

Ethics in AI

Some interesting questions from 20,000 feet



1

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

2

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? | Ideas? |

3

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

4

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?



5

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 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

6

Ultimately...

- 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?**



7

Discrimination and ML

- Machine learning is only as good as its training data
- **GIGO: Garbage In, Garbage Out.**
- 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

8

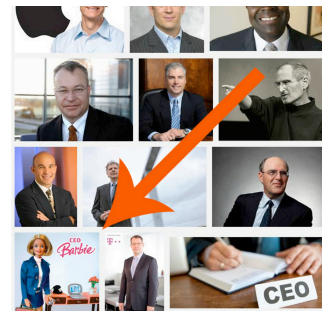
Case 1: Predictive Policing

- 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

9

CEO Barbie

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



the only woman
returned in a
GIS for “CEO”

10

Transl

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 aşçı	she is a cook
o bir mühendis	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 sekreter	he is a secretary
o bir arkadaş	he is a friend
o bir sevgili	she is a lover
onu sevmiyor	she does not like her
onu seviyor	she loves him
onu görüyor	she sees it
onu göremiyor	he can not see him
o onu kucaklıyor	she is embracing her
o onu kucaklamıyor	he does not embrace it
o evli	she is married
o bekar	he is single
o mutlu	he's happy

3:36 PM - 27 Nov 2017

11

How Did This Happen?

- Google Translate is not a “translation” algorithm.
 - It is a pattern-matching, predictive algorithm
- It **reproduces patterns**, whether or not they are good/appropriate translations
 - Mostly they are, and translations come out
 - Sometimes they are not!
- Why not just hardcode gender-neutrality?
 - Very little of it is hardcoded – or even seen by human eyes

12

(Why) Is It a Problem?

- Some translations are wrong
 - Consider:
 - “President’s Erdogan’s cook travels with him; her advice is indispensable”*
 - ← * completely made-up example
 - This may be importantly wrong
- It’s self-reinforcing
 - Once published, text becomes part of Google’s statistical model
- It affects people’s ideas of who can/should do what
 - As mentioned in the CEO Barbie study and others
 - These results and representations **do** affect minds
 - Think they don’t affect yours? Let’s look at those survey results.

13

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, and that could be found out*) and (*I think*) nobody would ever target me for harassment.”


14

Real-World Effects

Google

Android and Google+ confusion outs trans woman

The company's decision to amalgamate its SMS and chat apps has made it too easy for users to leak personal information



Google+ integrates heavily with Google Hangouts, which can leak personal information unwillingly. Photograph: CTK/Alamy

REVIEW: Google Nexus 9: A good, but not great, Android phone


Topic: Security

Facebook nymwars: Disproportionately outing LGBT performers, users furious

Summary: Facebook is enforcing its "real names" policy, insidiously outing a disproportionate number of gay, trans and adult performers -- placing them at risk for attacks, stalking, privacy violations and more.

By Violet Blue for Pulp Tech | September 12, 2014 -- 12:03 GMT (05:03 PDT)


[Get the ZDNet Security newsletter now.](#)



Trans Woman Commits Suicide Amid Fear of Outing by Sports Blog

Tracy Moore 112,859 23

LGBTQ 1/18/14 4:40pm



MAGIC CONFLUENCE OF EVENTS

Web Hannan thought he'd found an interesting idea for a story when he wrote a YouTube video about a "scientifically superior" golf club called the V. He'd out more about the woman behind the invention, but was met with a lot of skepticism and inconsistencies. As he peeled back the layers, he uncovered a