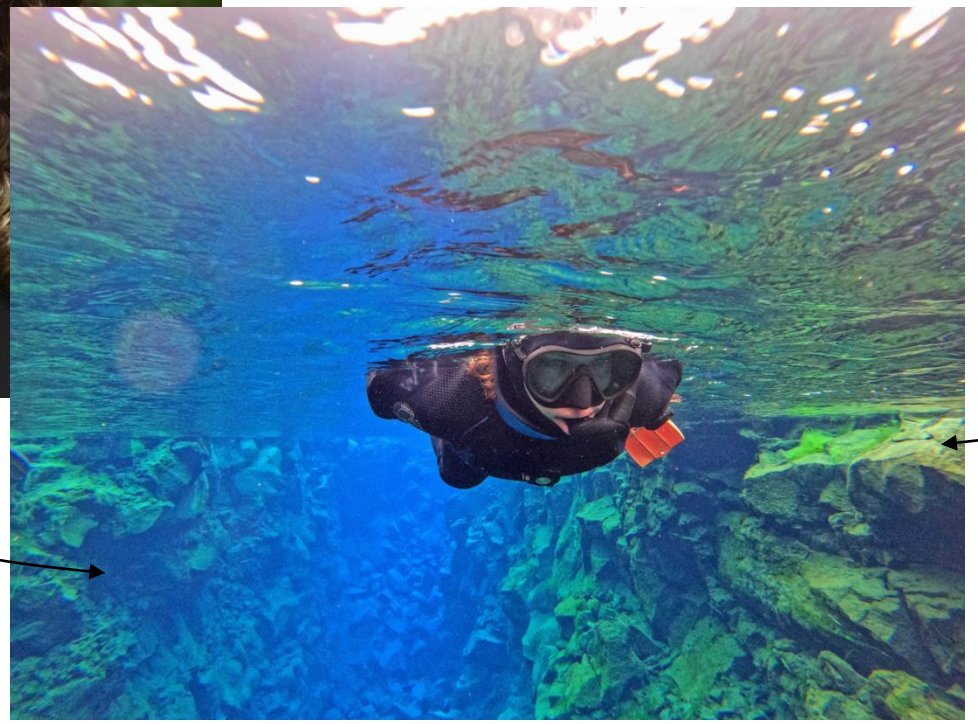
- Andrea *(if you are a grad student)*
- Prof. Bajcsy or Prof. B *(if you are undergrad)*

Office Location: NSH 4629
Office Hours: Wednesdays, 12:20-1:20pm *(after class)*
Email: abajcsy@cmu.edu

Professor

Andrea Bajcsy

Fun fact: I swam between the Eurasian and North American tectonic plates!



Silfra Fissure, Iceland

Teaching Assistant

Junwon Seo, PhD Student

Research Interests:

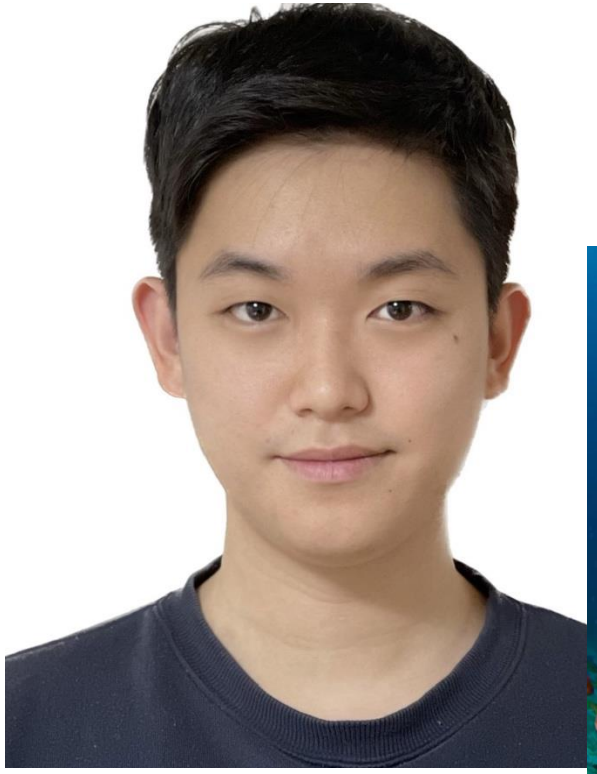
- theoretical foundations and algorithms for quantifying *uncertainty* in learning-based robots
- ensuring *safety* and *robustness* in complex, open-world environments

Office Location: NSH TBA
Office Hours: TBA

- **Please take survey on Canvas so we can select OHs that suit folks best**

Email: junwonse@andrew.cmu.edu

Teaching Assistant



Junwon Seo, PhD Student

Fun fact: I am a PADI Divemaster!

What is next?

Course Contents

Course Logistics

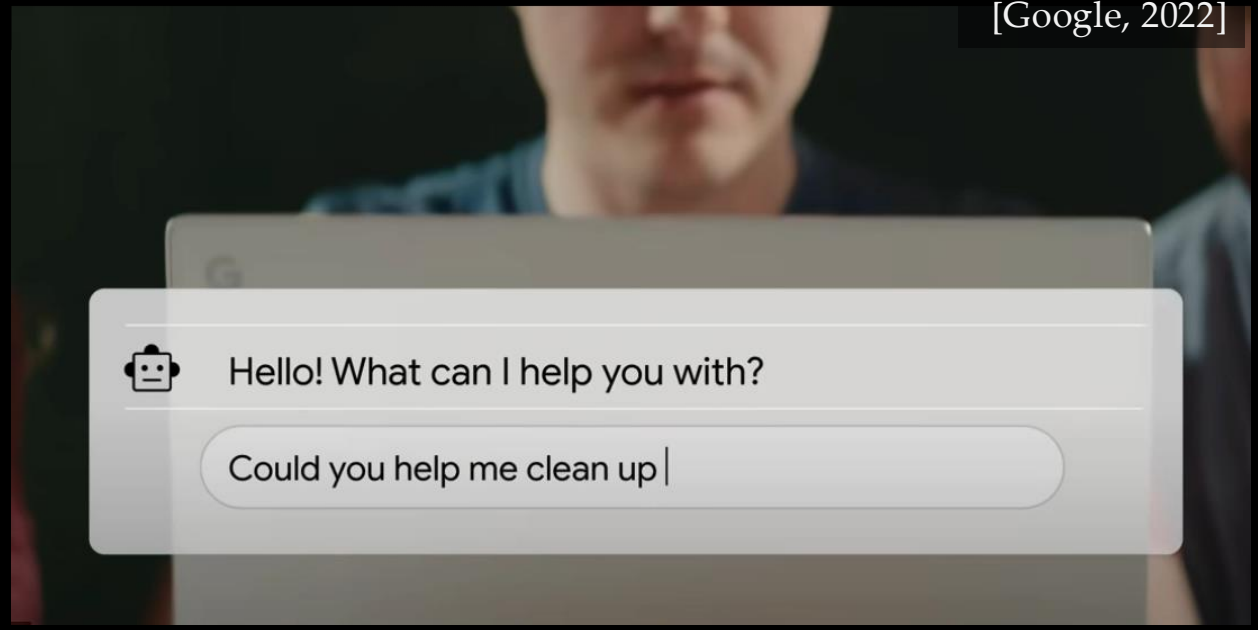
Intro Survey

(Intro to Sequential Decision-Making)

This class: Embodied AI Safety

"agents which interact with the environment to accomplish tasks"


[Waymo, 2023]

