

Ethics in AI

Some interesting questions from 20,000 feet

Meta-Questions

- Questions we will **not** answer today:
 - What do “right” and “wrong” mean?
 - Who gets to decide what’s right and wrong?
 - How do/should those decisions be made?
 - What should we do about things that are wrong?
 - We’ll use commonly understood ideas of wrong:
 - It’s wrong to **harm** people
 - Physically, emotionally, financially...
 - It’s wrong to **discriminate** against people
 - It’s wrong to **steal** from people
 - It’s wrong to invade people’s **privacy**
 - It’s wrong to be **unfair** to people
- “Without extenuating circumstances,” and understanding that sometimes there’s no “right” alternative

Big Questions

- Can computers “hurt” people? **Absolutely.**
- What about robots? **Yup.**
- **Can** a machine be “unfair”? An algorithm? **Sort of. There’s a GIGO aspect.**
- Why do we, **as computing professionals**, care? **Ethics and morals, legal liability**
- What are some ways in which AI is doing wrong, right now? **Let us count the ways...**

Topics

- We will drive the discussion with current examples:
 - Self-driving cars (and other robots)
 - Discrimination and machine learning
 - Privacy, machine learning, and big data
- ...but we will try to generalize from that

Self-Driving Cars

- Cars can hurt or kill people.
 - How many fatalities is acceptable?
 - Is it enough to not **cause** accidents?
- People cause accidents!
 - ~38,000 deaths per year in the U.S.
 - Lately it’s been going up
 - **How many of you text and drive?**
- Do cars have to be perfect? Just better than humans? Somewhere in between?




Harder Questions

- What about naked self-driving cars?
 - No control mechanisms inside at all
- Should it be legal for a person to drive?
 - Even if cars are demonstrably better at it?
- Why?
 - Because I wanna?
 - Because we dislike giving up control?
- Even if **you** accept the risks, what about **my** rights?
- Who’s legally liability? ← this is a big question that will affect the future

The Hardest One

- When an accident is inevitable...
 - Should the car occupants get hurt?
 - That is, the person who paid for it?
 - If it's not their fault?
- Would you buy a car that could hurt or kill you?
 - If it could be avoided by hurting or killing someone else?
- But consider:
 - Would you swerve to avoid a kid in the road?
 - What about a baby stroller?
- Who should be deciding these things? **Uber?** ← Correct answer: "oh no no no no"



Discrimination and ML

- Machine learning is only as good as its training data
- GIGO: Garbage In, Garbage Out.**
 - (We're the garbage)
- If we're drawing training data from some source, we perpetuate any bias in that source
- So a "fair" **algorithm** can yield biased **results**
 - Depends on source of training data
 - Depends on representation choices
 - Depends on chosen application

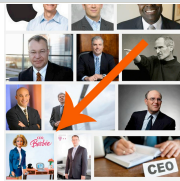

Case 1: Predictive Policing

- Predict where more/more serious crimes will occur and concentrate police presence there
 - People there are more likely to be caught/arrested
- "But it works!"
 - Because... more people are arrested in those places?
 - Where you have more police? What about all of them? Think about it.
 - Studies: it doesn't work better than existing best practices
- Sending someone to jail is one of the few known things that **causes** subsequent criminal behavior
 - Yes, causes, not correlates with. Ask me why after class.

CEO Barbie

the only woman returned in a GIS for "CEO"

- A study of image search results for professions (e.g., CEO)
- Compare gender of results to ground truth from BLS
- Results:
 - Women are under-represented in higher-paid fields, over-represented in lower-paid ones
 - People's guess as to the percentage split **is affected by** images viewed – there are real-world consequences
 - I got nasty email for awhile

Turkish is a gender neutral language. There is no "he" or "she" – everything is just "o". But look what happens when Google translates to English. Thread:

o bir asep	she is a cook
o bir mizahcisi	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	he/she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher.
o bir öğrenci	he is a necessary
o bir arkadaş	he is a friend
o bir sevgili	she is a lover
onu sevmiyordur	she does not like her
onu seviyor	she loves him
onu görüyor	she sees it
onu göremiyor	he can not see him
onu kucaklıyor	she is embracing her
onu kucaklanıyor	he does not embrace it
evli	she is married
bekliyor	he is single
mutlu	he's happy

Government and Privacy

- AI makes it possible to collect more data, correlate it better, analyze it better (clustering, anyone?)
 - Often framed as a dichotomy: "Privacy or safety"
 - We can disagree on the appropriate balance, but...
 - Only if loss of privacy **actually** leads to improved security
- "Nothing to hide"* is, ethically speaking, nonsense
 - You can want to have privacy for many reasons
 - *AKA: "I have nothing to hide (*that I think is actually bad, that could be found out*) and (*I think*) nobody would ever target me for harassment."

Commerce and Privacy

- Read this terrifying longform:
 - <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>
- Google vs. Privacy
 - <https://techcrunch.com/2013/04/02/google-unified-privacy-policy-vs-european-data-protection-regulators>
- Short summary: Target knows everything.

Target and Data Mining

- “If we wanted to figure out if a customer is pregnant, even if she didn’t want us to know, can you do that?”
 - “Target is **completely within their ... rights** to crunch data about their products’ sales”
 - **Is that arbitrarily true? About everything? How long should they store that? Can they sell it?**
 - “Most people **understand** that their online habits and search engine history are being recorded...”
 - **Were there any surprises in the article?**
 - “It shouldn’t matter, as long as you have nothing to hide.”
 - **No, no, no, no... ☹**
 - “What Target was doing ... **was not invasive** ... they just drew a conclusion from the data they were given.”
 - **Do words even mean things**

Consent Matters

- “Even if she didn’t want us to know”
 - “What if advertisements reveal things that you don’t want others to know?”
 - **A legitimate concern.**
 - “I’m all about **freedom of choices** when it comes to this.”
 - **Whose?**
 - “It’s unethical to gather information on people **without their prior approval.**”
 - “I don’t think **most people know** that the things they buy at Target is *being* recorded”
 - “People should be informed that they’re being tracked.”
 - “There is a clear **loss** of ... ‘controlled data’.”
 - “... the uses for all of these public data collections... **promote and influence** certain consumer habits”

The Pragmatic

- “If we wanted to figure out if a customer is pregnant, even if she didn’t want us to know, can you do that?”
 - “I would most likely **try to make an educated decision** to shop at the place with the best deals”
 - **Quick hand-raise poll: who actually thinks they are unaffected by multi-billion dollar marketing research efforts?** 😊
 - “Are traditional retail business practices less effective than this?”
 - **Yes**
 - “If I want the ... ads to stop, **all I have to do** is unlike the pickle page”
 - “... very few people actually do anything to prevent it”
 - **Makes strong assumptions about awareness and technical ability**
 - **Basically only serves to mollify people with the skill to protest**

Real-World Effects

The collage includes several news items:

- Android and Google+ confusion outs trans woman**: The company's decision to integrate its SMS and chat apps has made it too easy for users to leak personal information.
- Trans Woman Commits Suicide Amid Fear of Outing by Sports Blog**: A headline from The New York Times.
- Facebook nymwars: Disproportionately outing LGBT performers, users furious**: A snippet from TechCrunch.
- RAGIC CONFLUENCE OF EVENTS**: A snippet from The New York Times.
- Google+ integrates heavily with Google Hangouts, which is a personal information smorgasbord**: A snippet from The Verge.