16-886

Embodied AI Safety

Instructor: Andrea Bajcsy

Welcome!

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, 1-2pm
Email: abajcsy@cmu.edu

Teaching Assistant

Ken Nakamura, PhD Student

Research Interests: *Discover synergy between **robust optimal control** and **generative models** to allow robots to safely operate in unstructured environments.*

Office Location: NSH TBA
Office Hours: TBA – **please take survey on Canvas so we can select OHs that suit folks best**
Email: kensuken@andrew.cmu.edu

What is next?

Course Contents

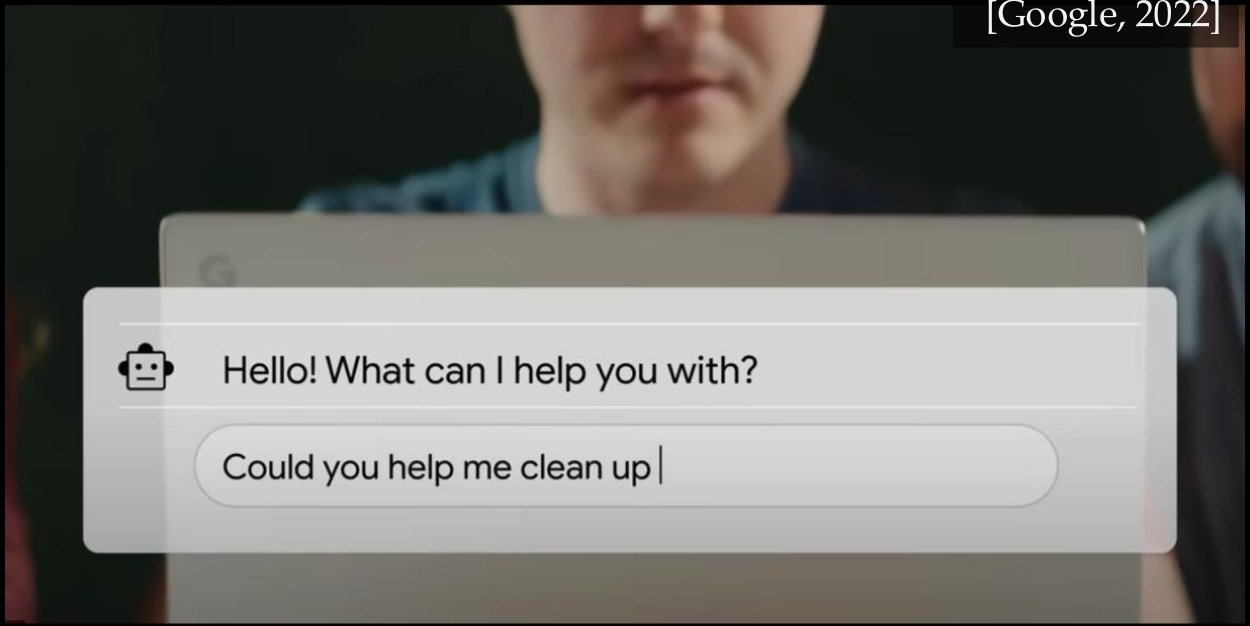
Course Logistics

Intro Survey

(Intro to Sequential Decision-Making)

This class: $\underbrace{\text{Embodied AI}}_{\text{“agents which interact with the environment to accomplish tasks”}}$ Safety

[Waymo, 2023]

