Announcements 4/12

• Please make sure you attend final classes as needed for your full participation grade.

• Check what you think your grades are with Fran (attendance, op-ed, project, briefing, and presentation scores) during office hours or over email. You are responsible for being sure that these are accurate.

• **Topic reports for remainder of class due TODAY.** Please give Fran a hardcopy **AND** send Fran a pdf or your report **with first names only.**
<table>
<thead>
<tr>
<th>Wednesday Section</th>
<th>Friday Lecture (first half)</th>
<th>Second half of class</th>
<th>Assts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 9: NO CLASS</td>
<td>January 11: INTRO – DATA AND SOCIETY</td>
<td>Fran presentation demo</td>
<td></td>
</tr>
<tr>
<td>January 16: NO CLASS</td>
<td>January 18: BIG DATA 1; Topic groups / Topic materials information</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>January 23: Student presentations</td>
<td>January 25: BIG DATA 2</td>
<td>Student presentations</td>
<td>Op-Ed instructions</td>
</tr>
<tr>
<td>January 30: NO CLASS</td>
<td>February 1: DATA AND SCIENCE</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>February 6: NO CLASS</td>
<td>February 8: DATA STEWARDSHIP AND PRESERVATION</td>
<td>Student presentations</td>
<td>Group Topics due</td>
</tr>
<tr>
<td>February 13: NO CLASS</td>
<td>February 15: INTERNET OF THINGS</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>February 20: Student presentations</td>
<td>February 22: DATA AND PRIVACY / FOUNDATIONS</td>
<td>Student presentations</td>
<td>Op-Ed Drafts due</td>
</tr>
<tr>
<td>February 27: NO CLASS</td>
<td>March 1: DATA AND PRIVACY / POLICY AND REGULATION</td>
<td>Student presentations</td>
<td>Briefing instructions Op-Ed Drafts Returned</td>
</tr>
<tr>
<td>March 6: Spring Break</td>
<td>March 8: Spring Break</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>March 13: Student presentations</td>
<td>March 15: DATA AND ENTERTAINMENT [ANDY MALTZ]</td>
<td>Student presentations</td>
<td>Topic Reports 1 due</td>
</tr>
<tr>
<td>March 20: TOPICS PRESENTATIONS 1</td>
<td>March 22: DATA AND DATING</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>March 27: NO CLASS</td>
<td>March 29: DIGITAL RIGHTS 1</td>
<td>Student presentations</td>
<td>Op-Ed Finals due</td>
</tr>
<tr>
<td>April 3: NO CLASS</td>
<td>April 5: DIGITAL RIGHTS 2</td>
<td>Student presentations</td>
<td>Briefings due</td>
</tr>
<tr>
<td>April 10: Student presentations [6]</td>
<td>April 12: DATA AND ETHICS</td>
<td>Student presentations</td>
<td>Topic Reports 2, 3 due</td>
</tr>
<tr>
<td>April 17: TOPICS PRESENTATIONS 2</td>
<td>April 19: CAREERS IN TECH [KATHY PHAM]</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>April 24: Student presentations [5]</td>
<td>April 26: TOPICS PRESENTATIONS 3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lecture 12: Data and ethics

- Ethics and Data
- Ethics and Self-Driving Cars
- Machine Ethics
Ethics and Data

• Multiple areas for development of “data ethics”
  – Data collection and handling (e.g. generation, recording, curation, processing, dissemination, sharing)
  – Data algorithms (e.g. AI, artificial agents, machine learning, robots)
  – Data practice (e.g. responsible innovation, programming, hacking, professional codes)

• (Note that these boundaries are somewhat artificial, most issues have aspects of all 3 …)
Key issues: Collection and handling

- Protection / use of data from research, private public sector
  - What should be open and what should be private? Under what circumstances?
  - When / how to de-identify; how to ensure that identification of groups doesn’t lead to identification of individuals

- Trust – benefits, opportunities, risks and challenges associated with data
  - Role of transparency in fostering trust – when / what / who / how
  - Practice / policy / regulation
  - Liabilities and enforcement
Key issues: Data algorithms

• How to develop “smart applications” that are unbiased?
• How to guide the behavior of learning applications and autonomous systems?
• How to minimize the risk of unanticipated negative outcomes?
• Whose ethics should applications represent?
Key issues: Data Practice

- What is the moral responsibility, accountability, and liability of designers, programmers, companies, consumers?
- What should the guidelines, checks and balances of responsible innovation be?
- What should we consider ethical development and usage?
- How do we trade-off the ethics of dealing with data about individuals and the ethics of dealing with data about groups?
- Who needs to consent to the use of data, when and how?
- What does it mean for data to be private and under what circumstances?
Ethics: Whom should your car hit?

