# 15-482 Autonomous Agents: Ethics Module

## Learning Objectives

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

The learning objectives build off of one another in the order presented above. First, our ethics module will make sure that students have the ability to recognize how an ethical issue can arise in a situation with an autonomous agent and to describe what factors contributed to the issue. Next, it will teach students the skill to perform a cost-benefit analysis and will make sure that students can do this regardless of the type of autonomous application. Additionally, students will then learn how to use the prior analysis to then decide and argue for a decision they will make during development and deployment. Finally, students will take all of their knowledge and skills and apply them from a solution-oriented perspective, designing a preventative measure that can actively discourage the development and deployment of a vulnerable autonomous agent.

## Plan

We plan to assess these objectives through the [8 questions](#_yya6ytow9ujr) we will add to the homework assignments and through the final in-class activity. Different assignment questions will target different learning objective(s). The plan is for students to become familiarized and gain practice with these skills so that they can successfully engage in the in-class activity that is to come later in the semester.

On the day of the ethics lecture, a [presentation](https://docs.google.com/presentation/d/1nHMlr5U-EJOXJmUdNhYU6bOoCgkcu7nmLil2pi97N1c/edit?usp=sharing) will serve as an informational review of all the topics covered in the assignment questions. The in-class activity will then be a group activity that will ask students to create and present to class a code of ethics for a company by which they are hypothetically employed and whose ethics committee they are a part of. Each group of students will together be writing the code of ethics from a developer/designer perspective, thinking about what kinds of guidelines and tests they wish were in place to encourage them towards an ethical development and deployment and what kinds of consequences they wish were in place to discourage the violation of the code. During this activity, we expect the students to be as transparent as possible and to voice all of their opinions and concerns during the group discussion as they devise this guiding framework. The goal of this activity is to have the students utilize all the skills they have learned through the assignment questions while benefiting from new perspectives that other group members will be able to supply. All of the learning objectives will be assessed through this activity.

**Schedule**

* Informational presentation: 20 minutes
* Group activity: 40 minutes
* Presentation/Discussion: 20 minutes

|  |  |  |
| --- | --- | --- |
| **Group** | **Companies** | **Topic** |
| Group 1 | Tesla | Self-driving vehicle |
| Group 2 | Northrop Grumman | LAWS |
| Group 3 | Apple | Apple Watch (biometric sensor) |
| Group 4 | Amazon | Alexa |
| Group 5 | DeepMind | AGI research |

Furthermore, we will gauge the effectiveness of the ethics module through a [pre-module survey](https://docs.google.com/forms/d/1LxqV_poacTgo7zJ5FpHH6PqnMNG4t7GSqrjQF1DxAZ4/edit?usp=sharing) and [post-module survey](https://docs.google.com/forms/d/1Ziw449nysLCNDzZc4GsHW10rBvCK5_CT06eU2zJJ0Yc/edit?usp=sharing). Both the pre and post-module surveys ask students to 1) describe some major ethical considerations to take into account in developing and deploying autonomous agents and to briefly explain their stance on the examples they give, and 2) quantitatively rate their comfort level with each of the learning objectives. By comparing the post-survey answers to those from the pre-survey, we can evaluate to what extent the students felt like they gained the skills outlined in the learning objectives through participation in our module, and if they actually did so. Moreover, the post-survey will additionally contain questions about the student’s level of satisfaction with the module itself. They are asked to rate their satisfaction with the overall module as well as with the different components that comprised it, namely the assignment questions, the informational presentation, and the group activity.

**Rubric for Survey Question #1**: What are some major ethical considerations to take into account in developing and deploying autonomous agents? Provide a few specific examples. Briefly explain your stance on the examples you gave.

|  |  |  |  |
| --- | --- | --- | --- |
| Poor | Adequate | Proficient | Excellent |
| The student does not state anything relevant to autonomous agents and/or ethics. | The student cannot describe any ethical considerations and/or formulate a logical argument. | The student can describe a few ethical considerations and formulate an argument but the response is vague and cannot support it with evidence. | The student provides concrete ethical considerations and formulates their argument with specific evidence. |

## Assignment Questions

### Assignment 1

* [Florida law enforcement scenario](https://www.engadget.com/2019-11-02-florida-police-obtain-alexa-recordings-in-murder-case.html#:~:text=Law%20enforcement%20in%20Hallandale%20Beach,and%20his%20wife%20Silvia%20Crespo.): Law enforcement in Hallandale Beach, Florida has used a search warrant to [collect](https://miami.cbslocal.com/2019/10/31/hallandale-beach-police-obtain-amazon-device-recordings-in-bizarre-murder-case/) Alexa recordings from two Echo Dots as part of a murder case. Investigators want to know if the smart speakers inadvertently picked up audio of a July altercation between Adam Crespo and his wife Silvia Crespo.
* As in the past, Amazon stressed in a statement to *CBS* that it doesn't hand over customer information unless mandated by a "legally valid and binding order," and that it resists "overbroad or otherwise inappropriate" requests.
* **Question:** 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

