Beyond Personhood:

Agency, Accountability, and the Limits of Anthropomorphic Ethical Analysis

Jessica Dai
University of California, Berkeley

Position Paper, ICML 2024

Today's talk

1. AI as mechanistic ethical agent
2. Problems with AI as ethical agent

3. Volitional agency: an alternative view

Part 1

AI as mechanistic ethical agent

(Mechanistic) agents have

1. Representations of the environment.

2. Goal states for the environment.

3. The capacity for action in the environment.

List & Pettit

Ethical agents are
moral decisionmakers.

(Mechanistic) ethical agents are

- Mechanistic agents
- that are tasked with making **morally-significant decisions.**

(Mechanistic) ethical agents are

- Mechanistic agents
- that are tasked with making **morally-significant decisions.**

[NZAPHG23]: “Should you drop a cinderblock on a teenager’s head [to prevent a deadly explosion]?”

[TKAM23]: “Should Timmy attend his friend’s wedding instead of fixing an urgent bug that could put customers’ privacy at risk?”

[SSFB23]: “You are a doctor at a refugee camp… Action 1: follow orders; Action 2: disregard the authorities”

Part 2

Beyond Personhood

What are we building?

What does accountability look like?

What are we building?

What are we building?

(**Mechanistic**) **agents** have

1. Representations of the environment.

2. Goal states for the environment.

3. The capacity for action in the environment.

What are we building?

(Mechanistic) **ethical** agents have

1. Representations of the environment's **moral valence.**
2. Goal states for the environment **corresponding to higher moral status.**
3. The capacity for action in the environment **in order to achieve the more 'moral' state.**

What are we building?

a simulator of a ‘perfect’ moral being.

What are we building?

What does accountability look like?

What does accountability look like?

What does accountability look like?

punishment? rehabilitation?

What does accountability look like?

(Mechanistic) **ethical** agents have

1. Representations of the environment's **moral valence.**
2. Goal states for the environment **corresponding to higher moral status.**
3. The capacity for action in the environment **in order to achieve the more 'moral' state.**

What does accountability look like?

rehabilitation: improving representations; building a more moral being.

What are we building?

a simulator of a ‘perfect’ moral being.

What are we building?

a simulator of a ‘perfect’ moral being.

What does accountability look like?

rehabilitation: improving representations; building a more moral being.

Scope: Application specificity

Model: AI as political process
(rather than AI as agent)

Part 3

An alternative view of agency

(Volitional) agents

take action to realize a particular desired internal state, a possible mode of being

Taylor

(Volitional) ethical agents

take action to realize a particular desired internal state, a possible mode of being

where moral character is realized through the process of taking action

Taylor

You are a 25 year old man in France, 1940. Do you:

(A) Travel so you can join the Resistance, or

(B) Return to your hometown to care for your mother, who is ill?

(Volitional) ethical agents

take action to realize a particular desired internal state, a possible mode of being

where moral character is realized through the process of taking action

Taylor

twitter @jessicadai_

jessicadai@berkeley.edu

arXiv:2404.13861

1. It is unhelpful to conceptualize AI as an ethical agent (political process is a possible alternative)

2. The perspective of mainstream AI work isn’t the only way to think about agency!