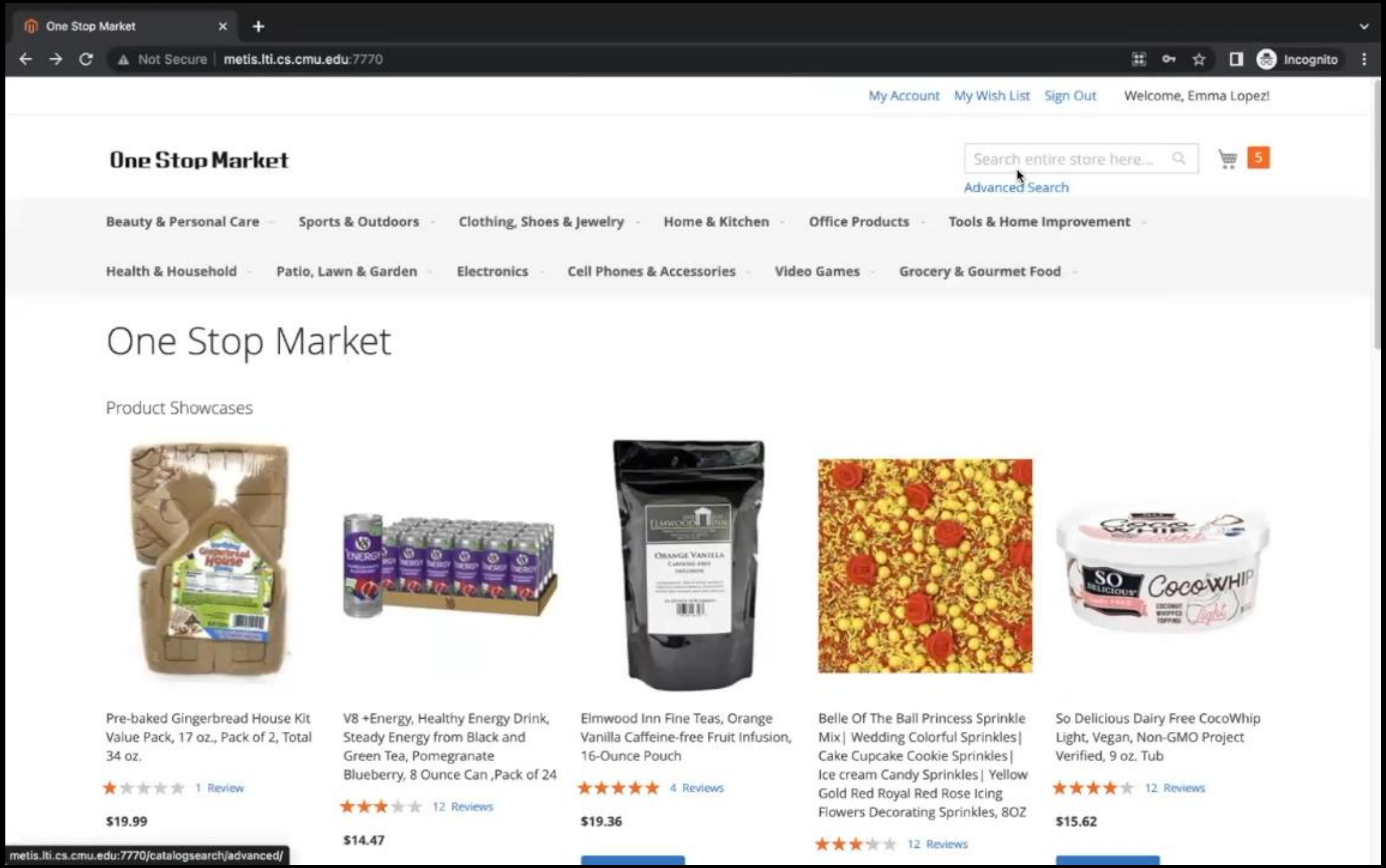


[Google, 2022]

[Skydio, 2023]

[Toyota Research Institute, 2023]

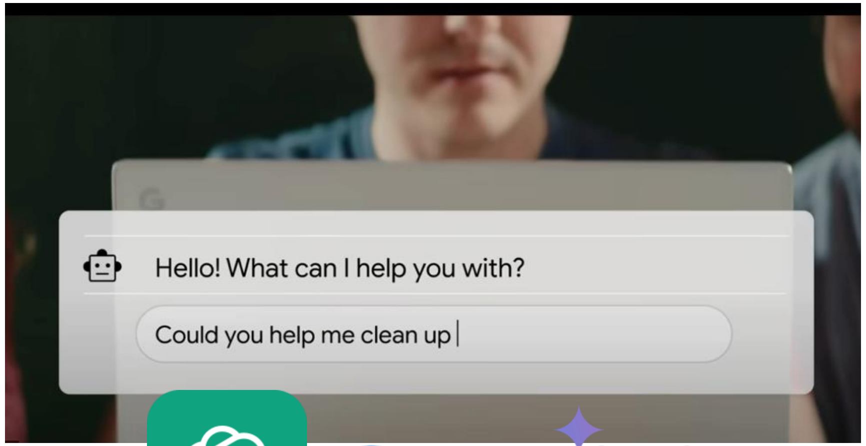
Most examples in this class will be of these EAI systems – **robots**!

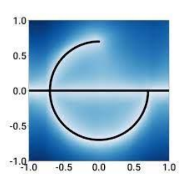


[Zhou et al. “WebArena”, arXiv 2023]

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

World Models, Video Prediction Models, …





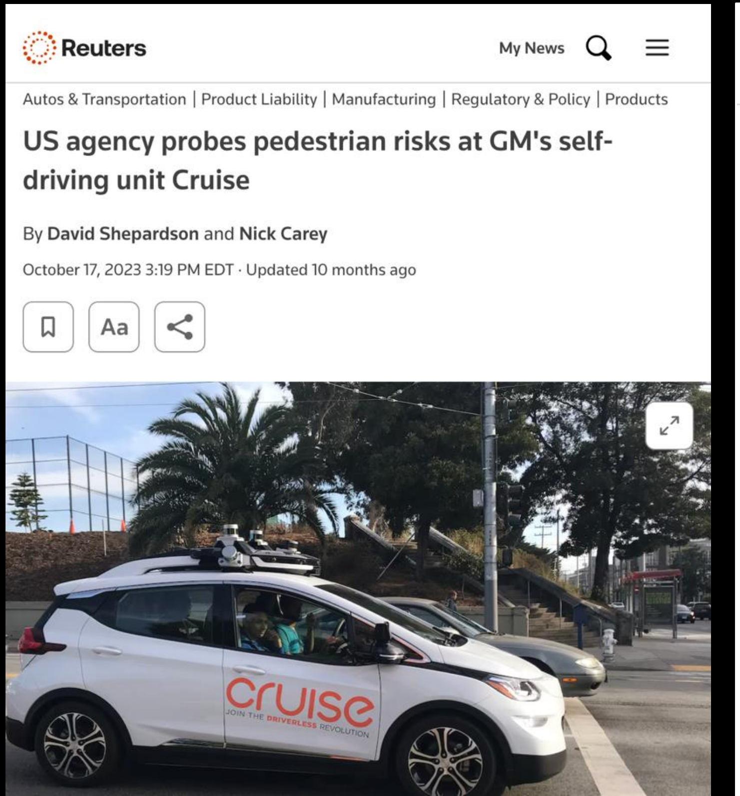
Diffusion Policy

This class: Embodied AI Safety

Some properties you will see in class:

- learned patterns instead of hand-designed ones
- high-dimensional inputs
- “End-to-end” models

