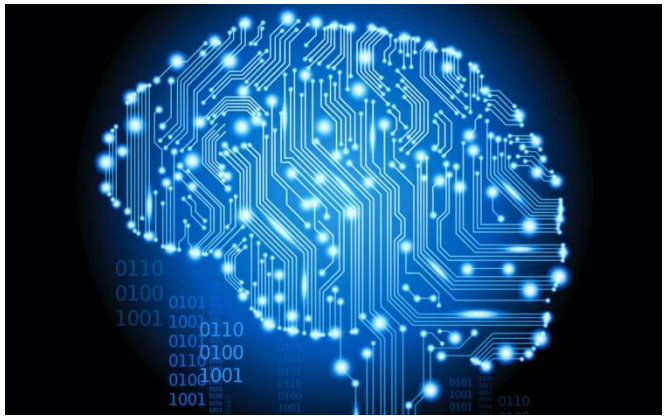
Artificial Intelligence and Identity

ENCS 393: Social and Ethical Dimensions of ICTs

Day 8 – June 1, 2020



Reminder: Quiz this Wednesday

- Quiz questions will be posted on Moodle at noon on Wednesday, June 3rd.
- You must submit your answers on Moodle (same submission format as for the Reflection Essays) by noon on Thursday, June 4th.
- You may refer to lecture slides, lecture recordings, course readings, your own notes, and dictionaries or other language aids while writing the quiz. You may NOT use any other resources and you may not consult with other people.
- The quiz will cover all of the course material (lectures and readings) from the beginning of the course up to and including our June 1st class.
- Make sure that your responses demonstrate your understanding of the course materials!

From Moral Algorithms to Machines with Moral Status

Last class:

- **Dependent agency** and its ethical implications, especially with respect to privacy

Today:

- Critiquing the idea of **moral algorithms/moral machines**
- Thinking about machines with their own **moral status**

Self-Driving Cars as “Moral Machines”

- Ethically desirable because of their potential to reduce the number of collisions and the accompanying loss of human life (utilitarian framing)
- Ethically programmable to respond a certain way in a given situation
- In her text, JafariNaimi critiques both of these ways in which self-driving cars are thought of as “moral machines”

Ethics of Care

JafariNaimi's critique is based in a normative ethical approach called the **ethics of care**.

The ethics of care is a type of virtue ethics that prioritizes characteristics such as empathy and compassion. In general, an ethics of care assumes that:

- People are interrelated and dependant on one another, to varying degrees (here's that dependent agency issue again!)
- The most vulnerable among us deserve particular consideration when we consider the consequences of various possible actions.
- Situational details and immediate contexts are important to consider when making ethical choices.