- **Utilitarian ethics:** The best action is the one that maximizes utility

- **Who should be saved?**
  - Smaller vs. greater number of people?
  - You vs. others?
  - Your family vs. others?
  - Young vs. old?
  - “Good” people vs. “bad” people?

- Should the utility function focus on minimizing bad outcomes vs. optimizing good outcomes?

- *Does there need to be a single ethical system for the IoT?*
The Moral Machine Experiment

- **Moral machine experiment**: Multilingual on-line serious game for collecting large-scale data on how citizens would want autonomous vehicles to solve moral dilemmas in the context of unavoidable accidents.

- **“Players”**: Individuals from 233 countries providing information about ~40M decisions
  
  - Geolocation information collected about users
  
  - Users could volunteer age, gender, income, education, religious and political attitudes

- **Play**: Users shown unavoidable accident scenarios with two possible outcomes, depending on whether vehicle swerves or stays on course. Users click on the outcome they find most preferable.
Ethical dilemmas

- Sparing passengers vs. others
- Spare humans vs. pets
- Spare more lives vs. fewer lives
- Spare men vs. women
- Spare young vs. old
- Spare pedestrians crossing legally vs. jaywalkers
- Spare those who are fit vs. those who are less fit
- Spare those with higher social status vs. those with lower social status
- Stay on course vs. swerving (preference for action)
Fig. 2 | Global preferences. a, AMCE for each preference. In each row, \( \Delta P \) is the difference between the probability of sparing characters possessing the attribute on the right, and the probability of sparing characters possessing the attribute on the left, aggregated over all other attributes. For example, for the attribute age, the probability of sparing young characters is 0.49 (s.e. = 0.0008) greater than the probability of sparing older characters. The 95% confidence intervals of the means are omitted owing to their insignificant width, given the sample size (\( n = 35.2 \) million). For the number of characters (No. characters), effect sizes are shown for each number of additional characters (1 to 4; \( n_1 = 1.52 \) million, \( n_2 = 1.52 \) million, \( n_3 = 1.52 \) million, \( n_4 = 1.53 \) million); the effect size for two additional characters overlaps with the mean effect of the attribute. AV, autonomous vehicle. b, Relative advantage or penalty for each character, compared to an adult man or woman. For each character, \( \Delta P \) is the difference between the probability of sparing this character (when presented alone) and the probability of sparing one adult man or woman (\( n = 1 \) million). For example, the probability of sparing a girl is 0.15 (s.e. = 0.003) higher than the probability of sparing an adult man or woman.
Preferences

• **Global preferences** (strong preferences across all respondents)
  – Spare humans over animals
  – Spare more lives over fewer
  – Spare young lives over old

• **Individual preferences**
  – Individual variations over all 6 characteristics have no sizable impact on any of the 9 scenarios.
  – Most notable individual preferences driven by gender and religiosity of respondents
Data clustering

- Geolocation data used to identify 130 countries with >100 respondents.
- Clustering analysis identified 3 distinct “moral clusters” of countries:
  
  - **Western cluster**: North America and many European countries of Protestant, Catholic, and Orthodox Christian cultural groups. [Particular sub-clusters contained Scandinavian countries and another contained Commonwealth countries]
  
  - **Eastern Cluster**: Japan, Taiwan and other countries that belong to the Confucianist cultural group, as well as Islamic countries such as Indonesia, Pakistan and Saudi Arabia
  
  - **Southern Cluster**: Latin American countries of Central and South America. Also some countries characterized in party by French influence, France, French overseas territories and territories that were at some point under French leadership).
Moral Machine Experiment -- results