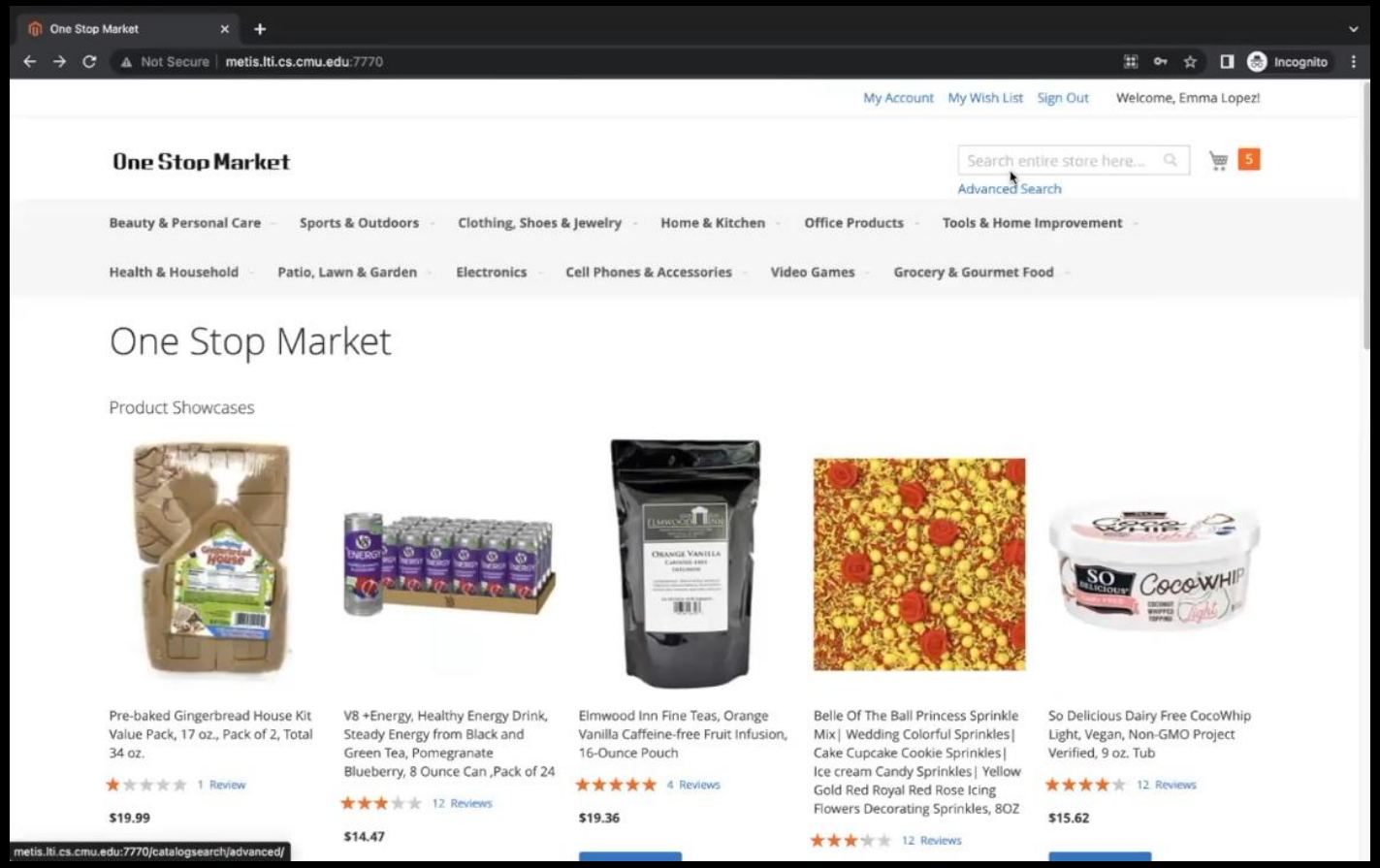


[Google, 2022]


[Skydio, 2023]


[Toyota Research Institute, 2023]

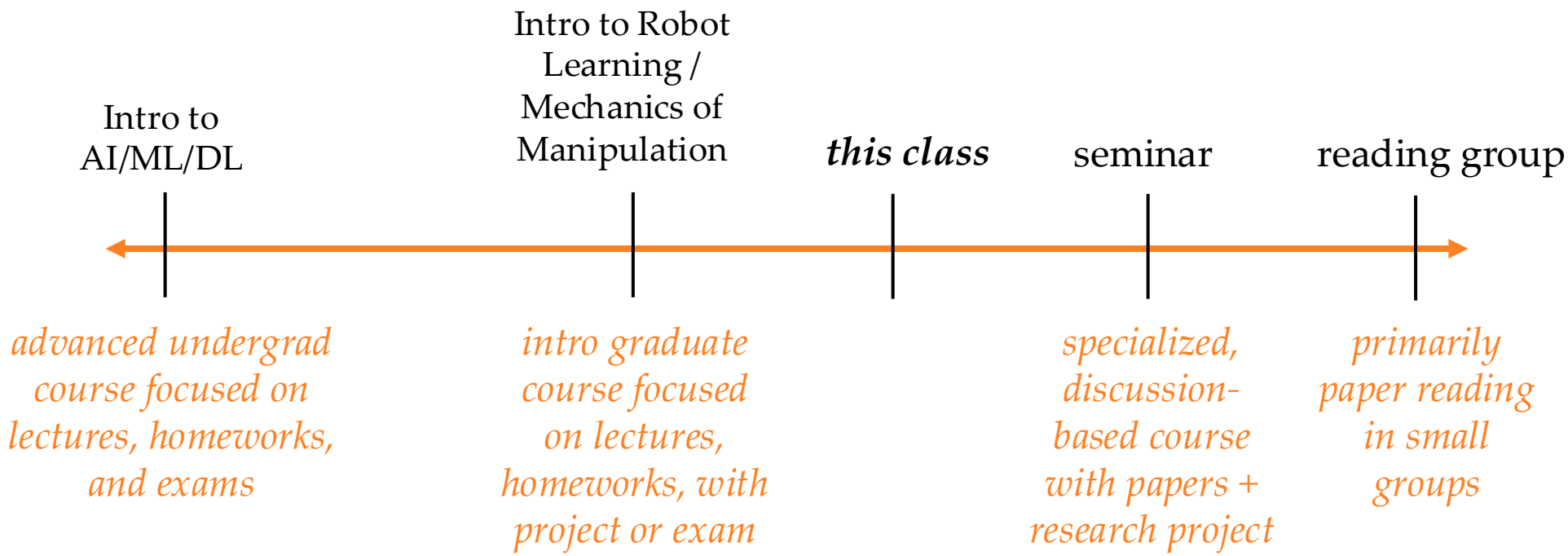
Most examples in this class will be of these EAI systems – **robots**!



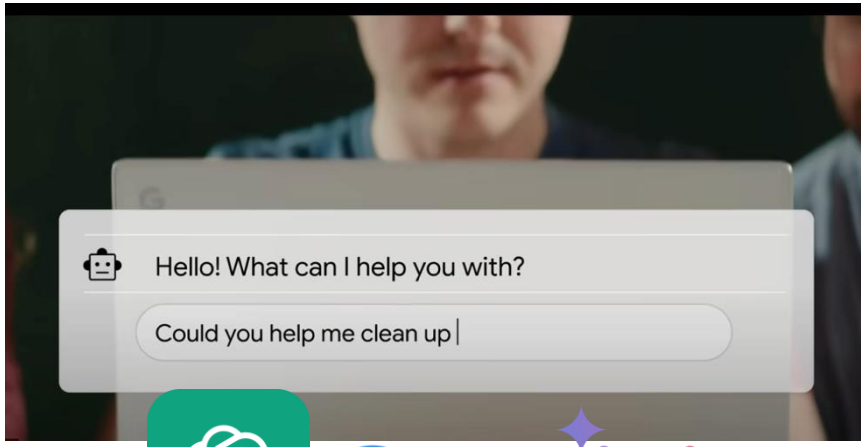
[Zhou et al. "WebArena", arXiv 2023]

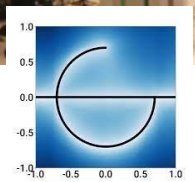
But the core ideas are also relevant to current & future EAI systems

What exactly is this class?



World Models, Video Prediction Models, …





Diffusion Policy

This class: Embodied AI Safety

Some properties of AI you will see in class:

- learned patterns instead of only hand-designed ones
- high-dimensional inputs (e.g., RGB images, natural language, HTML, thermal images)
- “end-to-end” deep models

Increased capabilities & deployment have escalated concerns about safety

MIT Technology Review

SUBSCRIBE

ARTIFICIAL INTELLIGENCE

Why humanoid robots need their own safety rules

Humanoid robots pose unique safety risks. That's driving a push for new standards before they start sharing our workplaces and homes.

By Victoria Turk

June 11, 2025

STEPHANIE ARNETT/MIT TECHNOLOGY REVIEW | ADOBE STOCK

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 could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and

MENU

This class: Embodied AI Safety

what is safety?

Group exercise!

Imagine you bought an autonomous humanoid robot for your home.

Group 1	**Imagine the robot's sole purpose is to walk through the inside of your house.** What are your safety concerns? How would you solve them and what are the bottlenecks?
Group 2	**Imagine the robot's sole purpose is to clean your kitchen.** What are your safety concerns? How would you solve them and what are the bottlenecks?
Group 3	**Imagine the robot's sole purpose is to prepare meals for you**. What are your safety concerns? What could cause these safety concerns?
Group 4	**Imagine that the robot gets into a vehicle and drives it to the store.** What are your safety concerns? What could cause these concerns?
Group 5	How would you "prove" that this robot is ready to deploy in a home? In other words, what "assurance" would you want from this system?

Neo, 1X Technologies

In the “open world” and as robots become more “generalist”, safety is a nuanced concept

First, let's think through "simple" safety specification….