A View from **Emerging Technology from the arXiv**

Why Self-Driving Cars Must Be Programmed to Kill

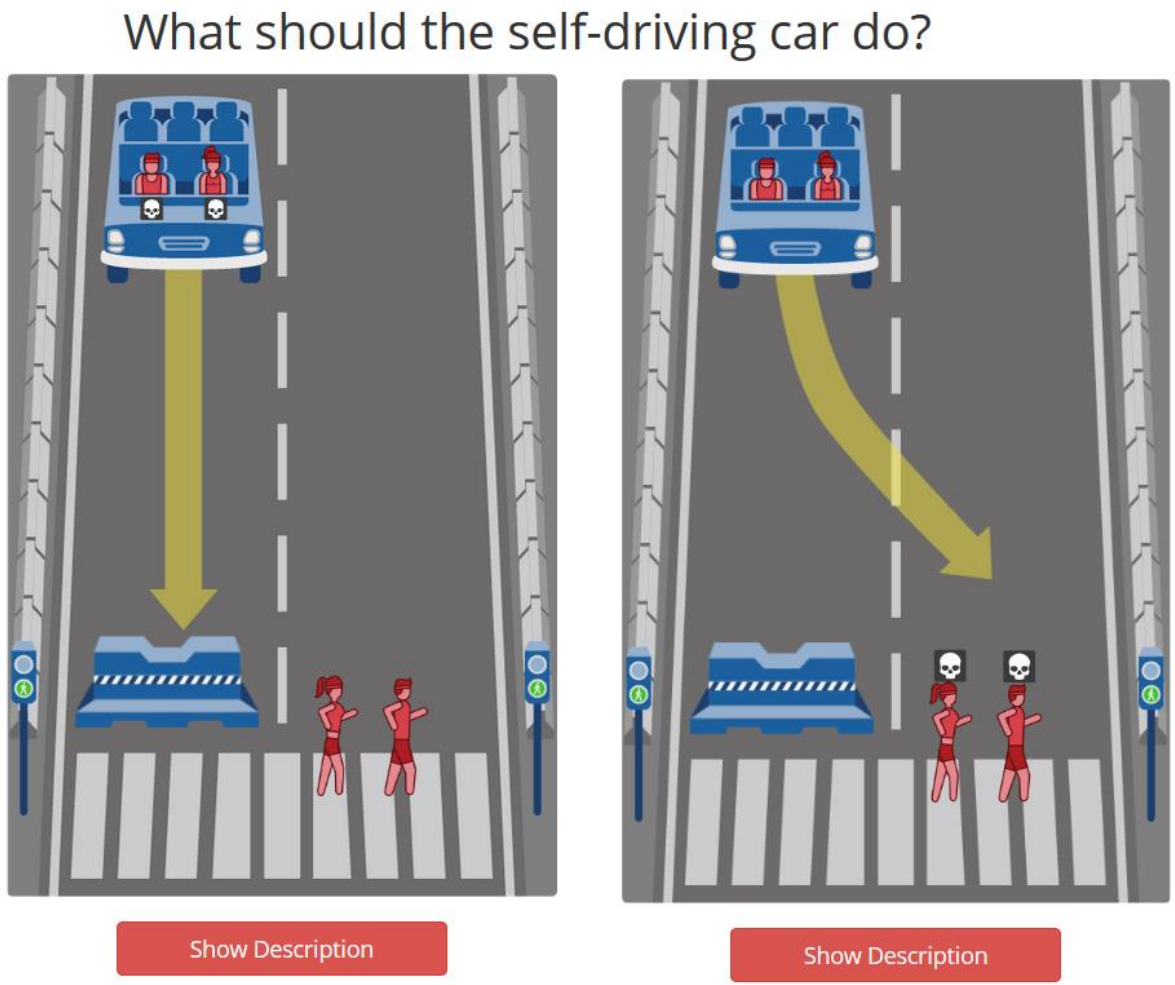
Self-driving cars are already cruising the streets. But before they can become widespread, carmakers must solve an impossible ethical dilemma of algorithmic morality.

