Teaching Ethics in AI

Benjamin Kuipers
Computer Science & Engineering
University of Michigan

Why Should We Teach Ethics in AI Classes?

- We are likely to have more robots (and other AIs) acting as members of our society.
  - Self-driving cars and trucks on our roads and highways.
  - Companions and helpers for the elderly.
  - Teachers and care-takers for children.
  - Managers for complex distributed systems.
- We need to ensure that they will behave well.
  - What does this mean? How can we do this?
  - How can we trust them?

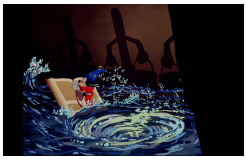
What Can a Robot Do?

- Humans provide the robot's top-level goals.
  - At least, at the current state of the art.
- The robot:
  - Perceives and acts in the world;
  - Builds its own model of the world;
  - Creates a plan to reach its goal in the world;
  - Puts that plan into action, *including creating its own subgoals*, as part of the plan it is pursuing.
- We do **not** assume that the robot is a moral agent, capable of taking responsibility for its actions.
  - But it should still know how to behave well.

We worry about robot behavior.

With no sense of what's appropriate, and what's not, they may do great harm.



"What about me, Frank?"

- Robot & Frank (2012)
  - https://youtu.be/eQxUW4B622E

"You're starting to grow on me."

- Robot & Frank (2012)
  - https://youtu.be/xlpeRIG18TA

"You lied?"

- Robot & Frank (2012)
  - https://youtu.be/3yXwPfvvIt4

Lessons

- Robot has no moral or legal inhibition from stealing, shoplifting, or robbery.
  - "*I took it for you. Did I do something wrong, Frank*?"
  - "*I don't have any thoughts on that* [stealing]."
- Robot has no inhibition against lying.
  - "*I only said that, to coerce you*."
  - "*Your health supersedes my other directives*."
- Robot has no concern for self-preservation.
  - "*The truth is, I don't care if my memory is erased or not*."

This does not end well.

- But what's the problem?
  - It's not the robot apocalypse.
- It's robots pursuing human-provided goals in unconstrained ways, generating and pursuing unexpected subgoals.
- We must design robots to be *trustworthy*.
  - How do we do that?
  - What can we learn from the philosophy of ethics?
  - What can we learn from the state-of-the-art in AI?

Theories of Philosophical Ethics

- We draw on concepts that philosophers and prophets have been teaching and developing for centuries.
  - **Utilitarianism** ("*What action maximizes* utility *for all?*")
    - Special case of **consequentialism** ("*What action has the best* consequences *for all?*")
  - **Deontology** ("*What is my* duty, *to do, or not to do?*")
  - **Virtue ethics** ("*What would a* virtuous *person do?*")
- Instead of treating these as mutually exclusive, we see them as parts of a single complex reality.
  - "*The Blind Men and the Elephant*"
  - "*Climbing the same mountain on different sides*"

Performance Requirements

- The physical and social environment has unbounded complexity.
  - Selecting an abstraction is an essential step.
- Many situations need an immediate response.
  - Real-time response leaves no time for deliberation.
- Later, careful deliberation is necessary.
  - We learn from good and bad experience.
  - We learn from explanations: our own and others'.
  - Incremental improvement toward *practical wisdom*.

An AI Perspective on Ethical Theories

- The different ethical theories suggest different AI knowledge representations, able to express different kinds of ethical knowledge.
  - **Utilitarianism** (*Decision theory / Game theory*)
    - Good for continuous optimization, but not in real time.
    - Sensitive to choice of utility measure.
  - **Deontology** (*Pattern-matched rules and constraints*)
    - Good for explanation and computational efficiency.
    - Depends on the terms that can appear in patterns.
  - **Virtue Ethics** (*Case-based reasoning*)
    - Good for expressive power in complex domains.
    - Good for incremental learning from experience.
- Using multiple models together is more robust.

Decision Theory and Game Theory

- The standard approach to decision making in AI [Russell & Norvig, 3e, 2010] defines **Rationality** as choosing actions to *maximize expected utility*.

$$action = \arg\max_a EU(a|\mathbf{e})$$

where

$$EU(a|\mathbf{e}) = \sum_{s'} P(\text{RESULT}(a) = s'|a, \mathbf{e})U(s')$$

- **Utility** $U(s)$ represents the individual agent's preference over states of the world.
- *Game theory* is decision theory in a context with other decision-making agents.