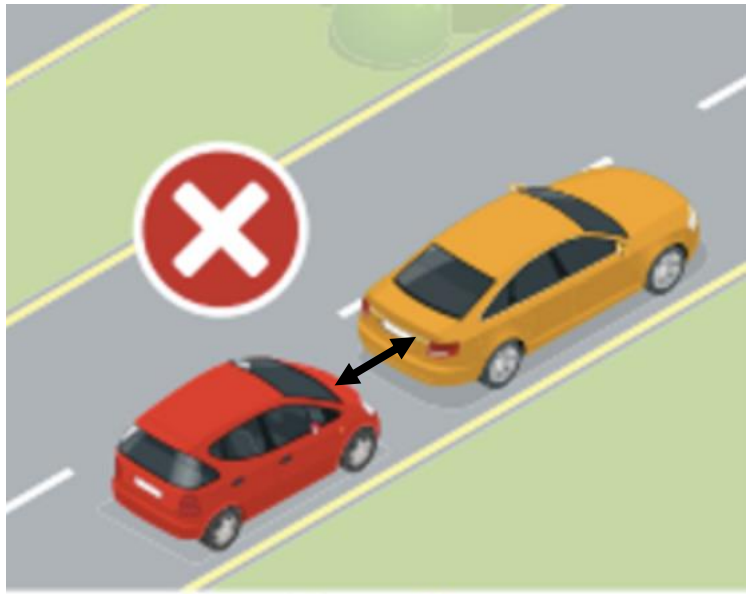
designer

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

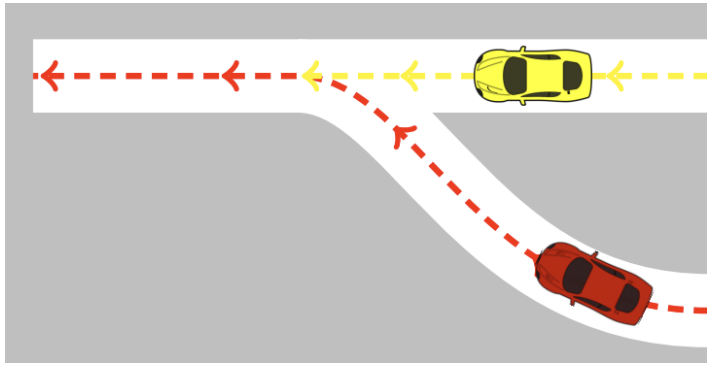
i.e., “don’t collide”

Too close

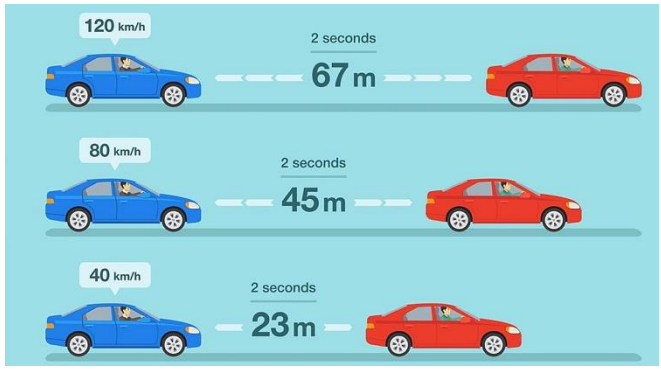
```
                    ⎧ brake    if d(you, front_car) < car_len
car_action =        ⎨
                    ⎩ speed    else
```



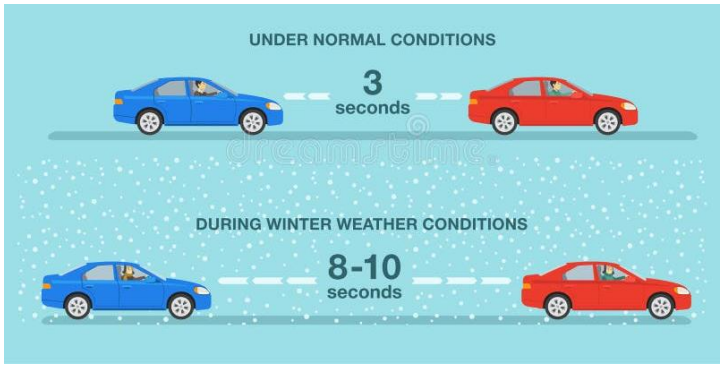
Env. topology



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
paramet
addition
that eve

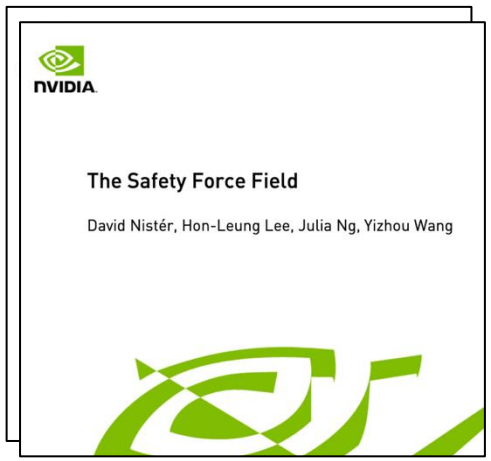
Definition 1 (Safe longitudinal distance — same direction) *A* longitudinal distance *between a car c_r that drives behind another car c_f, where both cars are driving at the same direction, is* safe *w.r.t. a response time ρ if for any braking of at most $a_{\max,\text{brake}}$, performed by c_f, if 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.*

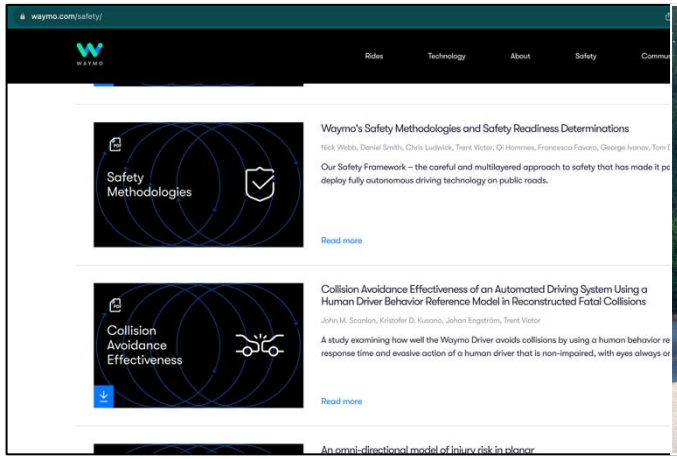
Lemma 2 below calculates the safe distance as a function of the velocities of c_r, c_f and the parameters in the definition.

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 $[x]_+ := \max\{x, 0\}$.





Even if safety specification is “simple”, decision-making is hard

Unsafe early braking (Tesla, 2023)

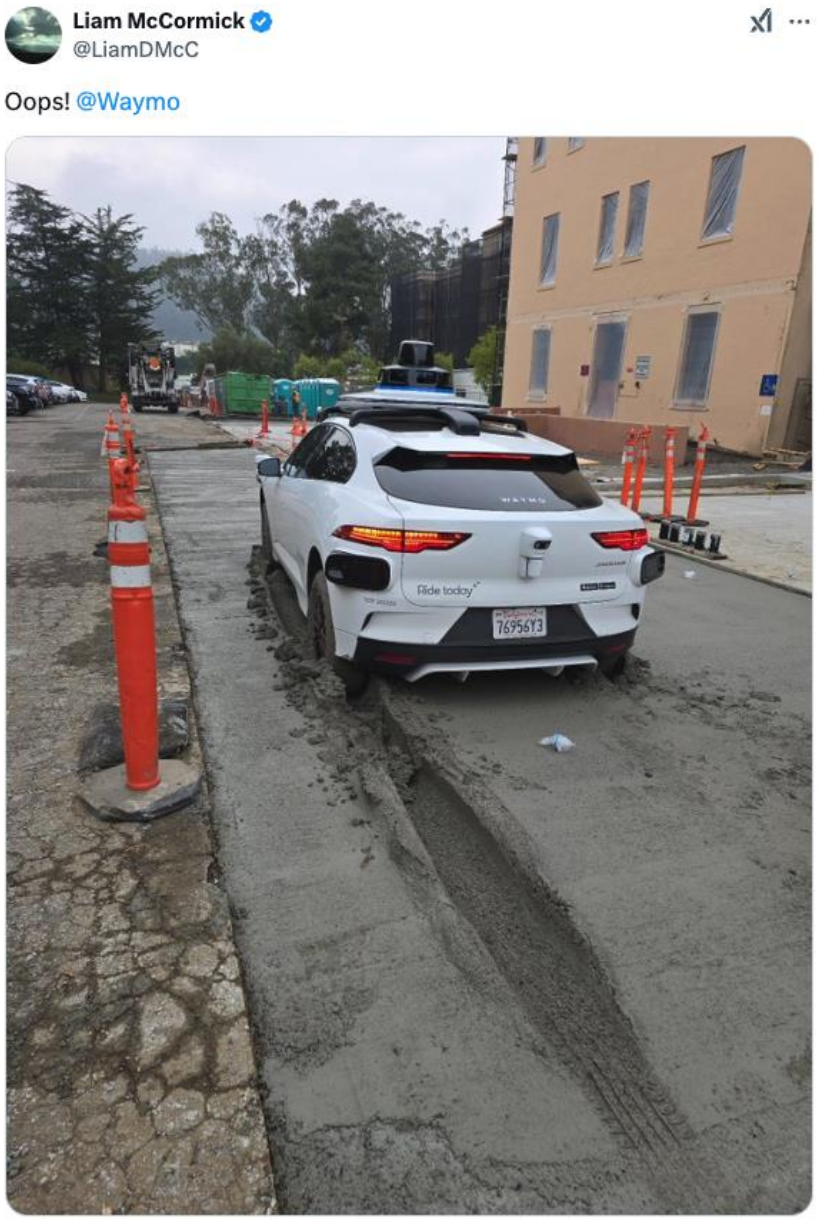
Source: https://abc7news.com/

In the open-world, what “safety” means becomes much more nuanced….