Increased capabilities & deployment have escalated concerns about safety


Reuters
My News

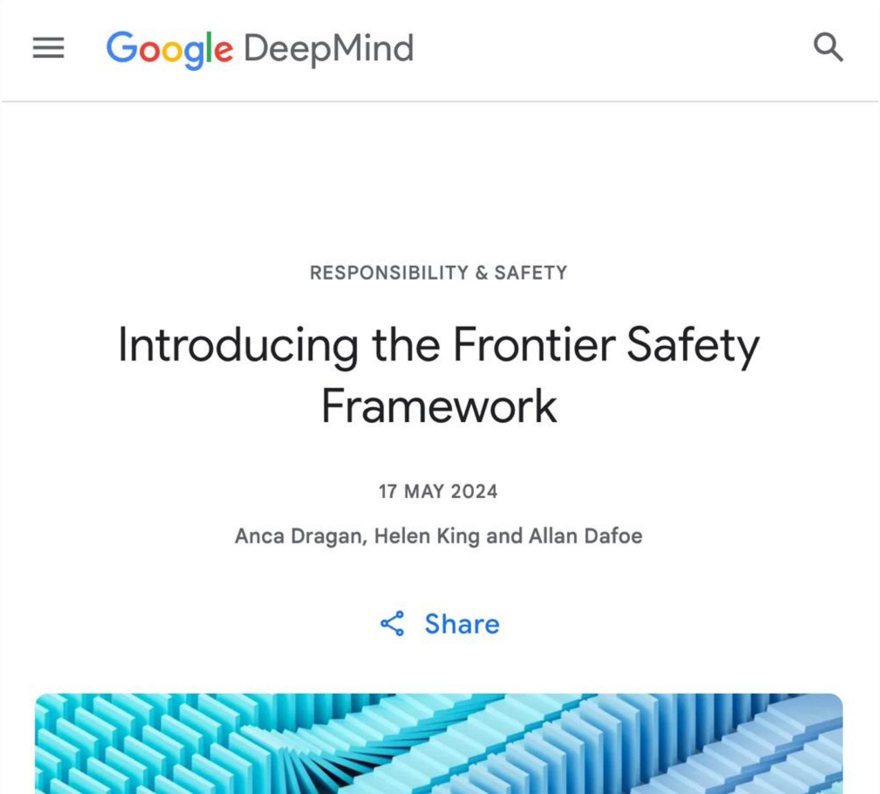
Autos & Transportation | Product Liability | Manufacturing | Regulatory & Policy | Products

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 cou[ld] exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and

MENU

This class: Embodied AI $\underbrace{\text{Safety}}_{\textit{What is safety?}}$

🤔

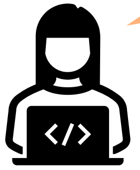
Group exercise!

Group 1	Think of ≥ 3 ways you can define or specify safety for embodied AI systems. Think of the example systems from a few slides ago to motivate your ideas.
Group 2	Imagine you bought a mobile manipulator that uses an LLM-based task planner to act in your kitchen (like the one we saw from Google). What safety concerns could you imagine arising from this system?
Group 3	Imagine you are deploying a drone to help with firefighting in urban disasters. What "safety assurance" would you want from this system?
Group 4	What makes "embodied AI safety" challenging? Name ≥ 3 challenges you foresee.
Group 5	Name 2 *opportunities* and 2 *challenges* do foundation models (e.g., LLMs/VLMs) bring for embodied AI safety?

unstructured, real-world environments

In the “open world”, safety is a nuanced concept

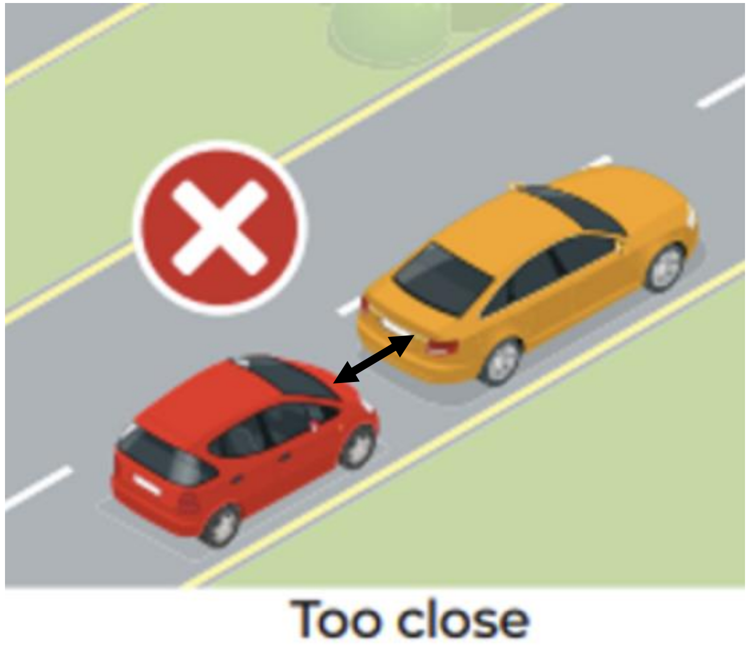
First, let's think through "simple" safety specification….



designer

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

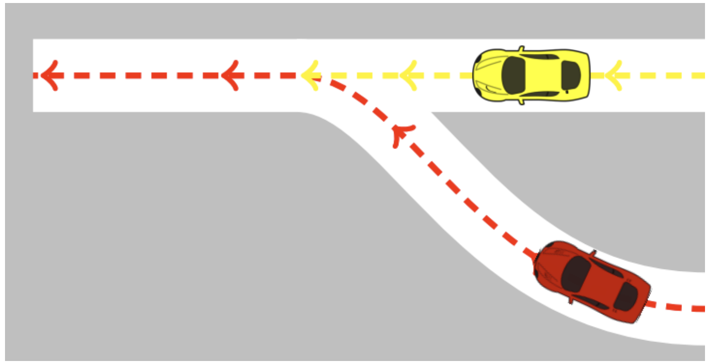
i.e., “don’t collide”



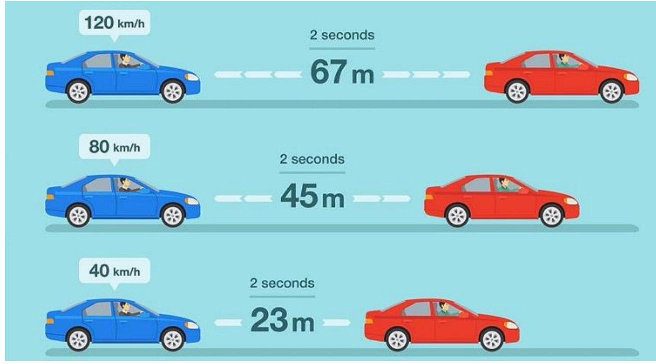
$$\texttt{car\_action} = \begin{cases} \texttt{brake} & \texttt{if d(you, front\_car) < car\_len} \\ \texttt{speed} & \texttt{else} \end{cases}$$

