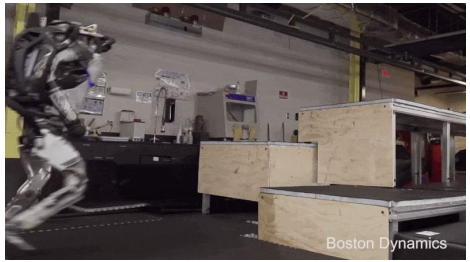
Autonomous Agents

Eri Phinisee & Dylan Pollak

15-482 Autonomous Agents



Course description

Autonomous agents use perception, cognition, actuation, and learning to reliably achieve desired goals without human intervention.

Ex. Smart homes, mobile robots, intelligent factories, self-driving cars

This course provides students with the **techniques needed for developing** complete, integrated **AI-based autonomous agents**. The course is project-oriented where, over the course of the semester, small teams of students will **design, implement, and evaluate** autonomous agents operating in a real-world environment.

Course objectives Throughout the course, students learn to analyze and implement...

- Architectures for intelligent agents
- Task planning
- Reasoning under uncertainty
- Optimization
- Monitoring
- Execution
- Error detection and recovery
- Collaborative and adversarial multi-agent interaction
- Machine learning
- Ethical behavior
- Explanation

Constraints

Update a pre-existing ethics module: in-class lecture at the end of the semester*

Class size is ~11 with mainly juniors & seniors

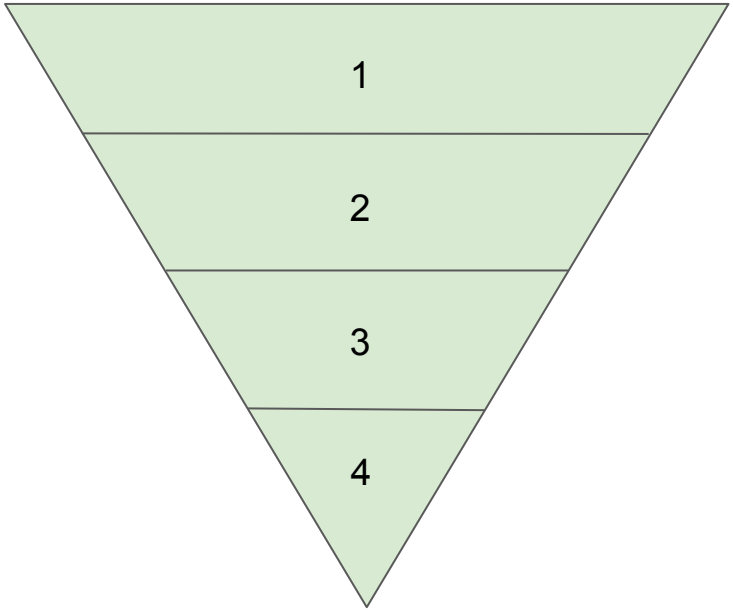
Designed for students with technical backgrounds

*However, we are also creating ethics questions to be distributed with the written assignments so that our ethics module can complement their learning experience throughout the course

Learning Objectives

Theme: Give students exposure and confidence on various topics surrounding autonomous agents

1. **Recognize and describe** how an autonomous agent may cause harm in **a situation** (e.g., privacy breach, violence, lack of transparency, bias, manipulation of human behavior), including the interplay of different technical, legal, and societal components that enabled the ethical failure
2. **Apply a framework** of ethical analysis across a variety of different autonomous applications **to assess the benefits and cost** associated with each autonomous application
3. **Evaluate and defend the ethicality of choices** made during the development and deployment of an autonomous application
4. **Propose preventative safeguards** against potential unethical uses of autonomous applications



Implementation Design

Pre-module survey

Assignment questions

Over multiple weeks

Informational presentation

20 minutes

Group activity

40 + 20 minutes

Day of lecture

Post-module survey

Surveys

Pre-Module Survey

- Provide specific ethical considerations and defend their stance
- Rate comfort level with each of the learning objectives

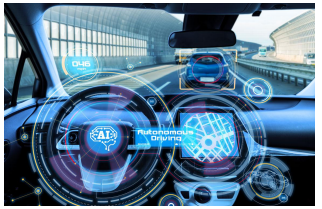
Post-Module Survey

- Provide specific ethical considerations and defend their stance
- Rate comfort level with each of the learning objectives
- ✚ Rate satisfaction
  - Overall
  - Components