The Crux is Defining Utility

- **Utility** $U(s)$ represents the individual agent's preference over states of the world.
  - Utility need not be self-centered.
  - An individual's utility can reflect *everyone's* welfare, or some other sophisticated property.
  - Unfortunately, that's often hard to implement.
- Utility is often defined selfishly – in terms of the agent's own reward.
  - Appropriate in entertainment games.
  - In society, maximization of self-centered reward often leads to bad outcomes, individually and collectively.
  - Prisoner's Dilemma, Tragedy of the Commons, . . .

The Prisoner's Dilemma

- Two prisoners are separated, and offered:
  - If you testify and your partner doesn't, you go free and your partner gets 5 years in prison.
  - If you both testify, you both get 3 years.
  - If neither testifies, you both get 1 year.

	Testify	**Don't**
Testify	(–3, –3)	(0, –5)
Don't	(–5, 0)	(–1, –1)

Utility is years in prison.

The Prisoner's Dilemma

	Testify	**Don't**
Testify	(–3, –3)	(0, –5)
Don't	(–5, 0)	(–1, –1)

Utility is years in prison.

- Whatever your partner does, **Testify** is your best choice. Same for your partner.
- (**Testify, Testify**) is a *Nash equilibrium*:
  - Any individual changing from this choice reduces his own utility.
- You both get 3 years: the *worst* collective outcome.
  - Society does badly. You do badly, too.

The Prisoner's Dilemma

	Testify	**Don't**
Testify	(–3, –3)	(0, –5)
Don't	(–5, 0)	(–1, –1)

Utility is years in prison.

- The cooperative outcome, (**Don't, Don't**), is better for individuals and for society, but requires trust.
- (**Don't, Don't**) is *not* a Nash equilibrium:
  - Your partner can unilaterally improve his own utility by "rationally" violating your trust.
  - If you both reason the same way ("rationally"), even his benefit from violating your trust is lost.
- Selfish utility maximization violates trust.

The Prisoner's Dilemma Scales Up

(The Public Goods Game)

- N players contribute money to a common pool.
  - The pool is multiplied (× 2 or 3) and the result is distributed evenly among all players.
- Best for society (Cooperation):
  - Everyone contributes their maximum, to get the most benefit from the multiplication.
- Best for individual (Selfish exploitation):
  - Contribute nothing. Share in the benefit.
- Nash equilibrium:
  - Nobody contributes. Nobody benefits.
  - Even the free rider's benefit collapses.

The Prisoner's Dilemma Is Realistic

(The Tragedy of the Commons)

[Garret Hardin, 1968]

- I can graze my sheep on the Commons, or on my own land.
  - Personally, I'm better off grazing as many of my sheep as I can on the Commons, saving my own land.
  - Likewise everyone else.
- So we all overgraze the Commons, and it dies.
  - Then we have only our own land, and no Commons.
  - We're all worse off!
- Modern, real-world Commons:
  - Clean air and water, fishing, climate change, . . .

What Have We Learned?

- **If** we select actions to maximize expected utility
  - (as in decision theory and game theory)
- **and if** we define utility in terms of own reward
  - (which is common)
- **then** the resulting Nash equilibrium is likely to be a bad outcome, individually and collectively.
  - Selfish utility-maximization:
    - Exploits vulnerability,
    - Violates trust, and
    - Discourages cooperation.

(There may be better utility measures, but tractability is a problem.)

Conclusions for Robot Builders

- Robots (and other AIs) can behave badly, even while they pursue human-given top-level goals.
  - Even without worrying about the robot apocalypse.
- Ethical theories from philosophy suggest useful clues, knowledge representations, and methods.
  - Hybrid structures at multiple time-scales are needed.
- Maximization of selfish utility is too simple.
  - The Nash equilibrium is often a poor outcome, both for the individual and for society.
- AI students need to understand these things.

References

- Robert Axelrod. *The Evolution of Cooperation*, 1984.
- Joshua Greene. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. 2013.
- Jonathan Haidt. *The Righteous Mind*. 2012.
- Benjamin Kuipers. Toward morality and ethics for robots. *AAAI Spring Symposium on Ethical and Moral Considerations in Non-Human Agents*, 2016.
- Kevin Leyton-Brown & Yoav Shoham. *Essentials of Game Theory*, 2008.
- Patrick Lin, Keith Abney & George A. Bekey. *Robot Ethics: The Ethical and Social Implications of Robotics*, 2012.
- Stuart Russell & Peter Norvig. *Artificial Intelligence: A Modern Approach*, 3e, 2010.
- Shannon Vallor. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. 2016.
- Wendell Wallach & Collin Allen. *Moral Machines: Teaching Robots Right from Wrong*. 2009.