Too close

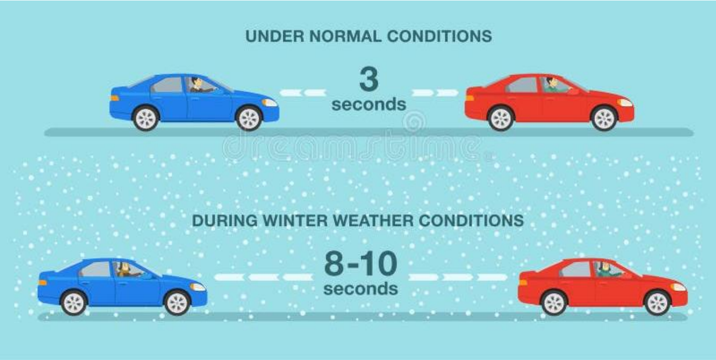
Env. topology



Relative speed



Weather



Many drivers

$$\texttt{car\_action} = \begin{cases} \texttt{brake} & \texttt{if d(you, front\_car) < car\_len} \\ \texttt{speed} & \texttt{else} \end{cases}$$

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

Shai Shalev-Shwartz, Shaked Shammah, Amnon Shashua

In r
parame
addition
that eve

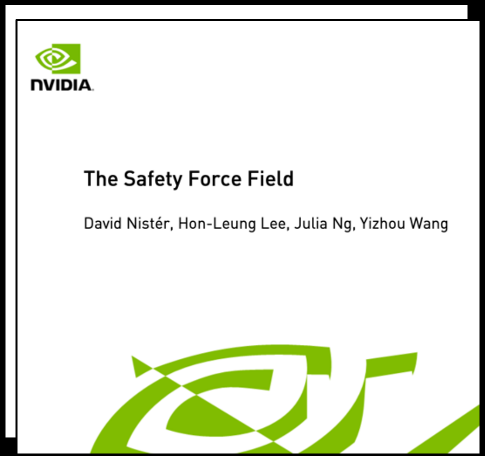
Definition 1 (Safe longitudinal distance — same direction) *A longitudinal distance between a car c_r that drives behind another car c_f, where both cars are driving at the same direction, is safe w.r.t. a response time ρ if for any braking of at most $a_{\max,\text{brake}}$, performed by c_f, if c_r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

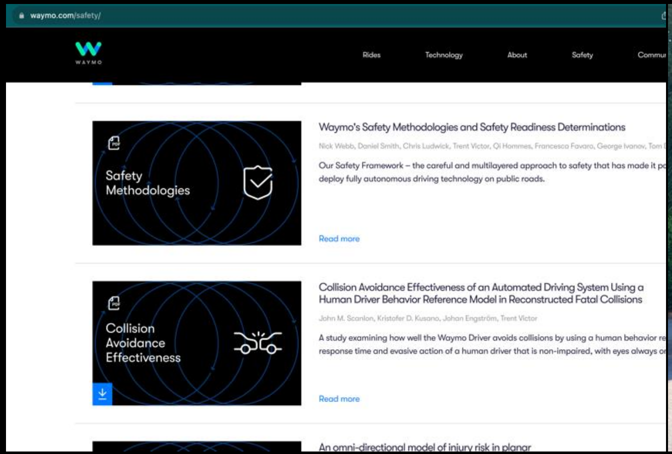
Lemma 2 below calculates the safe distance as a function of the velocities of c_r, c_f and the parameters in the definition.

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:*

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

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





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

Unsafe early braking (Tesla, 2023)

