16-886: Special Topics: *Embodied AI Safety*

Spring 2025

Time:	Mon & Wed, 11:00 - 12:20pm
Location:	GHC 4101
Professor:	Andrea Bajcsy (`abajcsy@cmu.edu`)
Office Hours:	Wed, 12:20 - 1:00 pm (NSH 4629) or *by appointment*
Class Website:	https://abajcsy.github.io/embodied-ai-safety/
Canvas Page:	TBA

1 Course Description

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 an understanding remains an open challenge.

In this graduate seminar class, we study the question of if (and how) 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 be several guest lectures from experts in the field. Students will practice essential research skills including reviewing papers, writing project proposals, and technical communication.

2 Prerequisites

The course is open to graduate students without strict prerequisites. Familiarity with differential equations, probability, and linear algebra is highly encouraged. Interested undergraduate students with a strong background may seek approval from the instructor.

3 Grading

Participation	5%	Regular attendance and engagement
Homework	10%	1-2 hands-on coding experiences
Paper Summaries	10%	1-2 paragraph summary of reading on Canvas
Paper Presentations	15%	In-class presentations on readings
Midterm Project Report	20%	Literature survey & preliminary results ($\sim$ 2 pages, 2 column)
Final Project	40%	Final project report ($\sim$ 6 pages, 2 column) and a final presentation

4 Learning Objectives

At the end of this course, you will have:

- gained knowledge about how to mathematically formalize safety problems arising within modern embodied AI systems (e.g., autonomous drones / vehicles, mobile manipulators, LLM/VLM agents, etc.),
- understood how to apply both decision-theoretic and machine learning safety techniques,
- understood the research frontiers of reliably deploying embodied AI systems in the "open world".

From a research perspective, you will be able to:

- plan a research project and take the first steps (e.g., preliminary results in toy scenario),
- critique research papers,
- contextualize technical work within the broader relevant literature,
- prepare a scientific presentation or talk.

5 Paper Summaries & Presentations

Paper Summaries. There will be several paper discussion days during which you will be assigned research papers to read. You are expected to complete all assigned readings before class and come prepared with comments and questions to discuss with the group. You will share 1–2 paragraphs with your takeaways or questions on each reading on Canvas, by **10am ET of the day the reading** will be discussed.

Paper Presentations. During paper discussion days, we will dive into two papers. During the very first discussion day, I will randomly assign you into groups that you will keep throughout the semester. On each paper discussion day, there will be a set of discussion topics I have generated for each of the papers. In your group, you will discuss the assigned topics. In each group, one person will be randomly assigned to be the group representative who, after the in-class discussion period, will come up and present on the group's conclusions. The whole class will engage the presenter on their conclusions and takeaways. (Note: this paper presentation structure is subject to change based on class size).

6 Class Project

Your class project can be either a thorough literature review ($\sim$ 50 relevant papers, organized so that it identifies gaps in the state of the art) or an exploration of an original research idea. You can choose to work individually or in groups of up to three. The deliverables for the project are as follows (due by **midnight ET**):

Project proposal (due on Feb. 3 | 0% of final grade). This is a brief (1 page) summary of your final project. Think of this as an extended abstract: you want to motivate the topic you have chosen and the technical questions that you want to investigate. By this stage, you should have decided if you are doing a project on your own or in a group. This required proposal will allow me to give you early feedback and help you refine the project scope.

Mid-term report (due on March 17 | 20% of final grade). 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). You can either submit a preliminary literature review, or preliminary exploration of your research idea.

Oral project presentation (to be scheduled for Apr. 21 & Apr. 23 | 10% of final grade). Presentations will be 10 minutes long with 5 minutes for questions. A good presentation will clearly identify the problem in the context of state of the art, state the key idea of the project, and include early results if applicable. *All presentation slides should be submitted by the day before presentations start (**11:59 pm ET, April 21**).*

Final project report (due on May 1 | 30% of final grade). The final report should present your final findings in a research or survey paper format. Target length should be a typical robotics conference paper ($\sim$ 6 pages, double-column).

All project deliverables will be submitted through Canvas under the corresponding Assignment. Only one submission per team is required as long as all team members are clearly identified. All presentation slides should be submitted by the day before presentations start (**11:59 pm ET, April 21**).

6.1 Project Proposal Tips

If you are doing a *research project*. You should 1) motivate the problem, 2) describe how state of the art tackles this problem, 3) what is missing from state of the art, 4) what is your key idea, and 5) describe the scope of what you would investigate (e.g., toy example that you want to study that exhibits your problem, simulator you will use, preliminary data you want to collect, etc.).

If you are doing a *literature survey*. You should 1) describe the topic area, 2) describe the method for how you will find papers (e.g., what conferences, keywords, etc.), 3) describe the inclusion criteria (i.e., what makes a paper relevant for your survey?), 4) provide 5-10 initial papers, 5) describe the key dimensions in which you are categorizing your topic area. For an example of a strong literature survey see:

> Rudenko, Andrey, et al, "Human Motion Trajectory Prediction: A Survey". The International Journal of Robotics Research 39.8 (2020): 895-935.

7 Resources

While there is no official textbook for this course, the following are companion textbooks can provide useful further reading:

* Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*
* Tamer Basar and Geert Jan Olsder, *Dynamic Noncooperative Game Theory*, 2nd Edition
* Dimitri Bertsekas, *Reinforcement Learning and Optimal Control*
* Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*
* Jorge Nocedal and Stephen J. Wright, *Numerical Optimization*

8 Attendance

Class attendance and participation are key for both your and your peer's success in this class. You are expected to attend class in person during the scheduled time, including the final presentations. I understand that occasionally you may have challenges attending (e.g., illness, religious observance, etc.). However, if you anticipate having a challenge regularly attending class, please contact me.

9 Academic integrity

Honesty and transparency are important features of good scholarship. On the flip side, plagiarism and cheating are serious academic offenses with serious consequences. If you are discovered engaging in either behavior in this course, you will earn a failing grade on the assignment in question, and further disciplinary action may be taken.

I encourage you to work together on projects and homework assignments and to make use of campus resources like Student Academic Success Center (SASC) to assist you in your pursuit of academic excellence. However, please note that in accord with the university's policy you must acknowledge any collaboration or assistance that you receive on work that is to be graded, either from a person, reference, or a tool (including AI-generation tools like ChatGPT).

10 Late Policy

All homeworks and assignments are assigned due dates and should be submitted through the relevant Canvas portal. If you cannot submit an assignment on time, my default will be to reduce the grade by 10% for each 24 hour period, up to three days, that the assignment is late. This will be automatically applied; you do not have to request it. After three days, the assignment will receive a zero. If you experience an unforseeable emergency and would like me to consider waiving the late penalty, please email me as early as possible to discuss this request. The 10% per day deduction does not apply to unexcused late presentations, which will receive a zero immediately, because they will affect our ability to hold class. Re-scheduling presentations will be based on schedule availability and the professor's discretion.

11 Accommodations for students with disabilities

If you would like to receive accommodation for a documented disability, please first contact Disability Resources (`access@andrew.cmu.edu` or 412-268-2013). Let me know as soon as possible so we can discuss reasonable accommodations. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at `access@andrew.cmu.edu`.

12 Student wellness

Take care of yourself. Do your best to maintain a healthy lifestyle by eating well, getting enough sleep, and taking some time to relax. This will help you achieve your goals and cope with stress. If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.