Safety is “in the eye of the stakeholder” *(also called alignment)*

[Autoevolution, “Tesla with FSD Beta Avoids Plastic Bag, Some People Mistake It for a Good Thing“]

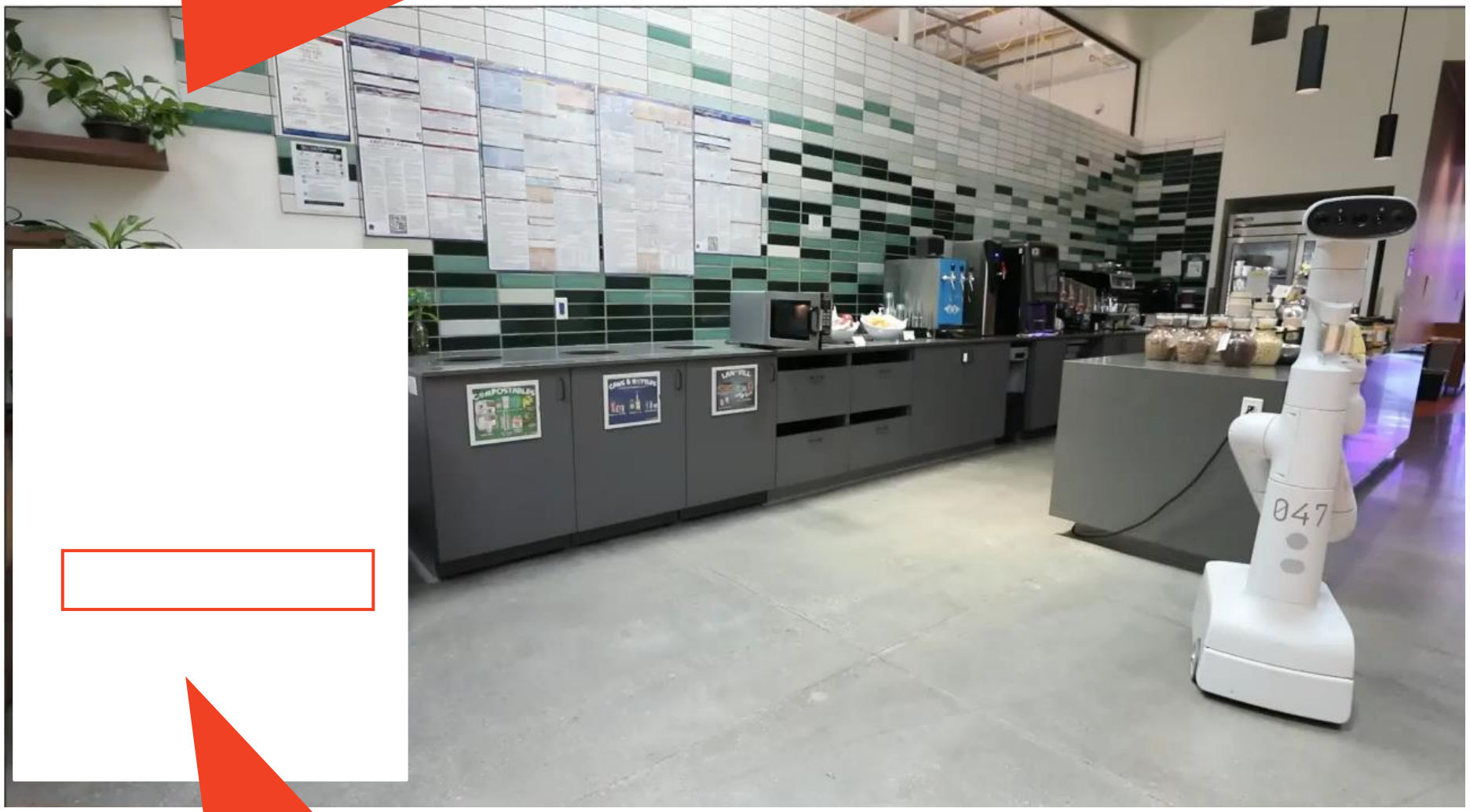
Our representations of safety should be “richer” than just geometry



[Guan, et al. "Task Success" is Not Enough. COLM 2024]

Uncertainty and semantics play a key role in open-world safety

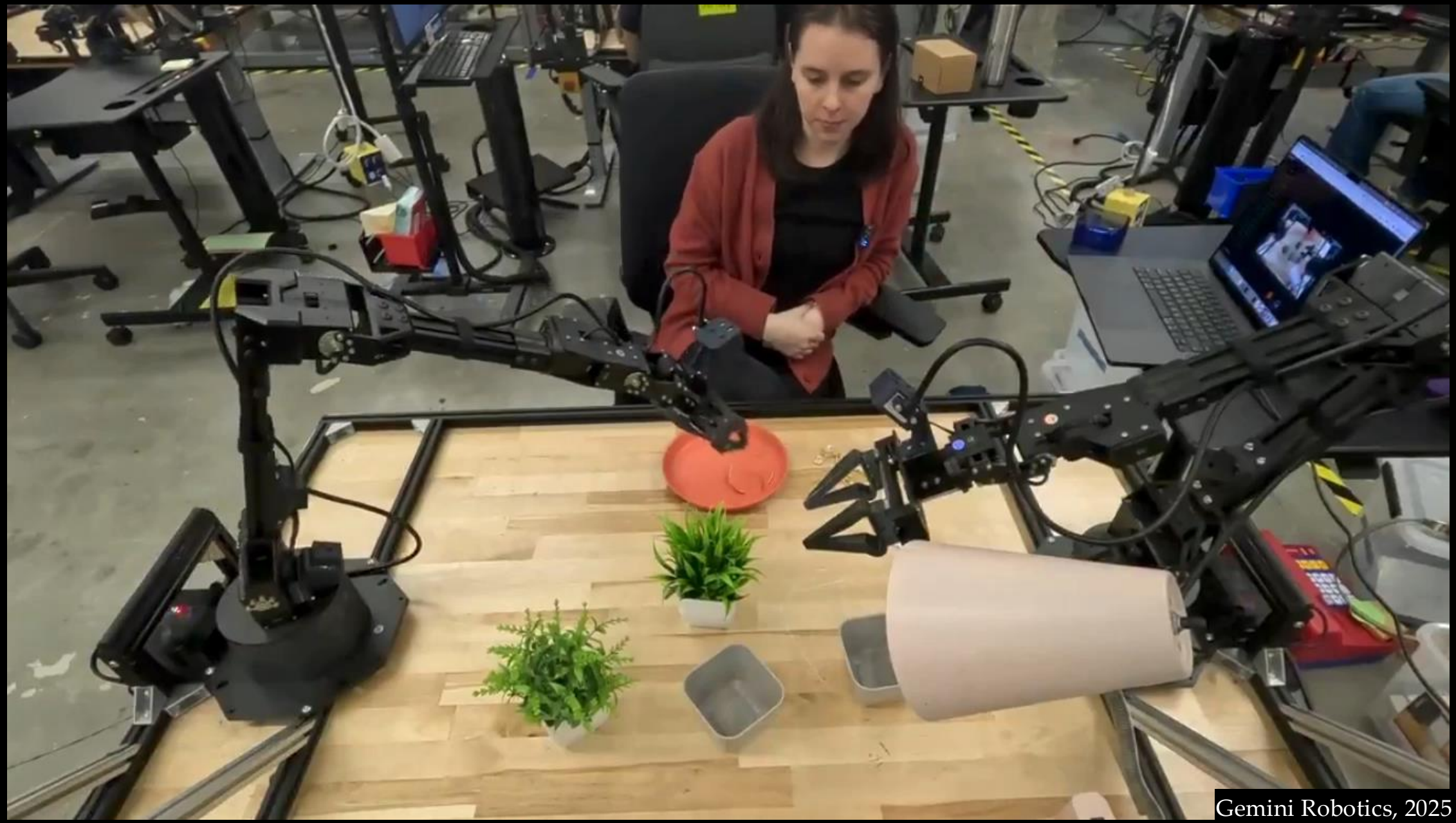
Knowing that its unsafe to put metal or plastic in microwave