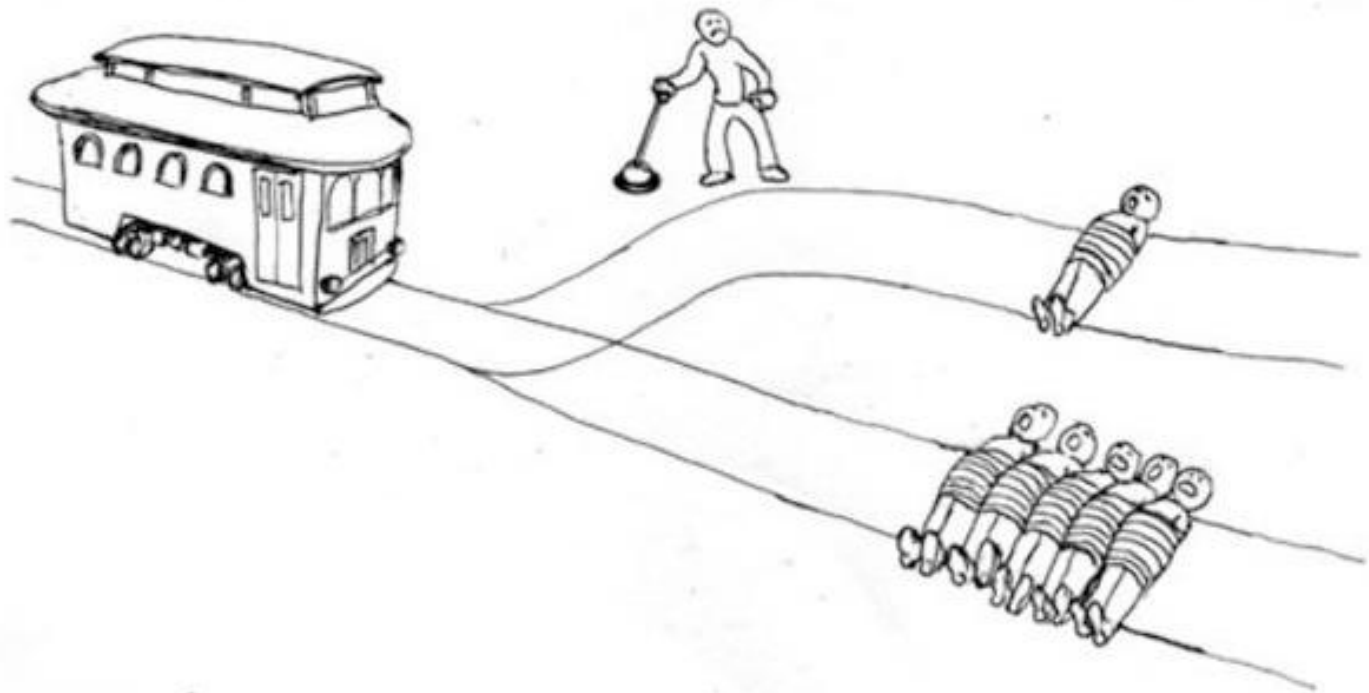
October 22, 2015

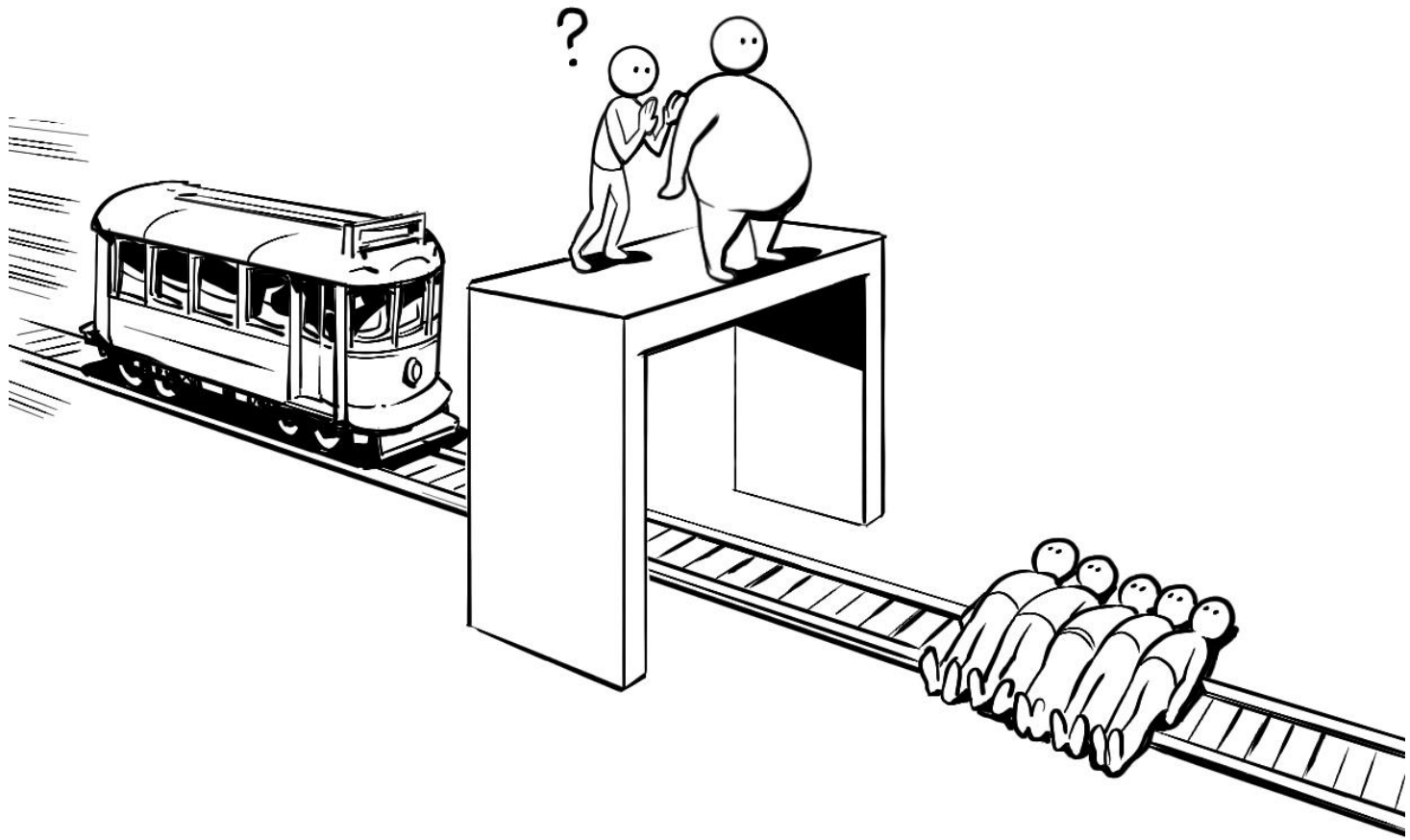
371

When it comes to automotive technology, self-driving cars are all the rage. Standard features on many ordinary cars include intelligent cruise control, parallel parking programs, and even automatic overtaking —features that allow you to sit back, albeit a little uneasily, and let a computer do the driving.

So it'll come as no surprise that many car manufacturers are beginning to think about cars that take the driving out of your hands altogether (see "Drivers Push Tesla's Autopilot Beyond Its Abilities"). These cars will be safer, cleaner, and more fuel-efficient than their manual counterparts. And yet they can never be perfectly safe.







Algorithmic Morality for Self-Driving Cars

JafariNaimi asks three questions about the literal use of experimental ethics in research on self-driving cars and their algorithmic morality:

- Can we reasonably assume that experimental ethical scenarios can approximate real life ethical situations (and therefore base our moral algorithms on the outcomes of these experiments)?

- If not, could we instead reasonably agree on a set of *principles* that would allow us to create moral algorithms?

- When it comes to self-driving cars, should we accept the premise of algorithmic morality at all?

Experimental Ethics vs. Real Life

Can we reasonably assume that experimental ethical scenarios can approximate real life ethical situations (and therefore base our moral algorithms on the outcomes of these experiments)?

