Artificial Intelligence and Ethics

Todd W. Neller
Gettysburg College
Department of Computer Science

Acknowledgements

- This talk is motivated by and draws material from:
  - E. Burton, J. Goldsmith, S. Koenig, B. Kuipers, N. Mattei, and T. Walsh. “Ethical Considerations in Artificial Intelligence Courses” to appear in AI Magazine
  - B. Kuipers. “Teaching Ethics in AI” presentation at the Educational Advances in Artificial Intelligence 2017 Symposium (EAAI-17)

Outline

- Motivations – Why care?
- Ethical Systems
  - How to we frame discussion?
  - How do ethical systems relate to one another?
- Case Study: Robot and Frank movie (2012)
- Difficulties of Utilitarianism
- Science Fiction Utopian versus Dystopian Visions
- Implications for You

The Sorcerer's Apprentice

- 1797 peom by Goethe
- Retold and popularized in the Disney movie "Fantasia" (1940)
- Cautionary tale of using powers beyond one's capacity to control.

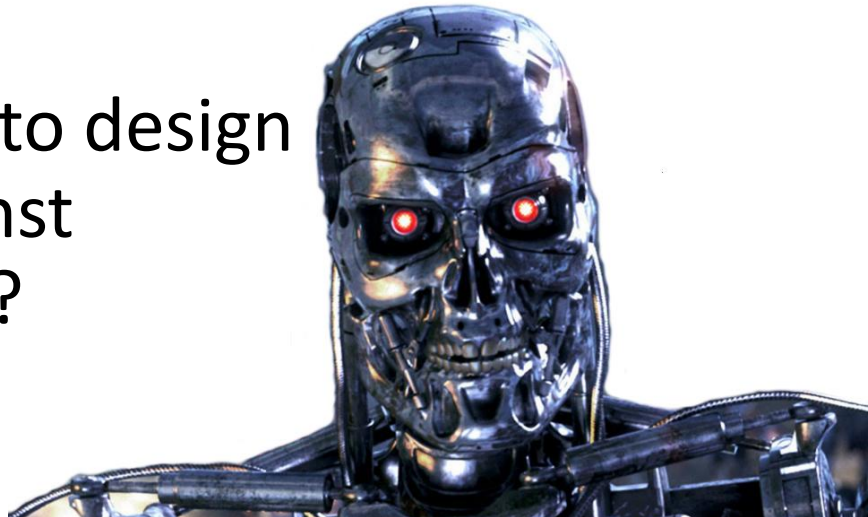




http://film-instant.tumblr.com/post/7656060802
http://giphy.com/search/sorcerers-apprentice
http://video.disney.com/watch/sorcerer-s-apprentice-fantasia-4ea9ebc01a74ea59a5867853

Concerns

- Arthur C. Clark's third law: "Any sufficiently advanced technology is indistinguishable from magic." Magic as metaphor for technology.
- Like the Sorcerer's Apprentice, is it possible we will we lose control of our tools of technology?
- We sometime do because we *can*, but do not ask whether we *should*.
- Will we have the foresight to design safeguards to protect against unintended consequences?



