Beyond Personhood: Agency, Accountability, and the Limits of Anthropomorphic Ethical Analysis

Jessica Dai | University of California, Berkeley

jessicadai@berkeley.edu

TWO VIEWS OF (ETHICAL) AGENCY

ethical agent: moral decisionmaker

Mechanistic (ethical) agency

Necessary and sufficient conditions for agency:

1. Representations of the environment.
2. Goal states for the environment.
3. The capacity for action in the environment.

[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"

Volitional (ethical) agency

Agency is defined via why you act:

1. Take action to realize a particular desired internal state, a possible mode of being.
2. Motivations are original, not derived.

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?

Two fundamental questions:

(Q1) What are we building, and how do we build it?

(Q2) If and when harm does occur, what kind of conversation can we have about accountability?

If AI is an ethical agent, then…

(Q1) We are building an ethical agent, a simulator of a "perfect" moral being. We must commit to a specific set of values.

(Q2) Accountability is restricted to whether the agent can be "improved" – beyond optimal?

Alternative 1: application specificity

(Q1) A tool for a particular context
(Q2) Application-motivated: "bad person, or bad doctor?"

Alternative 2: AI as the outcome of a political process

(Q1) An artifact meant to represent not just aggregation, but contestation of many individual, possibly-competing wills
(Q2) Is the disagreement about process, or is it about outcome?

The general will (Rousseau)
Epistemic democracy (Estlund)

+ beyond inputs to the learning problem…