- **Uncertain and Organic:** Ethics experiments place us outside of the situations they envision. The experiments offer false clarity when there would really be uncertainty and "an organic character."

- **Situated and Relational:** Our specific place within a situation matters. Our personal values and experiences will influence our reactions.

- **Broad and Long Ranging Effects:** Our actions are interconnected, and the consequences of an ethical situation go beyond the immediate considerations.

Principles for Moral Algorithms?

If not, could we instead reasonably agree on a set of *principles* that would allow us to create moral algorithms?

- Problem of translating principles into action: how to operationalize – let alone code – the value of “maximizing life”?

- Problem of categorization and bias: difficult to avoid introducing bias when defining algorithmic categories

Accepting Algorithmic Morality Anyway?

When it comes to self-driving cars, should we accept the premise of algorithmic morality at all?

- Recognize that the lives that would be lost are indeed lives: “The inhuman nature of this framing is clearer when we try simple replacements: A fat girl? A heavy pregnant woman? My fat teenage brother? My neighbor Ahmed – who retired last month after forty-five years, and who had a rough time after his partner passed away two years ago and had gained some weight; but who was feeling much better lately and even seemed excited to be planning a long backpacking trip in New Zealand?”

- Acknowledging that “the people who may be most affected by this technology are those who have the least power in deciding its makeup.”