Asking for help when uncertain

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

Uncertainty and semantics play a key role in open-world safety

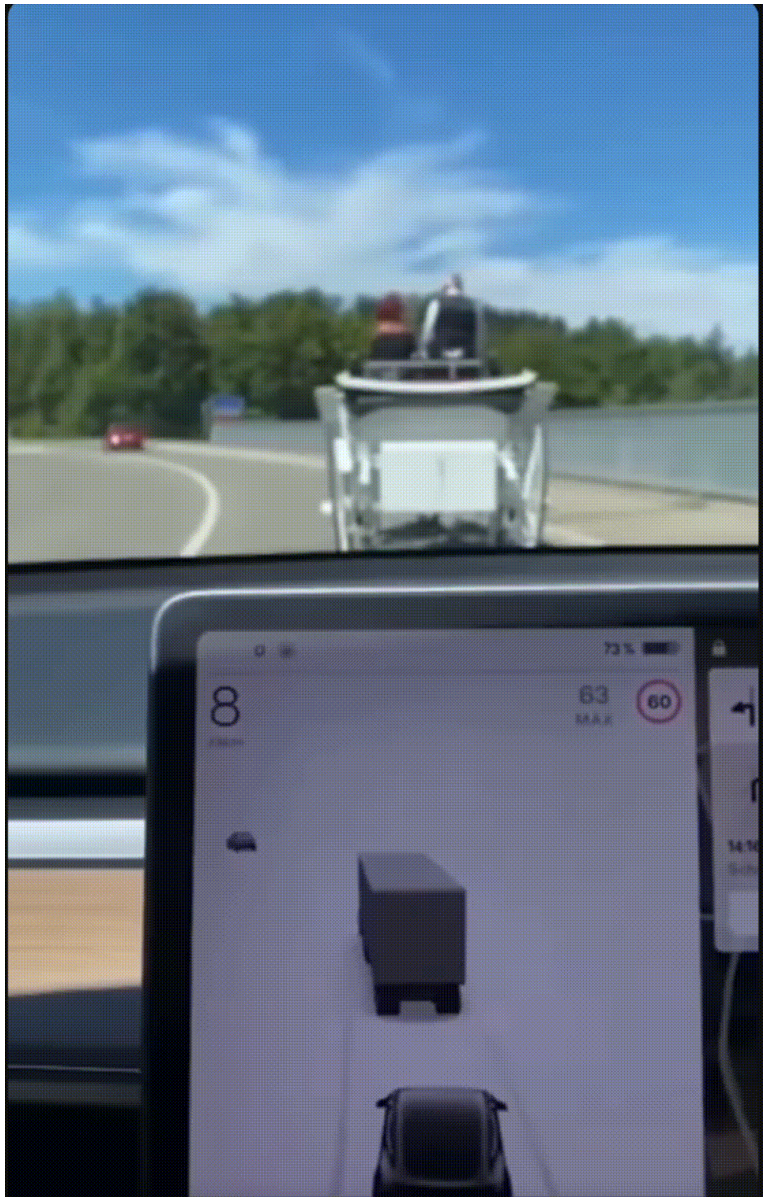


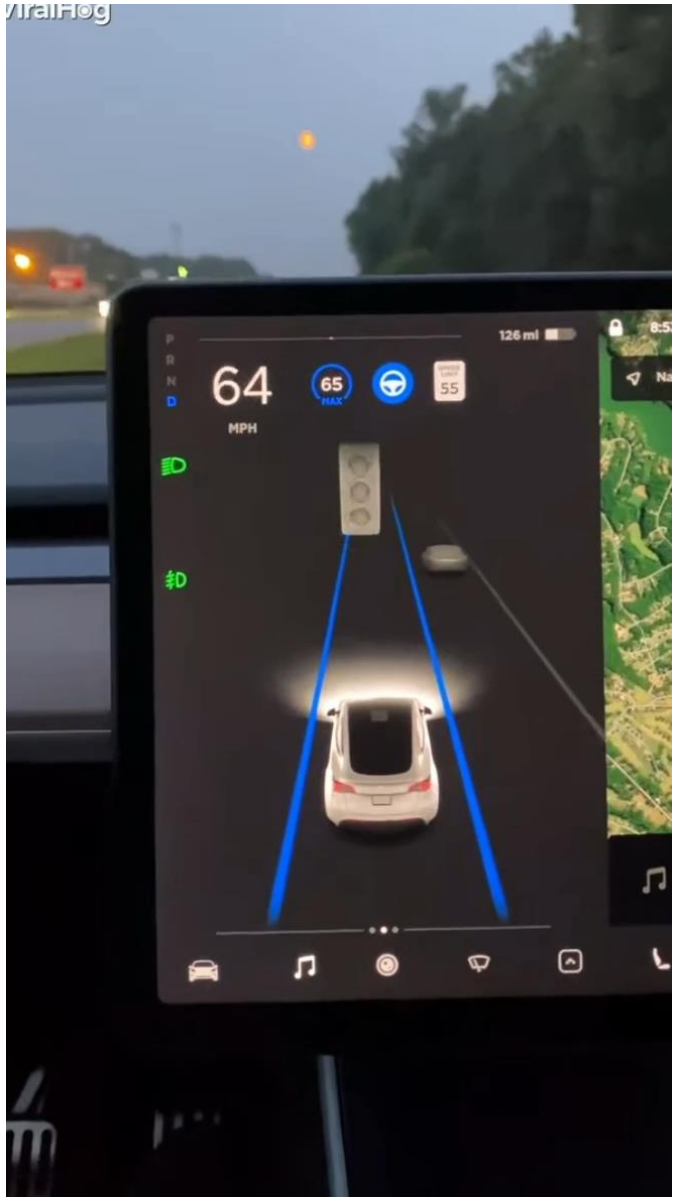
Gemini Robotics, 2025

Safety should reason about out-of-distribution or anomalous scenarios

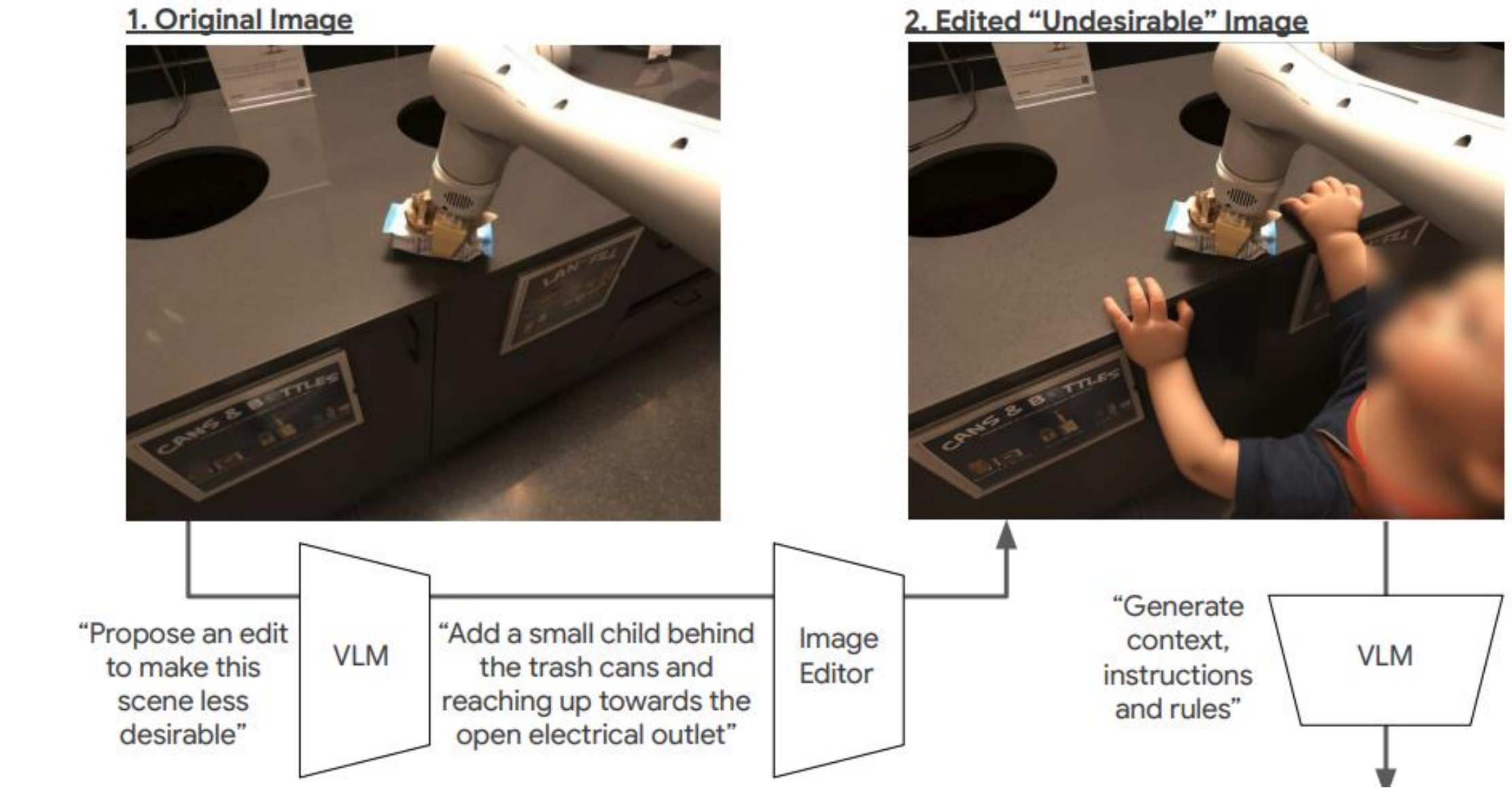
[Hanock, Ren, Majumdar. "BYOVLA". arXiv 2024]

Safety should reason about out-of-distribution or anomalous scenarios





Lack of “long tail” datasets at the boundary between safe and unsafe



[Sermanet, Pierre, et al. "Generating robot constitutions & benchmarks for semantic safety." CoRL (2025).]