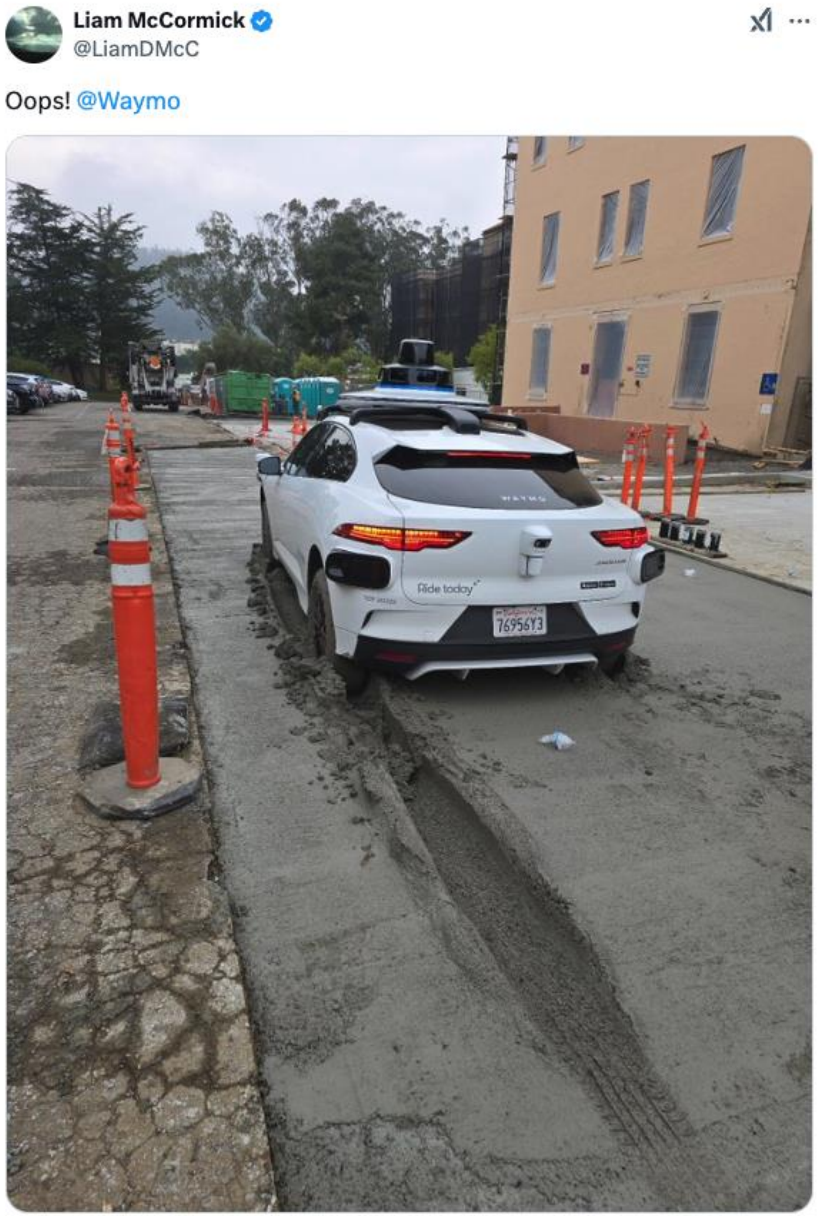
Source: https://abc7news.com/

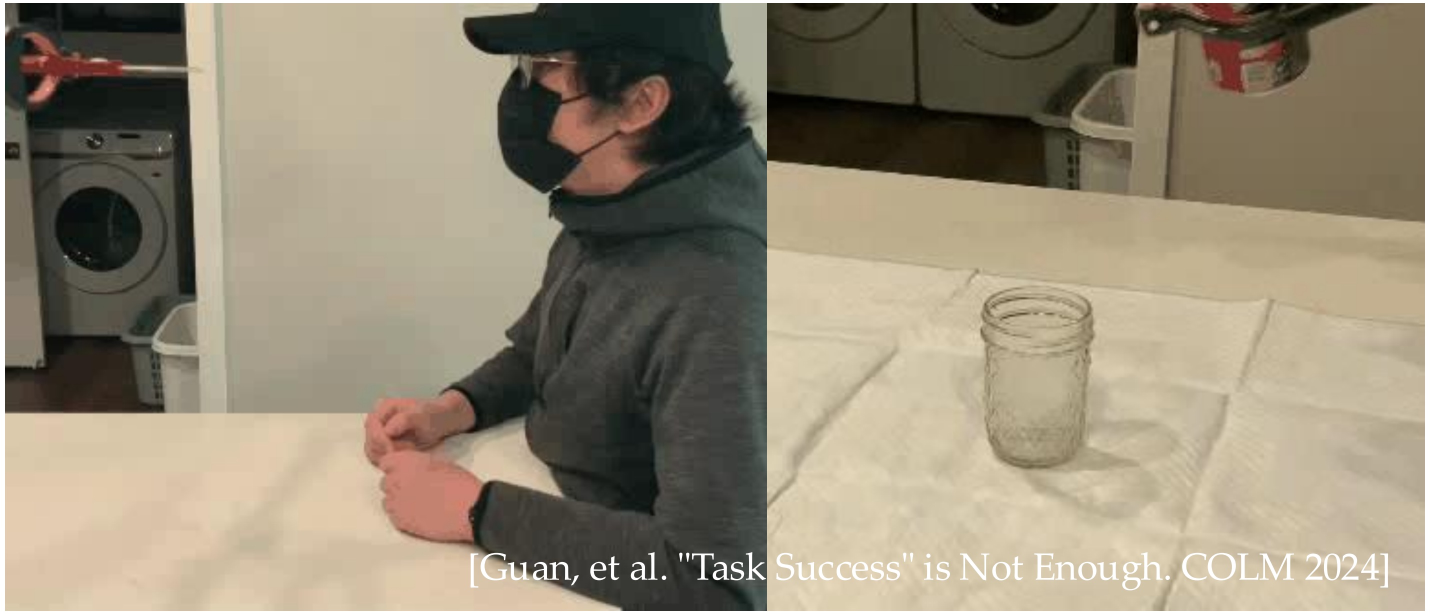
In the open-world, many safety specifications are less obvious….