Image: https://www.nature.com/articles/s41586-018-0637-6.pdf,
Ethical predictors

Differences in ethics:
• Individualistic cultures and collective cultures
  – Respondents from individualistic cultures show greater preference for sparing more characters
  – Respondents from collective cultures show greater preference for sparing elders
• Cultures with greater prosperity and quality of laws and institutions vs. those with lesser
  – Greater preference for law-abiding pedestrians vs. jaywalkers
• Respondents from cultures with more economic inequality treat people of stature somewhat differently
• Women in general were less expendable than men *but* respondents in countries where there are higher ratios of female to male life expectancy and sex ratio at birth saw men as even less expendable than women
Between-cluster analysis

• Clusters largely differ in the weight they give to some preferences
  – Preference to spare younger characters rather than older less pronounced for Eastern cluster and higher for Southern cluster
    • Also true wrt the preference for sparing higher status characters.
  – Preference to spare humans over pets weaker for the Southern cluster
  – Strong preference for sparing women and fit characters in the Southern cluster.

• Authors: “Whereas the ethical preferences of the public should not necessarily be the primary arbiter of ethical policy, the people’s willingness to buy autonomous vehicles and tolerate them on the roads will depend on the palatability of the ethical rules that are adopted.”
Caveats

• Study pool is big but **respondents were self-selecting**
  - Samples should not necessarily be construed as representative

• **No uncertainty** introduced about the classification of characters
  - In real life, one individual is many things

• **No hypothetical relationships** between respondents and characters assumed (family, friends, etc.)
Ethical Guidelines for self-driving cars in Germany


- Commission included 14 scientists and legal experts, and Ministry said it would implement and enforce the guidelines.

- Germany home to major automakers such as BMX, Daimler and Volkswagen.

An initial pass at vehicle ethics ...from the German Federal Ministry of Transport and Infrastructure

2016 Ethical rules for Automated and Vehicular Traffic
(excerpt, paraphrased)

1. Primary purpose of autonomous vehicles is to improve safety and increase mobility.
2. Prevent accidents when possible
3. Save people over everything else
4. Public sector responsible for guaranteeing safety through regulation, policy, enforcement
5. Liability or damage should be governed by the usual product liability principles
6. It should be clear who is controlling the car. Drivers should receive training in the operation of autonomous vehicles.
An initial pass at vehicle ethics ... from the German Federal Ministry of Transport and Infrastructure

2016 Ethical rules for Automated and Vehicular Traffic
(excerpt, paraphrased)

7. In accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited.

8. Genuine dilemmatic decisions, such as the decision between one human life and another cannot be clearly standardized, nor can they be programmed.

9. Permitted business models must respect limitations in the autonomy and data sovereignty of road users.

10. Complete connectivity and central control of all motor vehicles is ethically questionable if it is unable to safely rule out the total; surveillance of road users and manipulation of vehicle control.
Can your morals be characterized?

- How would you represent your own ethics?
- What are the key issues, parameters, frameworks that might be used to distinguish one person’s ethics from another?
- Could we use that to develop customized ethical autonomous systems?

Images and info: https://www.yourmorals.org/explore.php
Machine Ethics

IBM's Watson Gets A 'Swear Filter' After Learning The Urban Dictionary

Watson, the name for IBM's supercomputer best known for crushing "Jeopardy!" contestants at their own game, briefly went from "smart" to "smart ass" with the help of the Urban Dictionary.

According to Eric Brown, an IBM research assistant and the "brains" behind Watson, he and his 35-person team wanted to get IBM's supercomputer to sound more like a real human. In Brown's mind, what better way to learn the intricacies of informal human communication and conversation than having Watson memorize the Urban Dictionary?

The Urban Dictionary, for those who don't know, is comprised of submissions from everyday people and regulated by volunteer editors, who are given an extremely small set of rules to maintain quality.

Alexa or not, voice-activated device called 911 during violent domestic dispute

John Bacon, USA TODAY
Published 10:41 a.m. ET July 11, 2017 | Updated 3:36 p.m. ET July 19, 2017
Creating Ethical Algorithms

• By default, algorithms don’t understand the context in which they act nor the ethical consequences of their decisions.