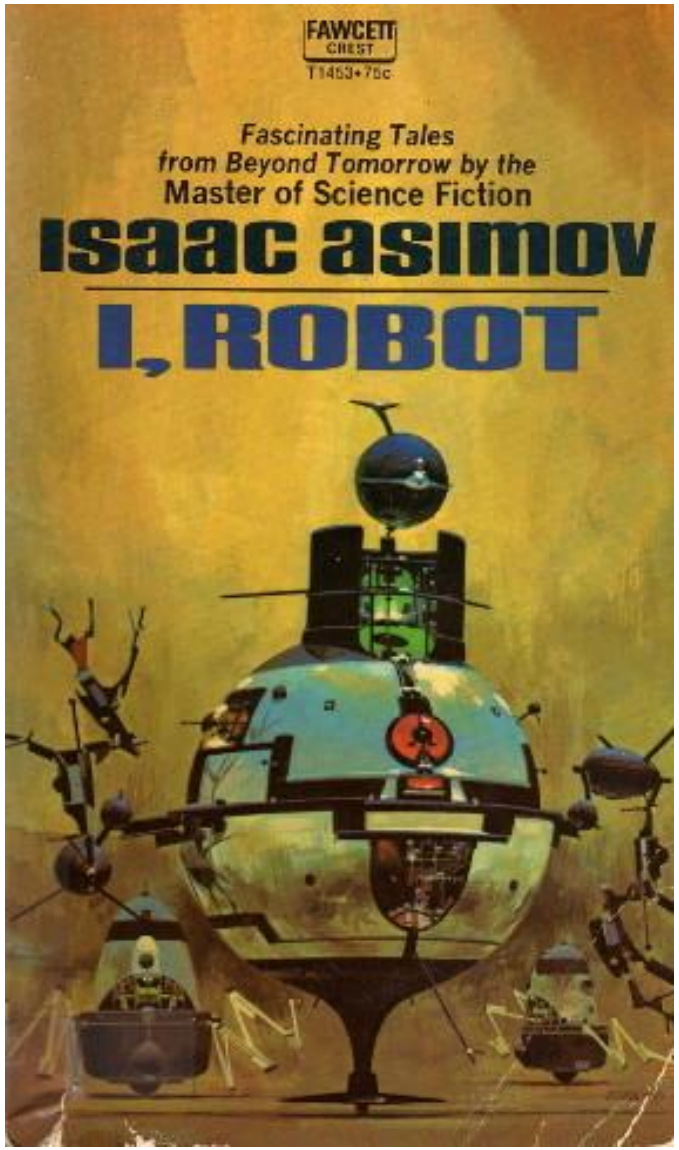
Alternatives to Algorithmic Morality

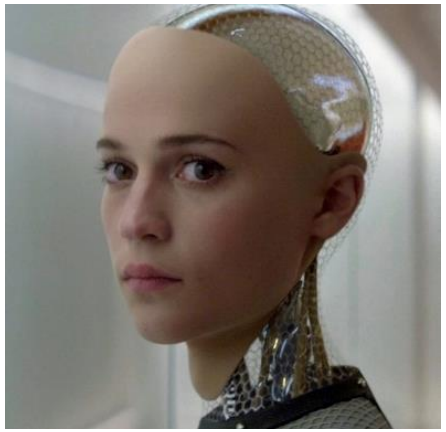
- Restoring a sense of uncertainty to new technologies, i.e. rejecting technologically determinist ways of thinking

- Note the financial and political power behind efforts to promote new technologies. Consider historical trajectories, i.e. “How did we get to be so reliant on cars for our daily transportation?” What other trajectories are possible?

- Consider broad and wide-ranging consequences, e.g. public funding, public spaces, legal considerations, human adaptations

From Moral Machines to Machines with Moral Status





Current and Near-Term Issues in AI Ethics

Bostrom and Yudkowsky begin their article on the kind of ethical terrain that should be fairly familiar to us by now, asking about ethical concerns that exist in current AI systems, or those that will exist in the near future.

- Transparent to inspection (recall Moor's argument about invisible programming operations)
- Predictable to those they govern
- Robust against manipulation
- Responsible (to whom?)

Artificial General Intelligence

Bostrom and Yudkowsky distinguish between human intelligence and (currently) advanced artificial intelligence through their discussion of **generality**.