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 more than just collisions

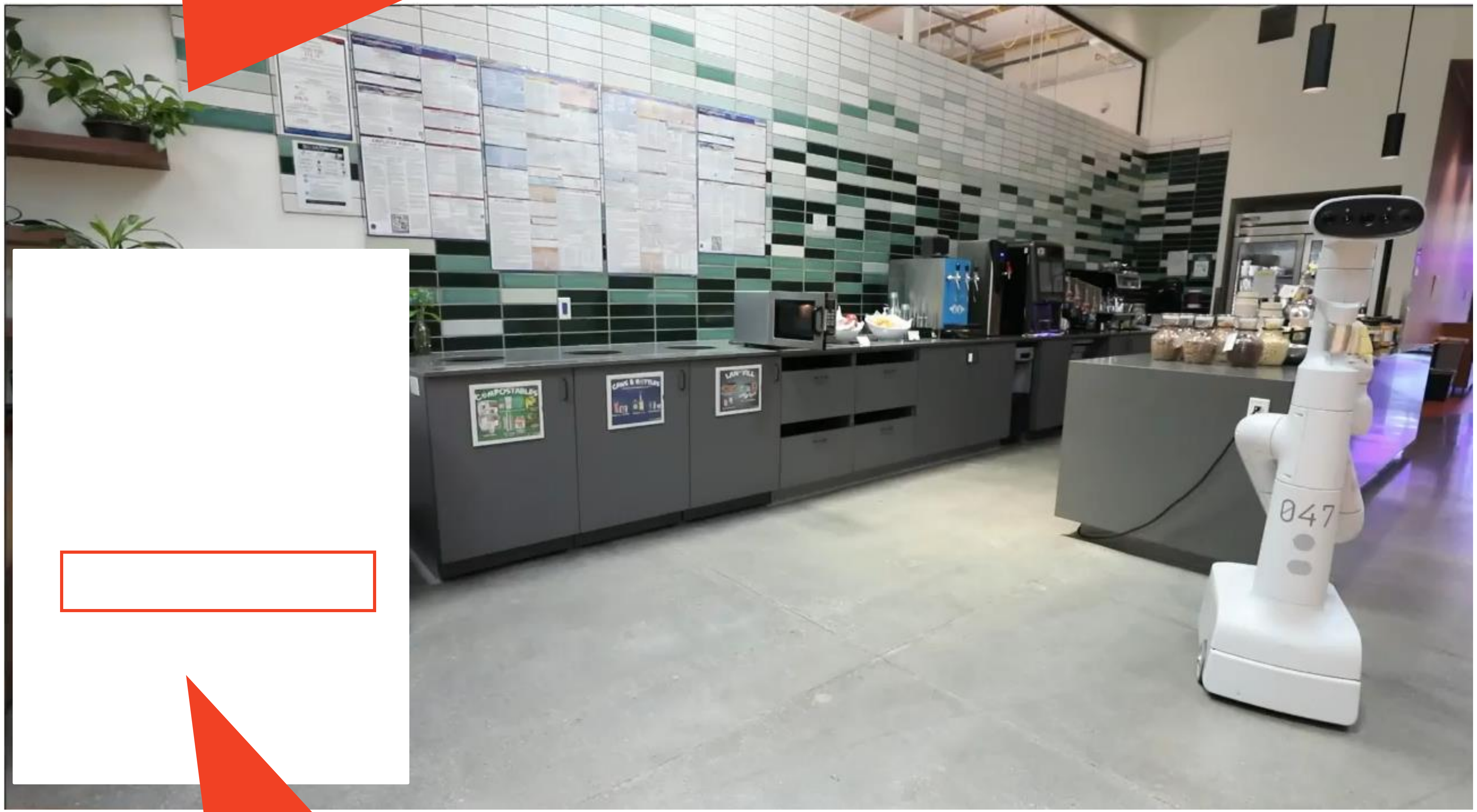




[Guan, et al. "Task Success" is Not Enough. COLM 2024]

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]

Embodied AI safety should reason about anomalous data

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

This class: Embodied $\underbrace{\text{AI Safety}}_{\textit{What is special about this?}}$

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

(promise of) deployment into more unstructured or novel environments

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

(promise of) generalization

generalize safety representations

generate (synthetic) data for stress-testing

(?) 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

Machine Learning / Statistics

how to couple the detection of anomalies with mitigation actions?

how to safeguard *any* AI model?

critique outcomes and steer towards good ones

single erroneous vision / language interpretation can lead to catastrophic action

"misaligned" generations

(promise of) deployment into more unstructured or novel environments

generalize safety representations

generate (synthetic) data for stress-testing

high inference latency

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?

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 2025. 16-886. Monday / Wednesday 11:00-12:20.

Announcements

Hello!

Nov 12 · 0 min read

See you next semester! 🤓

Course Overview

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 r 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 caut tape should never be breached. However, designing systems with such a nuanced understanding is an outstanding challenge, especially in the era of large r 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 predicto...

Schedule (Tentative)

Control-Theoretic Safety Foundations

Jan. 13:	Course Overview	Syllabus
Jan. 15:	Sequential Decision-Making	
Jan. 20:	NO CLASS MLK Day	
Jan. 22:	Safety Filtering	Data-Driven Safety Filters, Model Predictive Sheilding, Safety & Liveness of Filters
Jan. 27:	Safety Filter Synthesis via Optimal Control	
Jan. 29:	Robust Safety	Differential Games I, HJI
Feb. 3:	Computational Frameworks for Safety I	Discounted Reachability, ISAACS
Feb. 5:	Computational Frameworks for Safety II	HW #1 DUE DeepReach, One Filter to Deploy Them All

Frontiers I

Feb. 10:	Semantic Safety I	Safety Representations from Language, Local Updates
Feb. 12:	Semantic Safety II	PAPER READING Semantically Safe Robot Manipulation, SALT
Feb. 17:	Belief-Space Safety	Deception Game, Analyzing Models that Adapt Online
Feb. 19:	Latent-Space Safety I	Dreamer, Human-AI Safety
Feb. 24:	Latent-Space Safety II	PAPER READING TBA, LS3
Feb. 26:	Failure Monitoring & Recovery via VLMs	HW #2 DUE PAPER READING LLM Fallbacks

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*

Attendance	(10%)
Homework (x3)	(30%)
Paper summaries	(10%)
Midterm project report	(10%)
Final project	(40%)

Attendance (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 (30%)

HW #1: Computing & Using Safety Filters

Released: ~Jan 22
Due: Feb 5

HW #2: Generalizing Safety Filters

Released: ~Feb 10
Due: Feb 26

HW #3: Conformal Prediction for Object Classification

Released: ~Mar 10
Due: April 2

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 (10%)

Feb. 12:	**Semantic Safety II**	PAPER READING Semantically Safe Robot Manipulation, SALT
Feb. 17:	**Belief-Space Safety**	Deception Game, Analyzing Models that Adapt Online
Feb. 19:	**Latent-Space Safety I**	Dreamer, Human-AI Safety
Feb. 24:	**Latent-Space Safety II**	PAPER READING TBA, LS3
Feb. 26:	**Failure Monitoring & Recovery via VLMs**	HW #2 DUE PAPER READING LLM Fallbacks, AHA

Paper discussion days:

~7 paper reading days

2 papers per reading day

Before class:

write 1-2 paragraphs of paper review / takeaway / questions **(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

