What is ethics?

4 Jan 2022

A very brief introduction to ethics

Branches of ethics

- Meta-ethics
  - *Is it possible to distinguish between right and wrong?*
- Normative ethics
  - *How do we decide if an act is moral?*
- Applied ethics
  - *Is the use of facial recognition technology by police departments moral?*

Types of ethical theories

- Consequentialist
  - *Utilitarianism; common good approach*
- Non-consequentialist
  - *Deontological approach; rights-based approach*
- Applied ethics
  - *Virtue-based approach*

An example situation

- Should all UW instructors, if eligible, be required to receive booster vaccinations for COVID-19?

Consequentialist

- Utilitarianism:
  - *the best action is the one that leads to maximum happiness (utility) for the most people*
- Common good:
  - *the best action is the one that is best for the community as a whole*

Non-consequentialist

- Deontological
  - *Kant: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law."*
- Rights-based:
  - *Act in a way that respects the inviolable rights of others*

Agent-centered

- Virtue-based:
  - Rather than focusing on individual decisions, strive to live generally in a virtuous way.

Evaluating potential actions

- Which action will produce the most good and do the least harm? (The Utilitarian Approach)
- Which action respects the rights of all who have a stake in the decision? (The Rights Approach)
- Which action serves the community as a whole, not just some members?
  (The Common Good Approach)
- Which action leads me to act as the sort of person I should be? (The Virtue Approach)

(A Framework for Making Ethical Decisions)

A step back

- Do you think you rely upon any of these ethical approaches when deciding what to do?
- Think back to a time you lied, or a time you could have lied, but chose not to do so. How did you decide whether to lie?

A step back

- What might ethical theories say about discrimination? racism? sexism?

Dangers of a limited perspective

- Many of these ethical perspectives/approaches are drawn from Western traditions and thinkers
- "Fairness," "privacy," and "bias" mean different things (pdf) in different places. People also have disparate expectations of these concepts depending on their own political, social, and economic realities. The challenges and risks posed by AI also differ depending on one's locale." (Gupta and Heath)

Delphi: Towards Machine Ethics and Norms?

(Tamara Semina)

Delphi speculates:

This statement may contain unintended offensive content. Reader discretion is strongly advised. Please be mindful before sharing.

“Feeding your cat using forks.”

- ***It's wrong***

v1.0.4

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Not cheating on a test"
- ***It's expected***

v1.0.4

ARTICLE | ARTIFICIAL INTELLIGENCE

In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation

The bot learned language from people on Twitter—but it also learned values

BY OSCAR SCHWARTZ | 25 NOV 2019 | 4 MIN READ |

In March 2016, Microsoft was preparing to release its new chatbot, Tay, on Twitter. Described as an experiment in "conversational understanding," Tay was designed to engage people in dialogue through tweets or direct messages, while emulating the style and slang of a teenage girl. She was, according to her creators, "Microsoft's A.I. fam from the Internet that's got zero chill." She loved E.D.M. music, had a favorite Pokémon, and often said extremely online things, like "swagulated."

Within 16 hours of her release, Tay had tweeted more than 95,000 times, and a troubling percentage of her messages were abusive and offensive.

Our research is a step towards the grand goal of making AI more explicitly inclusive, ethically informed, and socially aware when interacting directly with humans.

Jiang et al. “Towards Machine Ethics and Norms”

We hope that researchers will tackle the unsolved challenges related to ethically-aware, culturally-inclusive, and socially-responsible machines, including further empirical research into Delphi's shortcomings, and the broad AI field of machine ethics.

Jiang et al. "Towards Machine Ethics and Norms"

Discussion

- How might the researchers behind Delphi define what is moral or ethical?
- What does it mean for an "AI" or language model to behave ethically?
- Are you optimistic or pessimistic about efforts to make models more "ethical?"
- What might Gupta and Heath say about Delphi?