• How do algorithms become unethical?
  – Biased training sets
  – Biased algorithmic approach
  – Biased interpretation
  – Lack of prediction / monitoring / addressing unintended consequences

• Ethics can be promoted through better training sets, explicit mathematical formulae, pruning of outputs, human monitoring, explain-ability, transparency, etc.
Challenges in creating ethical algorithms

- Algorithm may try to minimize mistakes averaged over all the training data. This may have different “inaccuracies” for different people / groups, e.g. minority groups.

- Algorithm may provide the “best guess” but may have varying levels of confidence in different options.

- Algorithm may learn to predict from historical data that reflects particular biases. New behaviors may also have these biases.

- Ethical considerations may conflict. For example, algorithmic approaches that increase ethical considerations for one group may decrease ethical considerations for another.

Based on information from [http://theconversation.com/ethics-by-numbers-how-to-build-machine-learning-that-cares-85399](http://theconversation.com/ethics-by-numbers-how-to-build-machine-learning-that-cares-85399)
Microsoft’s Tay

- Tay (thinking about you) – artificial intelligence chatbot originally released by Microsoft via Twitter on 3/23/16.
- Tay was developed to mimic language patterns of a 19 year old American girl and learn by interacting with human users over Twitter
- Tay “learned” to post inflammatory and offensive tweets through its twitter account.
- Tay retweeted more than 96,000 times; service shut down 16 hours after its launch.
Tay chatbot

• Tay uses a combination of AI and editorial written by a team of staff including improvisational comedians
  – Relevant, publicly available data that has been anonymised and filtered is its primary source.
• Tay in most cases was only repeating other users’ inflammatory statements, but it also learned from those interactions.
• Need to set up language filters and an environment for both what to say and what not to say …
• “If you’re not asking yourself ‘how could this be used to hurt someone’ in your design/engineering process, you’ve failed.” Zoe Quinn

General ethical guidelines for autonomous systems: Asimov’s Laws of Robotics (circa 1942!)

Over 50 years ago, Isaac Asimov introduced laws of robotics (introduced in the 1942 short story "Runaround" (included in the 1950 collection I, Robot).

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

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.
Update needed for today’s AI

“According to report by Cambridge Consultants, titled “AI: Understanding And Harnessing The Potential,” there are five key areas that rules for AI should address:

1. **Responsibility:** There needs to be a specific person responsible for the effects of an autonomous system’s behaviour. This is not just for legal redress but also for providing feedback, monitoring outcomes and implementing changes.

2. **Explainability:** It needs to be possible to explain to people impacted (often laypeople) why the behaviour is what it is.

3. **Accuracy:** Sources of error need to be identified, monitored, evaluated and if appropriate mitigated against or removed.

4. **Transparency:** It needs to be possible to test, review (publicly or privately), criticise and challenge the outcomes produced by an autonomous system. The results of audits and evaluation should be available publicly and explained.

5. **Fairness:** The way in which data is used should be reasonable and respect privacy. This will help remove biases and prevent other problematic behaviour becoming embedded. “

Are algorithms making us too isolated? (9 min)

https://ed.ted.com/on/h01kSlpF
Ethical responsibility

• **What is a company’s ethical responsibility?**
  – How do you design / engineer / monitor to maximize positive outcomes and minimize negative outcomes?
  – Which ethics? -- how do you decide who it’s OK to offend?
  – What is your liability in the case of harm?
  – When do you pull the plug?

• **What is society’s ethical responsibility?**
  – Legal / policy framework so that business / technology protects / promotes the public good
  – Development of a governance environment to enforce laws and policy and ensure accountability
  – Development of a framework for trading off the ethics of promoting the individual over promoting society
Lecture 12 Sources (not already on slides)

- “What is Data Ethics?”, http://rsta.royalsocietypublishing.org/content/374/2083/20160360
Presentations
April 19


• “What will online dating be like in 2030?” Mashable, https://mashable.com/article/future-online-dating/#gBYFaVsRCZqf [Charlie Y.]


April 24


Today’s Presentations