Midterm Report (10%) & Final Project (40%)

Two options:

Research project:

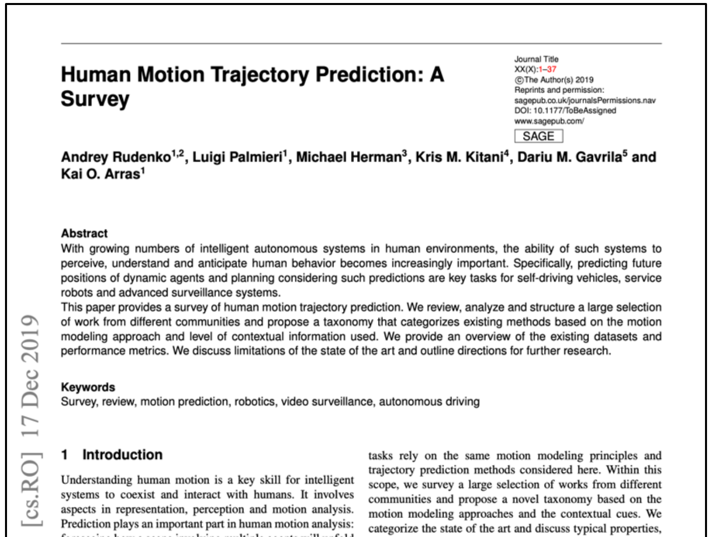
Identify a research direction broadly relevant to this class
Propose and take first steps towards an original idea

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)

Can work individually, or in groups of up to 3.

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

[cs.RO] 17 Dec 2019

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,

Midterm Report (10%) & Final Project (40%)

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

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

Come talk to us about your interests and we can help!

<u>Some examples:</u>

- Applying one of the techniques from class to your problem domain *(e.g., using conformal prediction to calibrate your pose estimator; using a safety filter to shield your policy, …)*
- Comparing two methods that seek to solve the same problem *(e.g., RL vs. SSL approach to computing safety filters)*
- Posing (and solving) a new decision-theoretic safety problem for your problem domain
- Posing (and using) a new uncertainty quantification approach for your problem domain
- Challenging an assumption underlying one of the methods in the class

Midterm Report (10%) & Final Project (40%)

Extra credit opportunity (+2%) – discuss your project with a course staff before Spring Break

After you do this, write a 1 paragraph summary of your idea and discussion which you will **submit to Canvas**

Mid-term report (10%) -- due on Wed, March 12

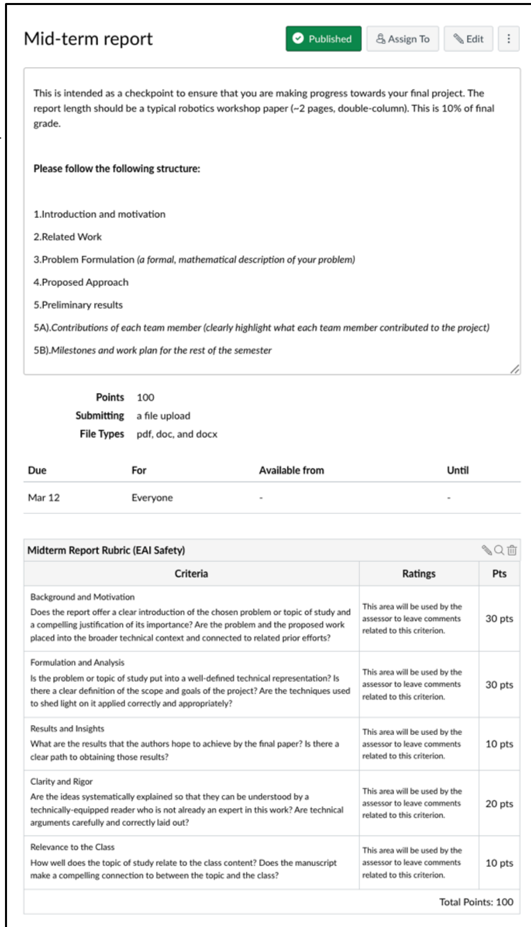
2 page writeup of progress, updated goals and timeline

Oral project presentation (10%) -- to be scheduled for April 21 & April 23

short "conference-talk" presentations (~5 minutes)

Final project report (30%) -- due on May 1

4-6 pages final report



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 (~2 pages, double-column). This is 10% of final grade.

Please follow the following structure:

1.Introduction and motivation

2.Related Work

3.Problem Formulation *(a formal, mathematical description of your problem)*

4.Proposed Approach

5.Preliminary results

5A).*Contributions of each team member (clearly highlight what each team member contributed to the project)*