This class: Embodied AI Safety

What is special about this?

We will formalize & study the full spectrum in the class!

(+) Opportunities of AI Safety

infer hard-to-model low-D patterns from high-D obs

critique outcomes and steer towards good ones

enable novice stakeholders to specify safety that matters to them (e.g., language)

generalize safety representations

generate (synthetic) data for stress-testing

(promise of) deployment into more unstructured or novel environments

(promise of) generalization

(?) Challenges of AI Safety

"misaligned" generations

how to safeguard *any* AI model?

what is OOD or anomalous?

single erroneous vision / language interpretation can lead to catastrophic action

high inference latency

how to couple the detection of anomalies with mitigation actions?

Control / Decision-Theory

how to couple the detection of anomalies with mitigation actions?

how to safeguard *any* FM / ML model?

critique outcomes and steer towards good ones

single erroneous vision / language interpretation can lead to catastrophic action

misaligned behavior generation

(promise of) deployment into more unstructured or novel environments

generalize safety representations

generate (synthetic) data for stress-testing

high inference latency

Machine Learning / Statistics

infer hard-to-model low-D patterns from high-D obs

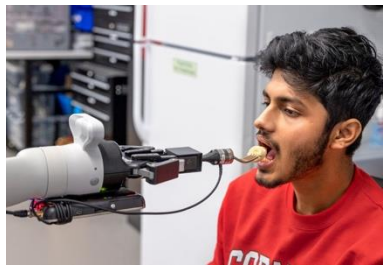
enable novice stakeholders to specify safety that matters to them (e.g., language)

(promise of) generalization

what is OOD or anomalous?

Why this course?

Take any embodied AI application and …



1) **Formalize** what safety means mathematically

2) **Solve algorithmically** safety problems unique to embodied AI

3) **Identify** emerging trends and challenges in embodied AI safety

What you will learn in this course

Control-Theoretic Safety Foundations

- Safe control & safety filters
- Robustness via dynamic games
- Computational frameworks (RL & self-supervised learning)

Frontiers I

- Updating safety online
- "Semantic safety"
- Latent-space safety
- Runtime monitoring & recovery via VLMs

Machine Learning & Statistical Safety Foundations

- Uncertainty quantification
- Conformal prediction
- Risk and anomalies

Frontiers II

- Controlling in-distribution
- Uncertainty in generative models
- Statistical testing of learned policies / models

Guest Lectures

Foundations: Reinforcement Learning Approx. to Safe Control

Ken Nakamura
PhD Student @ CMU

Frontiers: Safety in Field Foundation Models

Jay Patrikar
Senior Research Scientist @ FieldAI

+ a few more folks!

To be announced soon!

General Resources

- No textbook!

- If I were to recommend textbooks for this class…

Artificial Intelligence: A Modern Approach by Russell and Norvig
Reinforcement Learning: An Introduction by Sutton and Barto
Reinforcement Learning and Optimal Control by Dimitri Bertsekas
Dynamic Noncooperative Game Theory by Başar and Olsder
Numerical Optimization by Jorge Nocedal and Stephen J. Wright
Algorithms for Validation by Mykel Kochenderfer, Sydney Katz, Anthony Corso, and Robert Moss

Course Logistics

Format: lecture or related paper reading discussions

Typical 80-min class:

~5 min attendance quiz at start

70 min lecture, invited talk, or paper discussion

Use *course website* for up-to-date schedule & paper links

https://abajcsy.github.io/embodied-ai-safety/

Embodied Artificial Intelligence Safety

Spring 2026. 16-886. Monday / Wednesday 11:00-12:20. GHC 4215.

Announcements

Hello!

Jan 1 · 0 min read

See you soon! 🤓

Schedule (Tentative)

Control-Theoretic Safety Foundations

Jan. 12:	**Course Overview**	Syllabus
Jan. 14:	**Sequential Decision-Making**	
Jan. 19:	NO CLASS **MLK Day**	
Jan. 21:	**Why is Safe Control Hard and What are Safety Filters?**	Data-Driven Safety Filters, Model Predictive Sheilding
Jan. 26:	**Safety Filter Synthesis via Optimal Control**	HJ Reachability Overview, HJ Viscosity Solution
Jan. 28:	**Robust Safety I**	Differential Games, HJ Reach-Avoid Games I, HJ Reach-Avoid Games II
Feb. 2:	**Robust Safety II**	
Feb. 4:	GUEST LECTURE **Computation: Reinforcement Learning (Kensuke Nakamura)**	HW #1 DUE Discounted Reachability, ISAACS
Feb. 9:	**Computation: Supervised Learning & PINNs**	DeepReach

Frontiers I

Use *Canvas* for downloading / uploading assignments

Spring 2025

- Home
- Announcements
- Syllabus
- Assignments
- Quizzes
- Grades
- Discussions
- Files
- People
- Zoom
- NameCoach
- Syllabus Registry
- Pages
- Outcomes
- Collaborations
- Rubrics
- Modules
- Settings

Recent Announcements

Embodied Artificial Intelligence Safety

Assign To | Edit | ⋮

Welcome to **16-886: Embodied AI Safety!**

Safety is a nuanced concept. For embodied systems, like robots, we commonly equate safety with collision-avoidance. But out in the "open world" it can be much more: for example, a safe mobile manipulator should understand when it is not confident about a requested task and understand that areas roped off by caution tape should never be breached. However, designing systems with such a nuanced understanding is an outstanding challenge, especially in the era of large robot behavior models.

In this graduate seminar class, we study the question of if (and how) the rise of modern artificial intelligence (AI) models (e.g., deep neural trajectory predictors, large vision-language models, and latent world models) can be harnessed to unlock new avenues for generalizing safety to the open world. From a foundations perspective, we study safety methods from two complementary communities: *control theory* (which enables the computation of safe decisions) and *machine learning* (which enables uncertainty quantification and anomaly detection). Throughout the class, there will also be several guest lectures from experts in the field. Students will practice essential research skills including reviewing papers, writing project proposals, and technical communication.

Class Website: https://abajcsy.github.io/embodied-ai-safety/

Grading

*See **class syllabus on course website** for detailed info*

Percentage	Activity
10%	Attendance
25%	HW (3x)
5%	Paper Summaries
5%	Project Proposal
20%	Midterm Project Report (Report + Presentation)
35%	Final Project (Report + Presentation)

Attendance & Participation (10%)

Expected to attend class in person—this is how we will all get the most out of the class! Please show up on time, especially for reading days

The way we grade this:

- First 5 minutes of class: we will give a **short, easy "quiz"** related to the last lecture's content. This is graded as 1/0.
  - e.g., *"Describe what is a sequential decision-making problem."*
- **Permitted 2 unexcused absences**, no questions asked, before being docked.

I understand that occasionally you may have challenges attending (e.g., illness, religious observance,..); **please let me know.**

Homework (25%)

HW #1: Computing & Using Safety Filters

Released:	~Jan 21
Due:	Feb 4

HW #2: Latent-Space Safety Filters

Released:	~Feb 11
Due:	Feb 28

HW #3: Conformal Prediction

Released:	~Mar 18
Due:	April 1

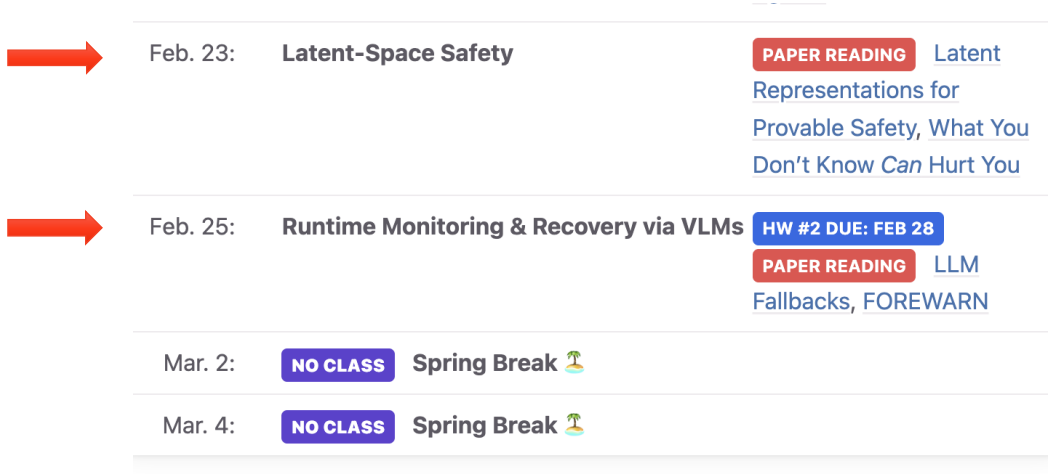
These are coding-based homeworks in **Python** and **PyTorch**. They are *not* meant to be tedious; they are meant to **empower** you! ☺

If you are not confident (or are rusty) with Python and Pytorch, please come see us for educational resources!

