# Maneesh Wijewardhana

## Introduction pg 13

- Humanity without technology is not a desirable proposition - It's not even a meaningful one. The only meaningful questions are:
  - ◇ Which technologies shall we create
  - ◇ with what knowledge and designs
  - ◇ **affording what**
  - ◇ **shared with whom** (which groups, what users, risk and benefit distributions)
  - ◇ **for whose benefit** (who decides, who decides to decide)
  - ◇ **and to what greater ends**

## Introduction pg 1

- Think carefully about how we shape tech and how it shapes us
- "How can humans hope to live well in a world made increasingly more complex and unpredictable by emerging technologies"
- what are the different answers that ethicists and moral philosophers have about living amongst AI
- **normative theory** (theories of right and wrong action)
- theories in science allow us to explain phenomenons and predict them as well
- ethical theories do the same thing
  - ◇ explains why a given action is right or wrong
  - ◇ explains why in the future, a given action would be right or wrong
  - ◇ theory of consequentialism (always act in a way where it maximizes the good and minimizes the bad for the most amount of people)

## Discussion

- What kinds of things would you want people to say about you in the eulogy at your funeral, or in your obituary?
  - ◇ Tell the truth of who I was and be honest since everyone has imperfections

### Virtue Ethics

- Characteristic moral virtues include generosity and honesty (what would the honest person do? What would the compassionate person do?)
- Characteristic intellectual virtues include open-mindnesses and humility
- both are dispositions (generous person is disposed to offer help when needed)
- dispositions of **character** (contrasted with personality)
- They are formed by habits and informed by exemplars (what would the person I truly admire do?)
- Further distinctions
  - ◇ executive vs substantive
  - ◇ high fidelity vs low fidelity
- Virtue ethics is about the person and character traits rather than action centered

### Refining Virtue

- Big 3: Kantianisms/Deontology, Consequentialism, Virtue Ethics
- "Religious laws and norms speak only to their believers, and thus are poor candidates for a global technological ethic"
- Consequentialism/Utilitarianism, which employs a **universal moral calculus designed to maximize the greatest good for all concerned**
- Decouples the moral worth of acts from the moral worth of persons (example: organ harvesting $\rightarrow$ 5 ppl living better than 1 living person but should you sacrifice?)
- Does not account for certain rights and focuses on the outcome without the process that leads to it
- **Advantage WRT to technology is that it gives us one rule to apply to code and software (simple and elegant)**
- Kantian deontology (what are the intentions of your actions and can those be universalized and shared amongst others and result in good)

### Normative Theory Heuristics

- Consequentialism/Utilitarianism: Outcomes

- Kantianisms/Deontology: Intentions
- Virtue Ethics: Character (What would x person do) ← Valor agrees with this

**What is AI**

- If a human did it then it's intelligent but because a machine did it, then it's artificial intelligent (human-centric)
- Also does things that no human can possibly do (high freq stock trading, internet search engine, etc.) this is why human centric definition is not accurate
- Should we try to make AI that approximates us? Or exceeds us?

**Scope of the book**

- Class of machine learning which is called predictive models
- Predictions can be user-facing (spotify, netflix, etc.) or behind the scenes (criminal justice, loans, etc.)
- Generally produce 'simple' outputs like probabilities, classifications, decisions
- Contrast with **generative models** which produce 'rich' outputs like text, images, video, etc.

**A Central Thesis: Continuity**

- 'in the media, AI is often portrayed as a brand new arrival something that's suddenly, and recently, started to affect public life'
- In the book: Want to emphasize that AI models currently are very much a continuous development of statistical methods
  - ◇ Lloyd's Register for shipping (1688)
  - ◇ Equitable Life Insurance Co. (1762)
- Is AI radically new? If so, do we need new frameworks?

**A Note on 'Neural' networks**

- Two tracks of AI (Deep learning)
  - ◇ Brute computation (transformers)
  - ◇ Biological plausibility (computational cognitive neuroscience)

**Evaluating Algos**

- Always will be a trade-off between false positives and false negatives and vice versa
- In different domains, we might want to err on one side or the other (example credit card companies should want to make sure fraudulent claims are 100% accurate)
- Example is 'innocent until proven guilty' (might have false negatives, but it is a trade-off)

**What is ChatGPT?**

- Large Language Model (100s of billions, even trillion parameters)
- **Generative** Pre-trained Transformer
- Earlier LM's were predictive