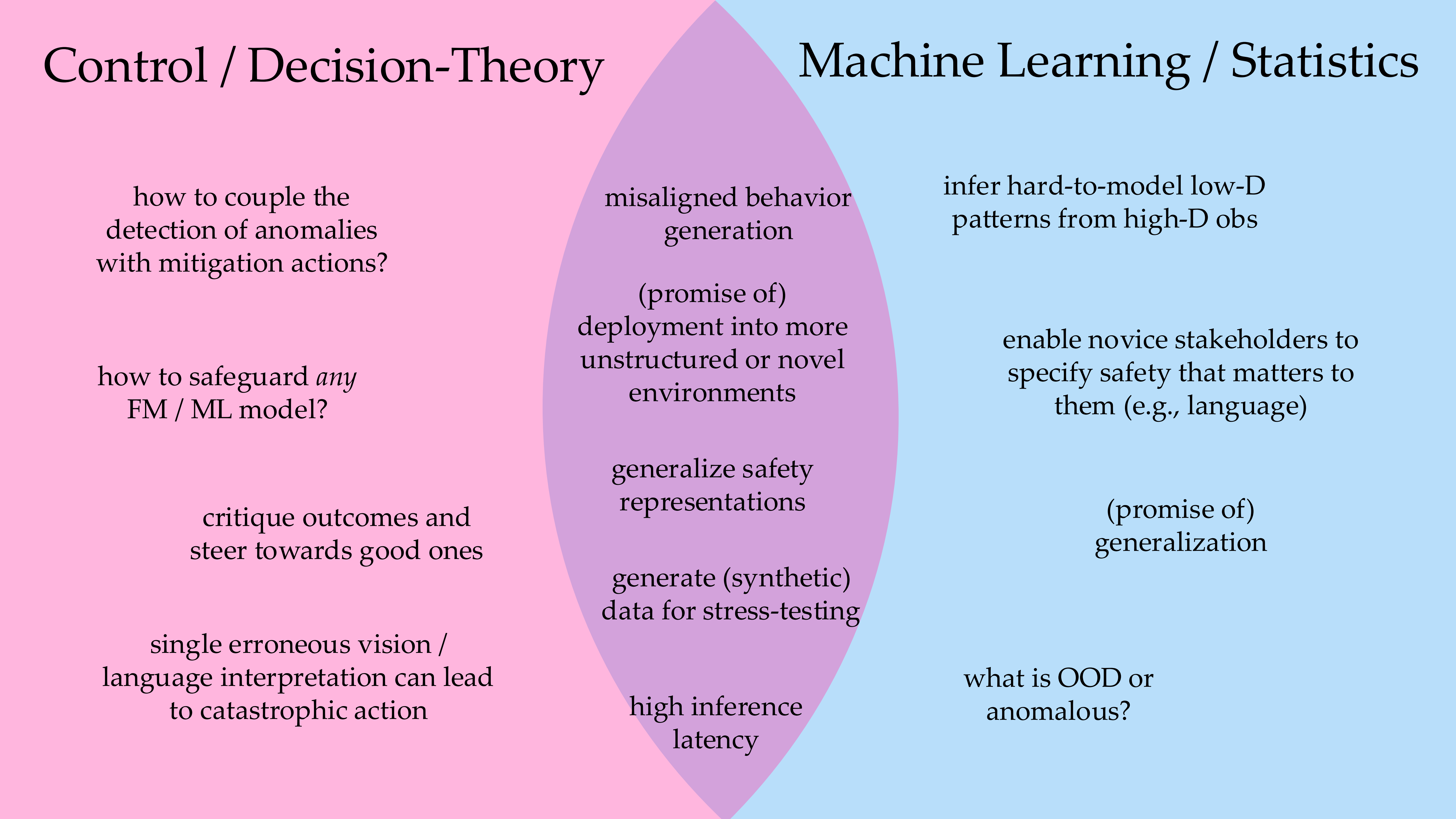
5B).*Milestones and work plan for the rest of the semester*

Points 100
Submitting a file upload
File Types pdf, doc, and docx

Due	For	Available from	Until
Mar 12	Everyone	-	-

Midterm Report Rubric (EAI Safety)

Criteria	Ratings	Pts
Background and Motivation Does the report offer a clear introduction of the chosen problem or topic of study and a compelling justification of its importance? Are the problem and the proposed work placed into the broader technical context and connected to related prior efforts?	This area will be used by the assessor to leave comments related to this criterion.	30 pts
Formulation and Analysis Is the problem or topic of study put into a well-defined technical representation? Is there a clear definition of the scope and goals of the project? Are the techniques used to shed light on it applied correctly and appropriately?	This area will be used by the assessor to leave comments related to this criterion.	30 pts
Results and Insights What are the results that the authors hope to achieve by the final paper? Is there a clear path to obtaining those results?	This area will be used by the assessor to leave comments related to this criterion.	10 pts
Clarity and Rigor Are the ideas systematically explained so that they can be understood by a technically-equipped reader who is not already an expert in this work? Are technical arguments carefully and correctly laid out?	This area will be used by the assessor to leave comments related to this criterion.	20 pts
Relevance to the Class How well does the topic of study relate to the class content? Does the manuscript make a compelling connection to between the topic and the class?	This area will be used by the assessor to leave comments related to this criterion.	10 pts
		Total Points: 100



Control / Decision-Theory
Machine Learning / Statistics
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
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?

What you will learn in this course

Control-Theoretic Safety Foundations

Safety filtering (theory and computation)
Computational frameworks for safety (RL & self-supervised learning)

Frontiers I

Semantic safety & the use of VLMs
Belief and latent-space safety

Machine Learning & Statistical Safety Foundations

Uncertainty quantification (e.g., ensembles, conformal prediction)
AI Alignment
Risk and anomalies

Frontiers II

Out-of-distribution detection & controlling in-distribution
Statistical assurances on learned policies / models

Guest Lectures

Latent Safety Filters

Ken Nakamura
PhD Student @ CMU

Conformal Prediction

Anushri Dixit
Prof @ UCLA

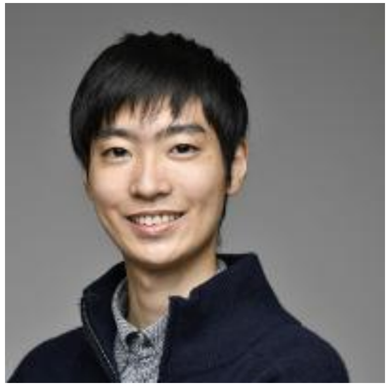
Mathematical Foundations of Robotic Behavior Cloning

Max Simchowitz
Prof @ CMU (MLD)

Out-of-Distribution & Failure Detection

Dr. Masha Itkina
Research Scientist, Toyota Research Institute

Statistical Assurances for Learned Policies



Dr. Haruki Nishimura
Research Scientist, Toyota Research Institute

Survey (5 min)

https://forms.gle/2eX9GZZrNPKe65T69

16-886

Embodied AI Safety

Instructor: Andrea Bajcsy

Carnegie Mellon University

intent ROBOTICS LAB