Autonomous Agents

Voice data collection intervention with domestic violence scenario

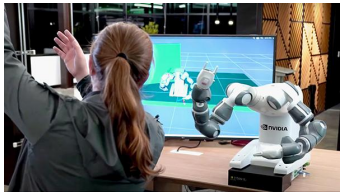
Lethal autonomous weapon



Safety vs. innovation in autonomous vehicle research

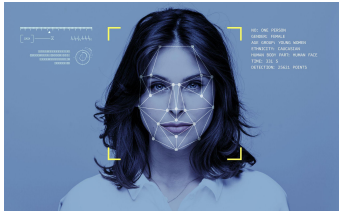
Autonomous optimization in ER triage

Interpretation of “appropriate” uses of technology



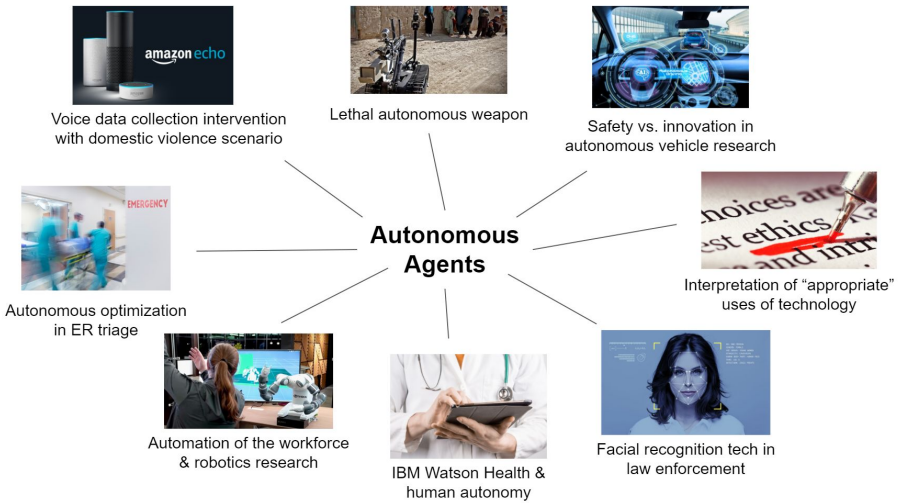
Automation of the workforce & robotics research

IBM Watson Health & human autonomy

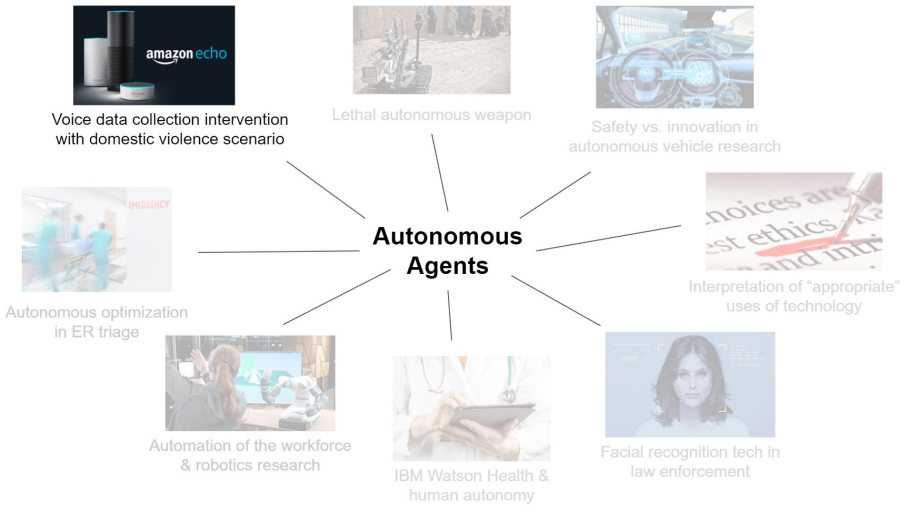


Facial recognition tech in law enforcement

Sample Assignment Questions



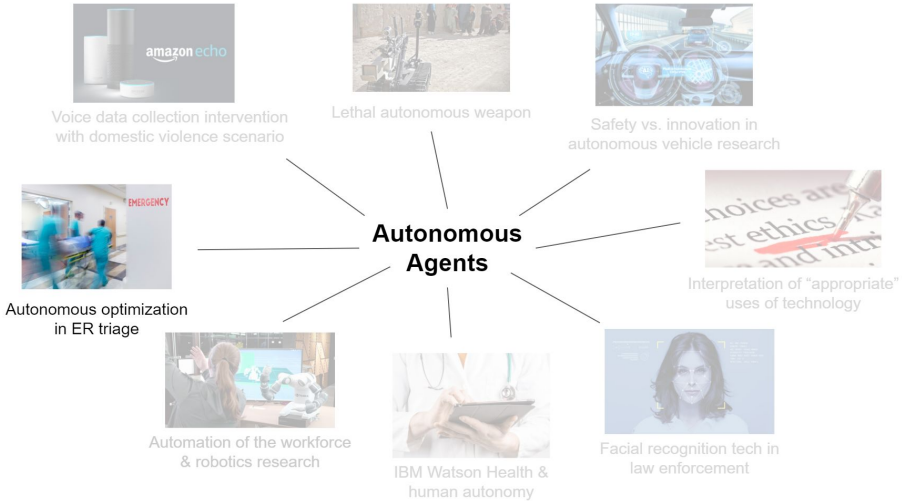
Sample Assignment Questions



Should a smart home device like Amazon Alexa turn on and record a conversation upon detecting signs of domestic violence (physical and/or emotional)? Furthermore, should it notify law enforcement? What safeguards can be implemented in order to avoid erroneously detecting movies and TV shows as violent dialogue as opposed to the actual people in the house?

