Week 2: Philosophical Underpinnings

Ethics in NLP: Including Society in Discourse and Design

January 17th, 2019

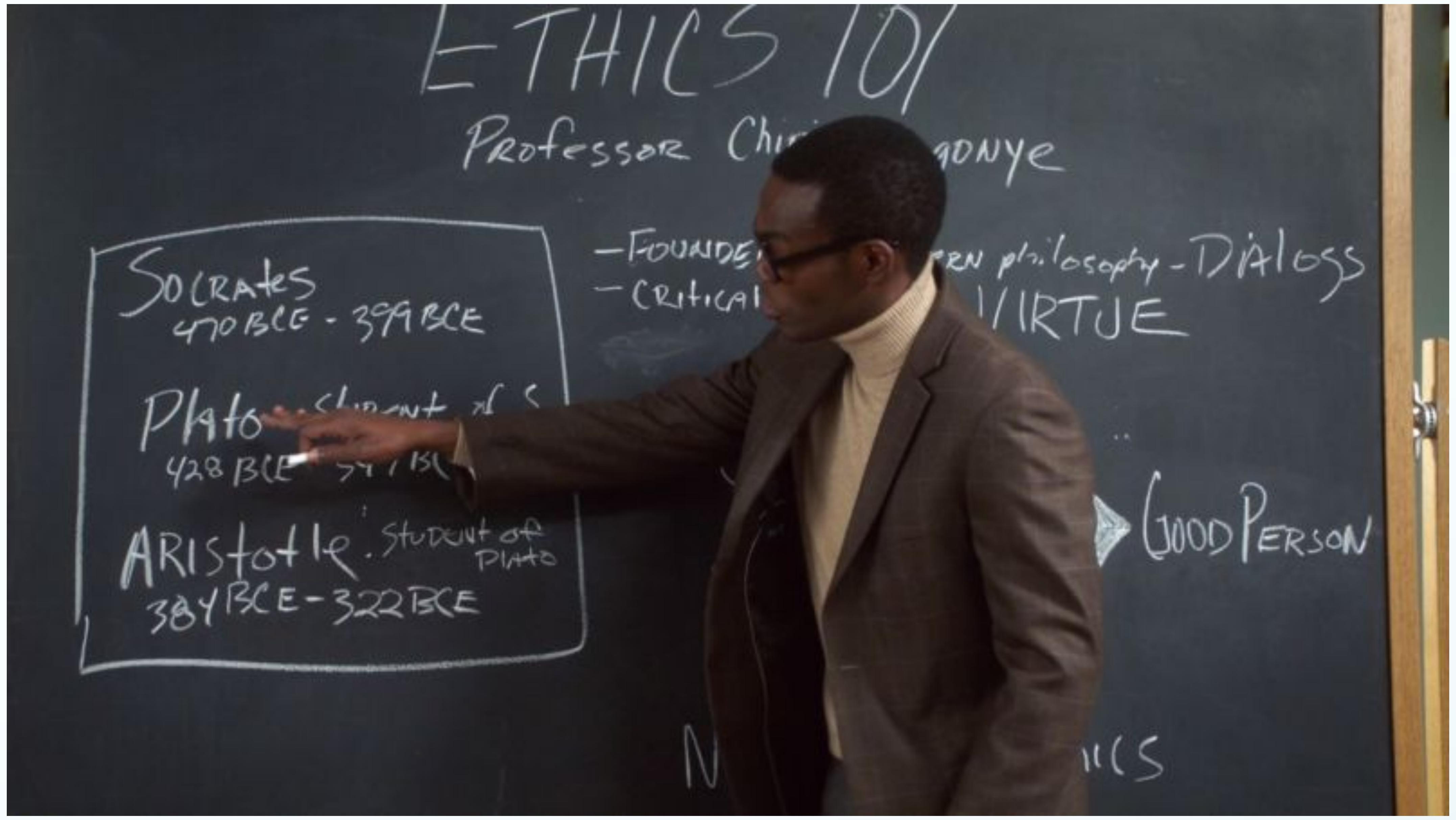
Ryan Georgi

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Philosophical Underpinnings: Purpose

- **Avoid** the philosophical rabbit-holes of arguing for a given framework
- **Focus** on familiarizing ourselves with rationale of each framework
  - Know what framework we are advocating
  - Know what framework others are advocating (if any)
  - Strengths of that framework
  - Weaknesses of that framework



Philosophical Underpinnings: Purpose

- Also, less interested in discussion of independence of choice for AI
  - …a question for philosophers, and is important
  - But I want to focus on what we, as system designers, build and value

Philosophical Underpinnings: Overview

- **Teleological** Frameworks
  - Frameworks that focus on the results of actions as defining their morality
    - **Egoism**
    - **Utilitarianism**
- **Deontological** Frameworks
  - Frameworks that focus on the actions themselves and rationale of the actions
    - **Kantian Philosophy**
    - **Virtue Ethics**

Philosophical Underpinnings: Overview

- **Teleological** Frameworks
  - Frameworks that focus on the results of actions as defining their morality
    - **Egoism**
    - **Utilitarianism**
- **Deontological** Frameworks
  - Frameworks that focus on the actions themselves and rationale of the actions
    - **Kantian Philosophy**
    - **Virtue Ethics**

  - Largely drawing upon the discussion in Edgar (2002)

Utilitarianism

- In essence: "***greatest amount of pleasure for the greatest number of people***"
  - Actions aren't categorically good or bad; it depends on the outcome of the action
- Other targets to maximize besides pleasure:
  - Knowledge
  - Health
  - Aesthetics

Utilitarianism

- **Benefits**:
  - Framing intrinsically takes into account society
- **Problems**:
  - If the actions themselves are neutral, this calculus could be used to justify any action.
    - e.g. — a society where many derive pleasure from watching the torture of others might come out ahead in this calculus
  - If morality is calculated according to a formula, are the actions moral?
    - Is independent action required?
    - Who gets to set the parameters of the formula?
  - How do we know what maximizes pleasure for others?

Egoism

- Every agent should act to maximize their own self-interest
- **Benefits:**
  - Individuals are in many ways the best situated to know what would help them the most
- **Problems:**
  - Individuals don’t always know what is their own best interest
    - Or, they know but don’t act in it

Utilitarianism vs. Egoism

- Should we design systems that make assumptions about the best interests of individuals? Or allow them to define that?
- **Example**:
  - For a social media feed where we analyze content with NLP — should we:
    - Optimize based on what users want to see, even if we have good evidence that what they say they want to see is detrimental to mental health?
    - Ignore user's wishes, and build a system that promotes content that our data suggests promotes the best mental health?
      - Other desirable outcomes?

Deontological Ethics

- The only things that are "good" are things that are good without qualification
- Kant defines a number of "Categorical Imperatives" to help define these:
  - Act only according to that maxim by which you can at the same time will that it should become a universal law
  - Act as though the maxim of your action were by your will to become a universal law of nature
  - Act so that you treat humanity, whether in your own person or that of another, always as an end and never as a means only
  - Act by a maxim which involves its own universal validity for every rational being
  - …

Deontological Ethics

- **Benefits:**
  - Discussing "rational beings" does generalize well to AI
    - (Seem familiar to Isaac Asimov fans?)
- **Problems:**
  - Defining everything as based on "rationality" and "independence of choice" still doesn't define moral content

Virtue Ethics

- “Virtue” is something that is practiced, and based on a character that is developed over time
- For an act to be virtuous, you must
  - (1) know that what you are doing is virtuous;
  - (2) choose the act;
  - (3) do that act for its own sake; and
  - (4) act according to a fixed, unchanging principle, or out of a fixed character