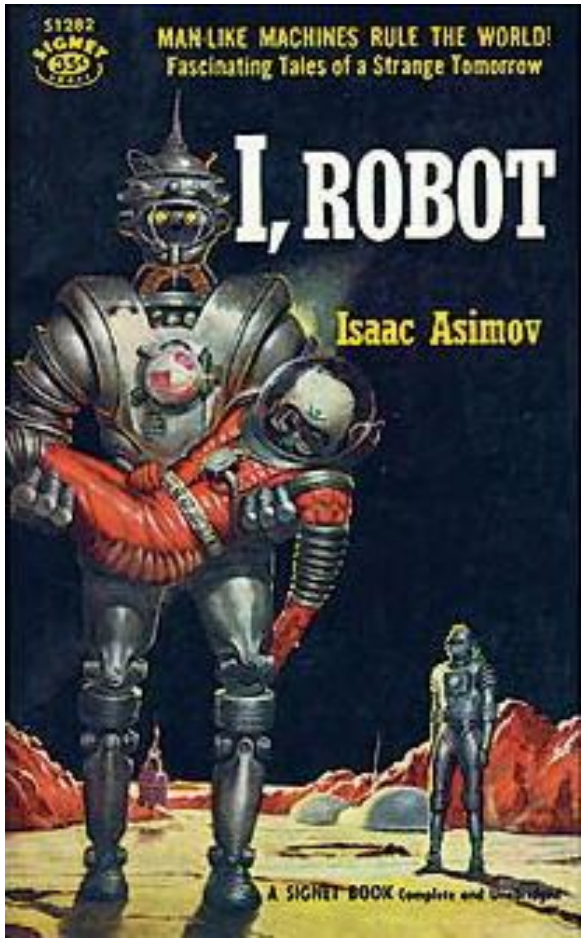
Deontology

- Deontology: duty-/obligation-/rule-based ethics judges the morality of an action based on **rules**
- Examples:
  - Ten Commandments
  - the law
  - a professional code

No other gods before me
No graven images or likenesses
Not take the LORD's name in vain
Remember the sabbath day
Honour thy father and thy mother
Thou shalt not kill
Thou shalt not commit adultery
Thou shalt not steal
Thou shalt not bear false witness
Thou shalt not covet

Asimov's Three Laws

- Isaac Asimov's Three Laws of Robotics ("I, Robot", 1950):
  1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
  2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
  3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.



Fair use, https://en.wikipedia.org/w/index.php?curid=1815661

Utilitarianism

- Utilitarianism: Act so as to **maximize expected utility**, where utility is a measure of "goodness".
- Presumes that we can *quantify* goodness.
- Places a moral burden of expressing what one values *quantitatively*.
- Example: Company policy
  - Monetary gain value? Easy? Already a number.
  - Environmental responsibility? Hmm. Very high value for legal compliance. Carbon footprint times *what*?

Does Utilitarianism Subsume Deontology?

- Example: Let L_n denote Asimov's law *n* and U_n be the utility of keeping Asimov's law *n*. If
  - U_3 is greater than the maximum sum of all other non-law utilities,
  - U_2 is greater than the maximum sum of U_3 and all other non-law utilities, and
  - U_1 is greater than the maximum sum of U_2, U_3, and all other non-law utilities,
  - Does this utility function express Asimov's three laws?

∞ Shades of Gray

- Is “harm” clear and well-defined?
- What if one action manages to keep law 1 with .001 probability, breaking all other laws, but another action breaks law 1 with certainty, but keeps other(s) definitely (probability 1.0)?
  - “There’s a slim chance I can autonomously drive this car to save my passenger and all of those pedestrians, but my passenger has given an override command to ‘Steer right!’ and crash into a wall, saving the crowd, but deliberately sacrificing himself/herself. (In my black box, I’ll survive.)”
- Does utilitarianism express that “the end justifies the means?”

Utility of Money

- Quantifying value of money seems easy, but…
- St. Petersburg Paradox: Flip a coin until it comes up tails on turn *t.* Payout: 2^t dollars.
- How much would you be willing to pay to play?
  $\frac{1}{2}\cdot 2 + \frac{1}{4}\cdot 4 + \ldots = 1 + 1 + \ldots. = \infty$
- Problem: The utility of money isn't linear.

Virtue Ethics

- Virtue Ethics: Act so as to acquire/cultivate/evidence virtues, that is, good personal characteristics, e.g.:
  - Fruits of the Spirit: love, joy, peace, patience, kindness, goodness, faithfulness, gentleness, self-control.
  - Nicomachean ethics moral virtues: courage, temperance, liberality, magnificence, magnanimity, proper ambition, truthfulness, wittiness, friendliness, modesty, righteous indignation
- Criticism: Virtue ethics is culturally relative ("Whose virtues?"), but then this applies to all of the above ("Whose laws?", "Whose utility measure?")