Learning objectives 1, 2, 3, 4

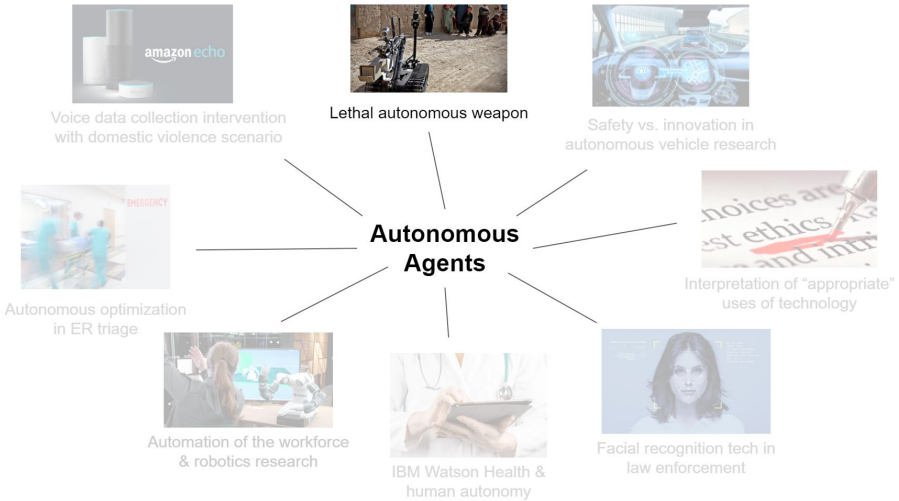
Sample Assignment Questions



Imagine you are developing an autonomous agent that assists with emergency room triage. What are some considerations that the agent should incorporate in making decisions about optimizing patient care, and what is an example scenario in which the optimization method may pose an ethical issue?

Learning objectives 1, 2, 3, 4

Sample Assignment Questions



A lethal autonomous weapon can operate on the battlefield without explicit instructions from a human. What are some ways in which the weapon may trigger inappropriate actions? In order to appropriately respond to cases in which it malfunctions or crashes, how can a human agent be effectively placed in the operational cycle of the weapon?

Learning objectives 1, 4

Informational Presentation

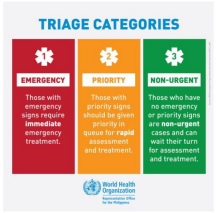
Emergency Room Triage

Assess patients' severity of injury or illness, assign priorities, and provide the appropriate treatment

What are some considerations an autonomous agent should incorporate in ethically optimizing the process?

Ex.

- Predictions of patient traffic
- Physical, mental, and emotional burden on ER staff
- Capacity at other hospitals
- Focus on the most serious cases first (normal) or save as many patients as possible (disaster)



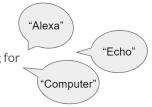
Lethal Autonomous Weapons (LAWS)

Systems that can target and fire deadly weapons without the need for human intervention

Pros	Cons
• Force multiplier • Expand the battlefield • Reduce soldier casualties • No emotional misjudgement or cognitive deterioration • Act as an objective reporter of ethical infractions	• Learning on the job -> Prone to error • Algorithmically unreliable • Violate the Principle of Distinction • Hard to determine accountability, does not meet *jus in bello* • Having humans in the loop defeats the purpose of eliminating human error

How can we address these concerns at the development and deployment stages?

Smart Home Device: Alexa



The Amazon Alexa is an "always-on listening device" waiting for

Police has used Alexa's recordings to pursue criminal cases

Should an Alexa intervene in the case of domestic violence and alert law enforcement as a preventative measure?

- Conversations unrelated to commands will be recorded
  - Potentially accessible to thousands of transcribers
- Lose privacy inside our own house
- Technical mistakes -> Unnecessary police intervention

Group Activity

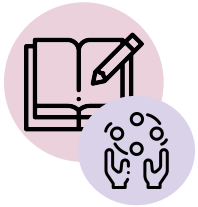
Create and present to class **a code of ethics** for a company they work for

- Think from a developer/designer perspective
- What kinds of guidelines and tests they wish were in place to encourage them towards an ethical development and deployment
- What kinds of consequences they wish were in place to discourage the violation of the code

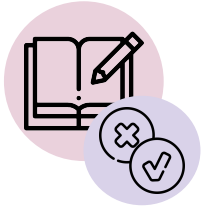
Goal: **Utilize all the skills** they have practiced while learning from **new perspectives**

Group 1: Tesla
Group 2: Northrop Grumman
Group 3: Apple
Group 4: Amazon
Group 5: DeepMind

Improvement



Focusing on agency



Balancing negatives with positives

Using controversial companies for the group activity

Adding a pre-module survey

We couldn’t have done this module without the support of Professor Stephanie, Reid, Illah & Victoria. Huge shoutout to Professor Stephanie for always encouraging us and allowing us to get really involved with the module implementation, we are both really grateful to all of you. We welcome any questions :)