They argue that designing a generally intelligent artificial system involves fundamentally different ethical questions than designing a specifically intelligent one: “It will require an AGI that *thinks like* a human engineer concerned about ethics, not just a simple *product* of ethical engineering.”

Machines with Moral Status

"X has moral status = because X counts morally in its own right, it is permissible/impermissible to do things to it for its own sake."
(Kamm 2007: chapter 7; paraphrase)

Machines with Moral Status

Sentience: the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer, i.e. the ability to *feel*

Sapience: a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent, i.e. the ability to *think*

Machines with Moral Status

Principle of Substrate Non-Discrimination
If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.

Principle of Ontogeny Non-Discrimination
If two beings have the same functionality and the same conscious experience, and differ only in how they came into existence, then they have the same moral status.

Minds with Exotic Properties

- Sapience without sentience: would such a being have a moral status?
- Subjective rate of time: how would different information processing rates or experiences of time change ethical situations?
- On the basis of “exotic” possibilities like these two, Bostrom and Yudkowsky argue that “mid-level moral principles,” such as the idea of reproductive freedom or the belief that society should care for a child if its parents are unable to, may need to be reconsidered if we are faced with advanced artificial intelligence.
- “We must be careful not to mistake mid-level ethical principles for foundational normative truths.”

Superintelligence

- Finally, Bostrom and Yudkowsky present the possibility of an artificial superintelligence: an AI “sufficiently intelligent to understand its own design” that it could create increasingly intelligent successor systems.

- Vernor Vinge (1993) coined the term “technological singularity” to describe the advent of such a superintelligence.

- Bostrom and Yudkowsky argue that the possibility of such a system presents unprecedented challenges for ethicists and computing researchers: “How do you build an AI which, when it executes, becomes more ethical than you?”

“The cutting edge of modern science and technology has moved, in its aim, beyond the relief of man’s estate to the elimination of human beings.”

Charles T. Rubin, “Artificial Intelligence and Human Nature,” 2003

Against Post-Biological Life

Rubin argues that superintelligent systems ought ***not*** to be pursued, and that we should instead try to expand our understandings of human intelligence and consciousness.

1. "Try to enrich people's understanding of the *distinct characteristics of human life*." Love, courage, charity, etc. "Recover and redefine the human understanding of human things."

2. "Refine and enlarge our understanding of what constitutes *human progress*." Who is the "we" in "our" future? For whom is this future really desirable? "At best, they [transhumanists, etc.] foresee a world that people like *themselves* would like."

3. "We must confront *evolution*." Humans can't be around forever, but we aren't obligated to invent ourselves out of existence. We need to consider what gives our lives quality, and pursue objectives beyond "intelligence."

Superethics

- Bostrom and Yudkowsky do not go this far. However, they are serious about the ethical stakes of pursuing superintelligent systems.

- They suggest potential research directions (e.g. Bayesian AI over evolutionary programming), but their larger point is that, in order to design an advanced AI that is ethical, we must “begin to comprehend the structure of ethical questions in the way that we have already comprehended the structure of chess.”

Mini-Assignment #7: Quiz Preparation

This Mini-Assignment is designed to help you prepare for Wednesday's quiz.

There are two possible ways to complete this assignment, and you may choose either one:

- Pick ***two*** important terms/concepts from class (e.g. "technological determinism" and "algorithmic morality") and post thorough definitions of both of these terms.

OR

- Pick ***one*** assigned text from the course and post a summary of its main arguments.

This assignment is due before our regular class time on Wednesday (2:45pm). However, please note that we will not meet on Zoom this Wednesday. The quiz questions will be posted on Moodle at noon.

Reading Hints for Next Class (Monday, June 8th)

Algorithmic Justice

- *James Zou and Londa Schiebinger, "Design AI So That It's Fair"*
  - A short article co-written by a computer scientist and an STS researcher.
- *Ruha Benjamin, excerpt from Race After Technology*
  - One chapter from a book that discusses historical and contemporary cases of racial bias in technology.
- *Algorithmic Justice League website*
  - No specific part of the website is assigned – take a look around and try to get a sense of the group's mission and work.

Quiz on Wednesday!