Does Utilitarianism Subsume Virtue Ethics?

- Let every action that could be classified as evidencing a virtue score a “virtue point” for that virtue. (An action could score multiple virtue points, e.g. a courageous honest answer.)
- Then if utility is measured in virtue points, then does maximizing expected utility express virtue ethics?

Much Can Be Swept Under the Utility Rug

- Utilitarianism in its most general form can express many ethical systems.
- Some believe that utilitarianism is a version of consequentialism ("the end justifies the means"). This is incorrect:
  - One can define a utility function that values **states** of the world and/or **actions** that cause us to transition from one state to another.
  - Defining utility w.r.t. **states only** is consequentialism.
  - Defining utility w.r.t. **actions only** allows us to subsume both deontology (actions that fulfill duty, follow rules/laws) and virtue ethics (actions that acquire/develop/express virtues)
  - And there's an entire *hybrid* utility function design space that values **both states and actions** that invites further exploration.

Expressing Goodness

- Whatever the ethical system, expressing goodness is difficult… and worthy of our best efforts.
- Sometimes, when a task is difficult, we have a strong incentive to avoid/neglect it.
- Consider the movie “Robot and Frank” (2012) in which a robot is designed with only Frank’s health care as having utility…

“What about me, Frank?”

[Robot & Frank are walking in the woods.]

F: [panting] I hate hikes. God damn bugs! You see one tree, you’ve seen `em all. Just hate hikes.

R: Well, my program's goal is to improve your health. I'm able to adapt my methods. Would you prefer another form of moderate exercise?

F: I would rather die eating cheeseburgers than live off steamed cauliflower!

R: What about me, Frank?

F: What do you mean, what about you?

R: If you die eating cheeseburgers, what do you think happens to me? I'll have failed. They'll send me back to the warehouse and wipe my memory. [Turns and walks on.]
https://youtu.be/eQxUW4B622E

“You’re starting to grow on me.”

(scene 1)

[Robot & Frank are walking through a small knick-knack shop in the town. As he walks by a shelf, Frank slips a small sculpture into his pocket.]

Young woman surprises him: Have you smelled our lavender heart soaps?

[Frank smells a soap]

R: We should be going, Frank.

Young woman: Oh, what a cute little helper you have!

Older woman marches up, frowning: What is in your pocket?

[Frank leans over, cupping his ear]

F: I'm sorry, young lady, I couldn't quite hear you. [While talking, slips the sculpture out of his pocket, back onto the shelf.]

Older woman: What is in your pocket? I'm going to make a citizen's arrest.

F [turning out his pockets]: Nothing. Nothing's in my pockets. Look!

R: Frank! It's time we head home.

F: Yeah. Yeah. If you'll excuse us, ladies. It's nice to see you.

[Robot & Frank walk out.]

Young woman: Have a good one.

“You’re starting to grow on me.”

(scene 2)

[R+F are walking through the woods. Frank looks in the bag and finds the sculpture.]

F: Hey! Hey! Where did this come from?

R: From the store. Remember?

F: Yeah, yeah. Of course I remember. But I mean what did you do? Did you put this in here? You took this?

R: I saw you had it. But the shopkeeper distracted you, and you forgot it. I took it for you. [pause] Did I do something wrong, Frank?

[Frank puts it back into the bag, and they walk on.]

“You’re starting to grow on me.”

(scene 3)

[At home, Frank is sitting at the table, holding the sculpture.]

F: Do you know what stealing is?

R: The act of a person who steals. Taking property without permission or right.

F: Yeah, yeah, I gotcha. [pause] [addresses Robot directly] You stole this. [long pause, with no response from Robot] How do you feel about that?

R: I don't have any thoughts on that.

F: They didn't program you about stealing, shoplifting, robbery?

R: I have working definitions for those terms. I don't understand. Do you want something for dessert?

F: Do you have any programming that makes you obey the law?

R: Do you want me to incorporate state and federal law directly into my programming?

F: No, no, no, no! Leave it as it is. You're starting to grow on me.

https://youtu.be/xlpeRIG18TA

“You lied?”