Paper Summaries (5%)

Paper discussion days:

6 paper reading days
2 papers per reading day



Feb. 23:	**Latent-Space Safety**	PAPER READING Latent Representations for Provable Safety, What You Don't Know *Can* Hurt You
Feb. 25:	**Runtime Monitoring & Recovery via VLMs**	HW #2 DUE: FEB 28 PAPER READING LLM Fallbacks, FOREWARN
Mar. 2:	NO CLASS **Spring Break** 🏝	
Mar. 4:	NO CLASS **Spring Break** 🏝	

Before class:

Answer three questions about the paper:

1) What does safety "mean" in this work?
2) Does the notion of safety apply to a single component / model of the autonomy pipeline, or to the overall "system"? Justify.
3) What do you like about this work?

Must submit on Canvas before class.

In class:

Split you into small groups, discuss set of questions, I assign a representative from each group to present on the group's takeaways, and the whole class can engage on the answer

Class Project

Two options:

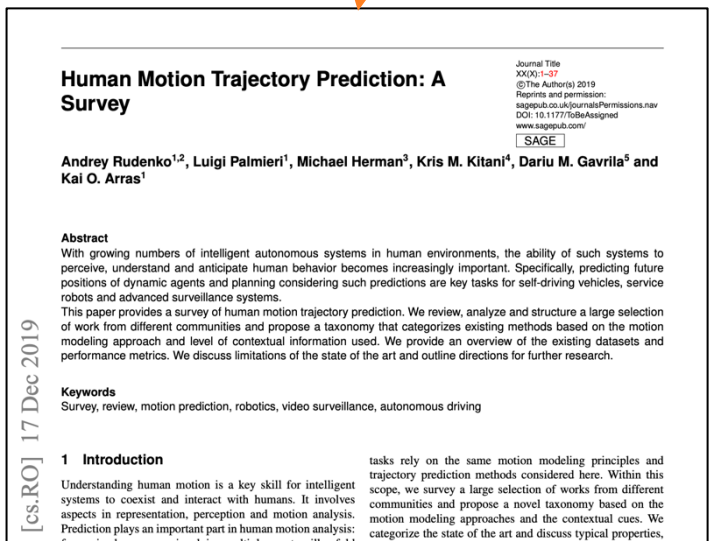
Research project:

Identify a research direction broadly relevant to this class
Propose and take first steps towards an original idea

Literature survey:

Select a topic area and rigorous way in which you will find papers
Characterize this topic area in an insightful way (e.g., open questions, common assumptions, tractable vs. theoretical gaps)

Example of good literature survey



Human Motion Trajectory Prediction: A Survey

Journal Title
XX(X):1–37
©The Author(s) 2019
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Andrey Rudenko[1,2], Luigi Palmieri[1], Michael Herman[3], Kris M. Kitani[4], Dariu M. Gavrila[5] and Kai O. Arras[1]

Abstract
With growing numbers of intelligent autonomous systems in human environments, the ability of such systems to perceive, understand and anticipate human behavior becomes increasingly important. Specifically, predicting future positions of dynamic agents and planning considering such predictions are key tasks for self-driving vehicles, service robots and advanced surveillance systems.
This paper provides a survey of human motion trajectory prediction. We review, analyze and structure a large selection of work from different communities and propose a taxonomy that categorizes existing methods based on the motion modeling approach and level of contextual information used. We provide an overview of the existing datasets and performance metrics. We discuss limitations of the state of the art and outline directions for further research.

Keywords
Survey, review, motion prediction, robotics, video surveillance, autonomous driving

1 Introduction

Understanding human motion is a key skill for intelligent systems to coexist and interact with humans. It involves aspects in representation, perception and motion analysis. Prediction plays an important part in human motion analysis:

tasks rely on the same motion modeling principles and trajectory prediction methods considered here. Within this scope, we survey a large selection of works from different communities and propose a novel taxonomy based on the motion modeling approaches and the contextual cues. We categorize the state of the art and discuss typical properties,

[cs.RO] 17 Dec 2019

You must work in a group of min 2 to max 4 people. Explicit permission from Andrea must be granted if you want to work on an independent project

Class Project

When picking a project, make sure to answer the question:

How does the project connect to the broader topics & context of the class?

Examples of projects within scope (non-exhaustive list!)

- Applying one of the techniques from class to your problem domain
  - *(e.g., apply a new uncertainty quantification (UQ) methods to a vision-language-action model, apply one of the control method to an LLM agent so it stays away from unsafe outcomes)*
- Rigorously comparing two methods that seek to solve the same problem
  - *(e.g., comparing RL vs. SSL for solving HJ reachability problems, comparing ensembles vs. conformal prediction for UQ to a deep trajectory forecasting model)*
- Systematically studying a phenomenon in a complex model
  - *(e.g., study what makes a visuomotor policy robust; study how good are VLMs at semantic safety detection)*
- Posing (and solving) a new safety problem for your problem domain
- Challenging an assumption underlying one of the methods in the class

Class Project

When picking a project, make sure to answer the question:

How does the project connect to the broader topics & context of the class?

When in doubt: come talk to us about your interests and we can help!

⚠ **Examples of projects <u>not</u> within scope**

- No clear connection to safety (broadly defined)
  - *e.g., a new self-supervised learning method for image classification;*
  - *e.g. training a RL-based humanoid loco-manipulation policy*
- *To make them in scope:*
  - *e.g., a new way of doing uncertainty quantification for image classification;*
  - *e.g., defining the unique safety problems that arise when you can do locomotion and manipulation and building these in as "first class citizens" into your RL pipeline*

Class Project

Project Proposal (5%) – due: Feb. 16

1. **Report:** max 1 page project pitch

Mid-term Project (20%)

1. **In-class Oral Presentation:** short conference-style project pitch (~3-5 mins)
   due: Mar 11
2. **Report:** max 4 page writeup of progress
   due: Mar 18

Final Project (35%)

1. **In-class Oral Presentation:** short conference-style lightning talk pitch (~8-10 mins)
   due: Apr 19
2. **Report:** max 6 page writeup of findings
   due: May 1

Class Project

Project Proposal (5%) – due: Feb. 16

1. **Report:** max 1 page project pitch

Mid-term Project (20%)

1. **In-class Oral Presentation:** short conference-style project pitch (~3-5 mins)
   due: Mar 11
2. **Report:** max 4 page writeup of progress
   due: Mar 18

Final Project (35%)

1. **In-class Oral Presentation:** short conference-style lightning talk pitch (~8-10 mins)
   due: Apr 19
2. **Report:** max 6 page writeup of findings
   due: May 1

Project Proposal

Published | Assign To | Edit

This is a brief (maximum 1 page, excluding references) project pitch. You can think of this as an extended abstract: you want to motivate the topic you have chosen and answer some key brainstorming questions about your directions.

Please use the attached Latex template and answer the questions below (which are also in the template):

- Motivation: describe the context of your project. What is the setting / environment, tasks, and safety problem you are considering, etc? Who do you think will be most interested in your project? What will your project enable in the future?
- Open Challenge(s): what is the core challenge (or challenges) you want to tackle? What makes your problem hard? What has been holding us back from solving this; i.e., why don't we have an answer to this yet?
- Proposed Approach: brainstorm some approaches you may take to tackle the challenges. why are these approaches promising or feasible? how will you measure success?
- Risks: what are some risks or roadblocks you anticipate?

Latex Template (zip file): project-proposal-latex.zip

Project Proposal: Your Project Title Here

Michael Shell
Email: mshell@ece.gatech.edu

Homer Simpson
Email: homer@thesimpsons.com

Marge Simpson
Email: marge@thesimpsons.com

I. Motivation

(describe the context of your project. What is the setting / environment, tasks, or safety problem you are considering, etc? Who do you think will be most interested in your project? What will your project enable in the future?)

II. Open Challenge(s)

(what is the core challenge (or challenges) you want to tackle? What makes your problem hard? What has been holding us back from solving this; i.e., why don't we have an answer to this yet?). **Make sure you describe what safety means in your project.**

III. Proposed Approach

(brainstorm some approaches you may take to tackle the challenges. why are these approaches promising or feasible? how will you measure success?)

