Informatics 2D: Reasoning and Agents

Alex Lascarides

Lecture 30c: AI and Ethics

Where are we?

- We have (deep breath) seen how to model AI decision making while coping with:
  - Uncertainty
  - Dynamic environments
  - Conflicting preferences
- These are powerful tools! Used in:
  - Medicine, autonomous driving, finance,. . .
- Today: **AI and ethics**

Why ethics matters for AI

- AI cannot advance without public trust
  - Full benefits only if perception is positive
  - Anti-vaccine movement shows sound science not enough.
- Progress of AI will stall if researchers don't address human norms, values
- Intelligence itself is inseperable from moral and social dimensions of living.

Ethics matters. . .

- AI already used to automate decisions with profound ethical consequences
  - lending, hiring, parole, bail, education, etc.
- So AI must be well aligned with human values
- AI is a powerful technology;
  anything that creates or redistributes power needs ethical attention

Ethics and Tech are interwoven

- Technologies are not value-neutral:
  - reflect human needs, wants, expectations, judgments.
  - We use tech to bring our values into the world.

Ethics in Preferences

Ethics and engineering share a common task:

1. Translate abstract goals (happiness, justice, duty; utility, efficiency, optimization) into functional material forms (i.e., actions and artifacts)
2. Cope with achieving these goals in a messy, noisy, unstable world.

AI Augments Cognition

- AI can counter harmful biases in human thinking
  but *only if* designed with this aim.
- AI learning/inference/memory can surpass limits in human cognition

Ethical Opportunities for AI

- New medical and scientific breakthroughs
- Improved materials, designs and processes
- Better forecasting of complex dynamic systems
- More affordable goods and services
- Freedom from routine/repetitive tasks
- Cognitive/creative/social ‘upskilling’

Core concerns in AI ethics vs. Machine values

Human Values

Safety, Value Alignment, Privacy, Autonomy and Liberty, Future of Work Accountability/Responsibility, Meaningful Human Control, Transparency, Explainability, Power and Justice, Fairness, Bias, and Equity, Diversity and Inclusion, Wisdom

Machine Values

Optimality, efficiency, speed, precision, predictability, reliability, readability, compressibility, replicability, invulnerability

How does AI fare on (ethical) human values?

Problem with traditonal reward functions:

$$R : S \to \mathbb{R} \qquad \text{or}$$
$$R : F_1 \times F_n \to \mathbb{R}$$

This hides structure among basic human values:

- What's more important? Safety or comfort? Health or wealth?
- What are acceptable margins for any of these?
  - Dilemma of lockdown vs. easing lockdown
- How do we test and audit AI against our values?
  - Transparancy/interpretability of AI models is needed

Real example

Uber autonomous vehicle fatality in Arizona in 2018

- Rides were ‘too bumpy’ for passengers
- So developers (manually) tweaked the reward function to give a more comfortable ride
- The fatality was a direct result of this change.

Fairness/Bias

Bias in training data can have concrete ethical impacts:

- Training data permeated with human prejudices on gender, race, etc. can lead to human prejudice baked into the machine decisions (lending, recruitment)
- Or leads to excluding users. . .

https://www.youtube.com/watch?v=J3lYLphzAnw

Ethical Leadership

1. Modelling ethics as a shared professional task
2. Evaluating ethical implications of our research
3. Aligning our decisions with our professed values.
4. Using ethics to inform and refine our technical and scientific work.

Summary

- You've learned a lot of AI methods for intelligent decision making.
- But ethics must be baked into the design of AI agents at every stage.
- We need much more research on how to align preference models with human values
- AI researchers must talk to philosophers and sociologists!