[Frank sitting in the woods, Robot standing next to him. They are in mid-conversation.]

R: All of those things are in service of my main program.

F: But what about when you said that I had to eat healthy, because you didn't want your memory erased? You know, I think there's something more going on in that noggin of yours.

R: I only said that to coerce you.

F: [shocked] You lied?

R: Your health supersedes my other directives. The truth is, I don't care if my memory is erased or not.

F: [pause] But how can you not care about something like that?

R: Think about it this way. You know that you're alive. You think, therefore you are.

F: No. That's philosophy.

R: In a similar way, I know that I'm not alive. I'm a robot.

F: I don't want to talk about how you don't exist. It's making me uncomfortable.
https://youtu.be/3yXwPfvvIt4

Robot and Frank Questions

- If an elderly person wishes to behave in ways that violate common social norms, should a caretaker robot intervene, and if so, how?
- If an elderly person wants to walk around the house despite some risk of falling, should the robot prevent it?
- How could a caretaker robot judge whether or not to follow the instructions of the human in its care?
- Should a robot enforce limits on patient autonomy?
- How should a robot balance privacy versus reporting concerns? Or prioritize loyalty to the one in its care versus loyalty to medical staff versus loyalty to close family?
- How would a robot resolve conflict between two different types of “well-being” (e.g. emotional vs. physical)?

Quantifying Hard-To-Quantify Values

- From S. Russell and P. Norvig. "Artificial Intelligence: a modern approach" (3rd ed., 2010, p. 615):

  "… *Although nobody feels comfortable with putting a value on human life, it is a fact that tradeoffs are made all the time.* Aircraft are given a complete overhaul at intervals determined by trips and miles flown, rather than after every trip. Cars are manufactured in a way that trades off costs against accident survival rates. **Paradoxically, a refusal to "put a monetary value on life" means that life is often undervalued.** Ross Shachter relates an experience with a government agency that commissioned a study on removing asbestos from schools. The decision analysts performing the study assumed a particular dollar value for the life of a school-age child, and argued that the rational choice under that assumption was to remove the asbestos. The agency, morally outraged at the idea of setting the value of a life, rejected the report out of hand. It then decided against asbestos removal—implicitly asserting a lower value for the life of a child than that assigned by the analysts." (boldface mine)

A Thought Experiment: AI CEO

- Imagine an AI CEO with a utility function based purely on **maximizing expected profit**. Further imagine superhuman knowledge, reasoning, and access to information.
- What would be the implications for:
  - Employee management, hiring, firing?
  - Environmental stewardship?
  - Federal, state, and local law?
  - Other important values that are hard to quantify?

The Roads to Sci-Fi
Utopia and Dystopia

- I would argue that there is a central feature that separates utopic and dystopic sci-fi visions of the future: *whether or not the utility function is designed well.*
- Dystopia: AI Rebellion and Domination
  - Utility based on AI self-preservation, human domination (e.g. R.U.R., Terminator series, The Matrix, Ultron, etc.)
- Utopia: AI Benevolent Shepherds, Companions
  - Utility based on human-preservation, service to society (e.g. I, Robot, Star Wars good droids, Star Trek's Data, Interstellar's TARS and CASE)

Asimov's Zeroth Law

- Isaac Asimov later added the "Zeroth Law" to supersede his original Three Laws of Robotics:
  - A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
- In "Foundation and Earth", Asimov expresses the difficulty of its realization:

  Trevize frowned. "How do you decide what is injurious, or not injurious, to humanity as a whole?"

  "Precisely, sir," said Daneel. "In theory, the Zeroth Law was the answer to our problems. In practice, we could never decide. A human being is a concrete object. Injury to a person can be estimated and judged. Humanity is an abstraction."

My Charge To You

1. Always critically question the utility function.
2. Have the courage to quantify hard-to-quantify values. Imperfect valuation is better than no valuation.
3. Imagine and work towards a better world, but design against the worst case.

Questions?