### Assignment 2

* Background [article](https://www.techrepublic.com/article/what-a-potential-sale-of-ibms-watson-health-means-for-ai-in-healthcare/) on AI IBM Watson Health
  + “Watson, the family of AI-based question-answering systems marketed by IBM, represents the state of the art in this field of AI-powered machines that exhibit high levels of human skill. Physicians themselves, faced with the success of Watson at diagnosing human disease from symptoms with very high accuracy as compared to their own peers, have gone on the record to herald the machine’s application in medicine and to recognize the dawn of a new medical age. These are physicians, not computer scientists, who have watched Doctor Watson’s behavior and are convinced that AI’s role in medicine is genuinely revolutionary.” **AI & Humanity, Chapter 5, Illah Nourbakhsh**
* **Question:** Does human autonomy shift when a machine can make more accurate diagnoses than a human? Do you think IBM Watson Health is helpful or not? Please justify your answer.
* Learning objectives 1, 2

### Assignment 3

\*We decided to consolidate our assignment question for assignment 3 with that of assignment 6 since the topics were similar.

### Assignment 4

* *The New Yorker* article: [“Did Uber Steal Google’s Intellectual Property?”](https://drive.google.com/file/d/1fhhPUWkv8TLI_-SxGq48kehkLurS9VKf/view?usp=sharing)
  + In the article, Anthony Levandowski describes his priorities on safety versus innovation: “If it is your job to advance technology, safety cannot be your No. 1 concern... If it is, you’ll never do anything. It’s always safer to leave the car in the driveway. You’ll never learn from a real mistake.”
  + **Question:** With your group members, discuss Levandowski’s stance on how to balance safety with the need to improve performance. Does one need to be prioritized over the other? How might you respond to a colleague holding Levandowski's view?
* Learning objectives 1, 2

### Assignment 5

* Part 3 (READ ME section)
* **Question:** In the part where it asks, “Explain why you believe these models should be appropriate for the problem,” have the student transparently discuss their ethical considerations when choosing the model and define the word appropriate in their own way
* Learning objectives 2, 3

### Assignment 6

* Part 4, written section in explaining the why, how and why not explanations
* Articles on facial recognition technology bias
  + [Machine Bias](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)
  + [A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear.](https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/)
    - “The algorithm, called COMPAS, is used nationwide to decide whether defendants awaiting trial are too dangerous to be released on bail. In May, the investigative news organization ProPublica claimed that COMPAS is biased against black defendants. Northpointe, the Michigan-based company that created the tool, released its own report questioning ProPublica’s analysis. ProPublica rebutted the rebuttal, academic researchers entered the fray, this newspaper’s Wonkblog weighed in, and even the Wisconsin Supreme Court cited the controversy in its recent ruling that upheld the use of COMPAS in sentencing. It’s easy to get lost in the often technical back-and-forth between ProPublica and Northpointe, but at the heart of their disagreement is a subtle ethical question: What does it mean for an algorithm to be fair? Surprisingly, there is a mathematical limit to how fair any algorithm — or human decision-maker — can ever be.”
* Draw a concept diagram visualizing the relationships between all the entities involved in the deployment of a facial recognition technology in law enforcement. Then, describe how biases can propagate through the system, how transparent they are and can be, and how the opinion of each entity should be weighed so that the system can minimize the risk of a wrongful arrest. Describe a case in which the technology actually counteracts the human biases and leads to a more objective investigation.
* Learning objectives 1, 2, 3

### Assignment 7

* Part 3, Diagnosis section
* The potential dangers associated with the deployment of AI-empowered robotic weapons begin with the fact that much of the technology involved is new and untested under the conditions of actual combat, where unpredictable outcomes are the norm. [Source](https://www.armscontrol.org/act/2019-03/features/autonomous-weapons-systems-laws-war)
* **Question:** 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

### Assignment 8

* Video: [Boston Dynamics](https://www.youtube.com/watch?v=Ns__YpRneBA)
* **Summary:** Boston Dynamics focuses on creating robots with advanced mobility, dexterity and intelligence. We have long held that mobility sufficient to access both the natural and the built world required legs. We began the pursuit of this dream over 30 years ago, first in academia and then as part of Boston Dynamics because it was an exciting technical challenge and because to build a highly mobile robot required it. We wanted to build a robot that could go where people go. The commonly referred to “dull, dirty and dangerous” tasks don’t occur solely on a neatly organized factory floor, they pop up in the natural world and in existing infrastructure. These are places where being effective requires deftly maneuvering through rocky trails, staircases, catwalks, doors or narrow cluttered passages. The environment can’t be conformed to the machine, and therefore the machine must be capable in the environment as it comes.
* **Question:** How can you envision robotics research that will complement human workers so that it's not just humans versus robots? The goal would be for humans to be more productive and to find fulfillment in their jobs. Furthermore, how do you envision autonomous agents playing a role in aiding the job search for the laid off workers that were replaced by robots? Please provide your opinion.
* Learning objectives 2, 4

### Assignment 9

* **Question:** 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