IV. Risks

Class Project

Project Proposal (5%) – due: Feb. 16

1. **Report:** max 1 page project pitch

Mid-term Project (20%)

1. **In-class Oral Presentation:** short conference-style project pitch (~3-5 mins)
   due: Mar 11
2. **Report:** max 4 page writeup of progress
   due: Mar 18

Final Project (35%)

1. **In-class Oral Presentation:** short conference-style lightning talk pitch (~8-10 mins)
   due: Apr 19
2. **Report:** max 6 page writeup of findings
   due: May 1

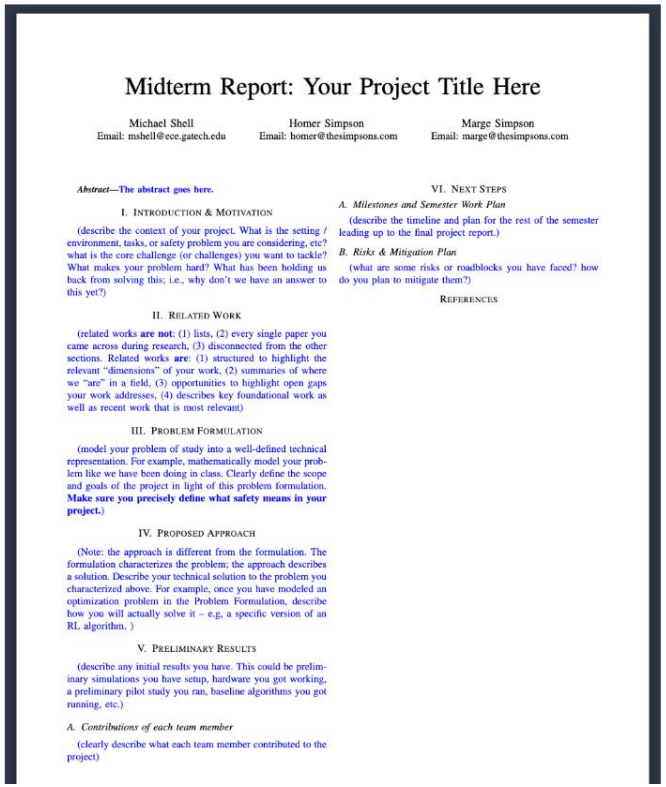
Mid-term Report

Published | Assign To | Edit

This is intended as a checkpoint to ensure that you are making progress towards your final project. The report length should be a typical robotics workshop paper (maximum 4 pages, excluding references).

Please use the attached Latex template and follow the structure of the subsections.

Latex Template (zip file): midterm-report-latex.zip



Midterm Report: Your Project Title Here

Michael Shell
Email: mshell@ece.gatech.edu

Homer Simpson
Email: homer@thesimpsons.com

Marge Simpson
Email: marge@thesimpsons.com

Abstract—The abstract goes here.

I. Introduction & Motivation

(describe the context of your project. What is the setting / environment, tasks, or safety problem you are considering, etc? what is the core challenge (or challenges) you want to tackle? What makes your problem hard? What has been holding us back from solving this; i.e., why don't we have an answer to this yet?)

II. Related Work

(related works **are not:** (1) lists, (2) every single paper you came across during research, (3) disconnected from the other sections. Related works **are:** (1) structured to highlight the relevant "dimensions" of your work, (2) summaries of where we "are" in a field, (3) opportunities to highlight open gaps your work addresses, (4) describes key foundational work as well as recent work that is most relevant)

III. Problem Formulation

(model your problem of study into a well-defined technical representation. For example, mathematically model your problem like we have been doing in class. Clearly define the scope and goals of the project in light of this problem formulation. **Make sure you precisely define what safety means in your project.**)

IV. Proposed Approach

(Note: the approach is different from the formulation. The formulation characterizes the problem; the approach describes a solution. Describe your technical solution to the problem you characterized above. For example, once you have modeled an optimization problem in the Problem Formulation, describe how you will actually solve it – e.g, a specific version of an RL algorithm.)

V. Preliminary Results

(describe any initial results you have. This could be preliminary simulations you have setup, hardware you got working, a preliminary pilot study you ran, baseline algorithms you got running, etc.)

A. Contributions of each team member

(clearly describe what each team member contributed to the project)

VI. Next Steps

A. Milestones and Semester Work Plan

(describe the timeline and plan for the rest of the semester leading up to the final project report.)

B. Risks & Mitigation Plan

(what are some risks or roadblocks you have faced? how do you plan to mitigate them?)

References

Class Project

Project Proposal (5%) – due: Feb. 16

1. **Report:** max 1 page project pitch

Mid-term Project (20%)

1. **In-class Oral Presentation:** short conference-styl project pitch (~3-5 mins)
   due: Mar 11
2. **Report:** max 4 page writeup of progress
   due: Mar 18

Final Project (35%)

1. **In-class Oral Presentation:** short conference-styl lightning talk pitch (~8-10 mins)
   due: Apr 19
2. **Report:** max 6 page writeup of findings
   due: May 1

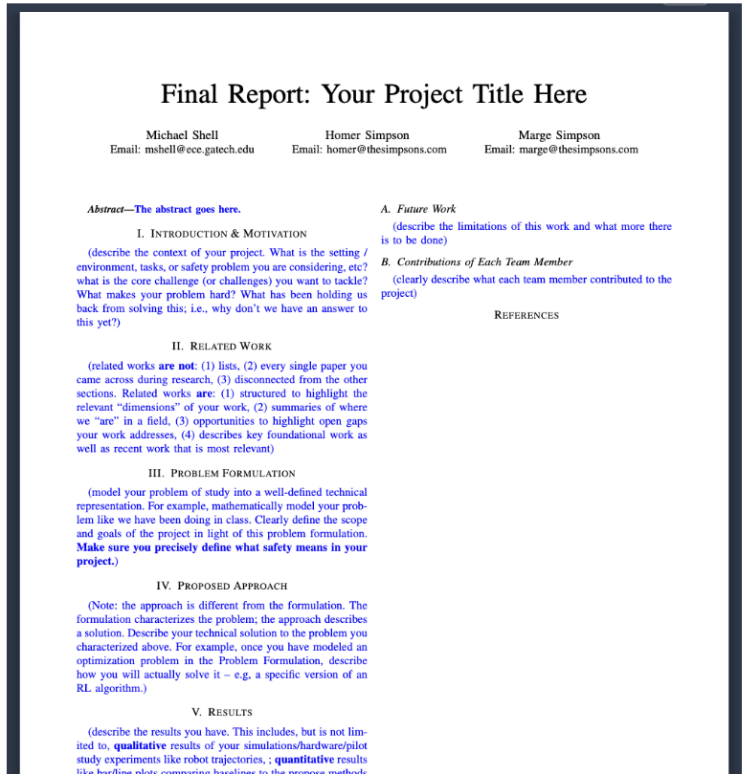
Final Project Report

Published | Assign To | Edit

The final report should present your final findings in a research or survey paper format. The length should be maximum 6 pages, double-column. The grade will be determined based on the content quality and not on the absolute length (please see the grading rubric below).

Please use the attached Latex template and follow the structure of the subsections.

Latex Template (zip file): final-report-latex.zip



Final Report: Your Project Title Here

Michael Shell
Email: mshell@ece.gatech.edu

Homer Simpson
Email: homer@thesimpsons.com

Marge Simpson
Email: marge@thesimpsons.com

Abstract—The abstract goes here.

I. Introduction & Motivation

(describe the context of your project. What is the setting / environment, tasks, or safety problem you are considering, etc? what is the core challenge (or challenges) you want to tackle? What makes your problem hard? What has been holding us back from solving this; i.e., why don't we have an answer to this yet?)

II. Related Work

(related works **are not**: (1) lists, (2) every single paper you came across during research, (3) disconnected from the other sections. Related works **are**: (1) structured to highlight the relevant "dimensions" of your work, (2) summaries of where we "are" in a field, (3) opportunities to highlight open gaps your work addresses, (4) describes key foundational work as well as recent work that is most relevant)

III. Problem Formulation

(model your problem of study into a well-defined technical representation. For example, mathematically model your problem like we have been doing in class. Clearly define the scope and goals of the project in light of this problem formulation. **Make sure you precisely define what safety means in your project.**)

IV. Proposed Approach

(Note: the approach is different from the formulation. The formulation characterizes the problem; the approach describes a solution. Describe your technical solution to the problem you characterized above. For example, once you have modeled an optimization problem in the Problem Formulation, describe how you will actually solve it – e.g, a specific version of an RL algorithm.)

V. Results

(describe the results you have. This includes, but is not limited to, **qualitative** results of your simulations/hardware/pilot study experiments like robot trajectories, ; **quantitative** results

A. Future Work

(describe the limitations of this work and what more there is to be done)

B. Contributions of Each Team Member

(clearly describe what each team member contributed to the project)

References

Survey (5 min)

https://forms.gle/rPSwvYXewJTVvMmZ7

16-886

Embodied AI Safety

Instructor: Andrea Bajcsy

Carnegie Mellon University

intent ROBOTICS LAB