AI and Mental States

Katja Maria Vogt

ValuesLab

Plan for Today

- Beliefs, and more generally, mental states
- Why beliefs, rather than consciousness, intelligence, and so on?
- Mental states versus ascriptionism

The Language of Mental States

The cat is *hunting*. It *believes* that there is a mouse over there. It *intends* to catch it.

Ascriptions of Mental States

When we talk like this, we ascribe mental states to the cat.

- "… is hunting": we ascribe an *end* to the cat, that it does something (walk noiselessly) for the sake of something (catching the mouse).
- "believes": we ascribe a belief (that there is a mouse over there)
- "intends": we ascribe intention, that it does what it does in ways that are guided by her ends.

Do Animals Have Beliefs?

Do cats have beliefs? Isn't belief a mental state that is distinctive of human minds?

VOTE 1:

YES (cats have beliefs), NO (cats don't have beliefs)

We'll take more votes later…

Analogous Questions About AI

It is contested whether AIs (AI systems, models, LLMs) have beliefs.

- Like animals, AIs don't have human minds.
- Perhaps AIs are even more deeply different from us than cats, because they are not *natural* creatures; they are *artifacts*.
- On the other hand, AIs can produce answers to questions in linguistic form, which cats can't.

Option 1: Ascriptionism

- If a behavior is best explained by ascribing beliefs to X, X has beliefs.
- What about the worry that the AI doesn't have a mind?
- Forget about the mind!

Option 1: Ascriptionism—Why?

- It's not clear what it would mean to ascribe beliefs, desires, intentions, etc., to AIs.

- But it's also not entirely clear what it means to ascribe beliefs, desires, intentions, etc., to us!

Option 1: Ascriptionism—Why? (Continued)

- Talk about beliefs, desires, intentions, etc., plays an *explanatory* role.
- We explain behavior by ascribing these states.
- If that's an important dimension of talk about mental states, why not do the same w.r.t. AIs?

Ascriptionism (based on Schwitzgebel 2024)

Interpretationism: A revision of dispositionalism.

Daniel Dennett: When we explain observable behavior, we take three stances: physics, design stance (about functions of organs, etc.), or the intentional stance, where we ascribe intentions and other mental states.

According to interpretationism: "[t]he system has the particular belief that P if its behavior conforms to a pattern that can be effectively captured by taking the intentional stance and attributing the belief that P." (Schwitzgebel 2024)

Ascriptionism (based on Schwitzgebel 2024)

Dispositionalism: Beliefs are behavioral dispositions. For someone to believe that P is for them to have some behavioral dispositions pertaining to P.

For example, for the cat to believe that there is a mouse over there is to walk a certain way, listen closely, and so on.

Objection 1: This is reductionism. Why reduce beliefs to behavior?

Objection 2: What about beliefs that don't have obvious behavioral correlates?

Option 2: Traditionalism

- Traditionalist: Beliefs are mental/psychological states that are distinctively human.
- Analogously, desires, intentions, and other mental states are distinctive dimensions of *our* mental lives.

Option 2: Traditionalism

Murray Shanahan (2024):

It is tempting to say things like “the AI believes,” but it is a form of anthropomorphism.

AIs simply aren’t the kind of thing that has beliefs. They are “generative mathematical models of the statistical distribution of tokens in the vast public corpus of human-generated text…”

Traditionalism vs. Ascriptionism

Whether we are ascriptionists or traditionalists determines our answer to whether AIs have beliefs:

- Ascriptionism: yes
- Traditionalism: no

Traditionalism vs. Ascriptionism

Whether we are ascriptionists or traditionalists hangs on:

- What we think beliefs/etc. are (today)—What *are* beliefs?
- What we think the human mind is (next class)

Minimal Idea 1: Holding True

Suppose you hold true that the window is open.

You *represent* a state of affairs, that the window is open.

Such representations are a basic feature of the human mind.

- Does the cat represent that there is a mouse over there? Perhaps!
- Does the AI represent such-and-such? We postpone that question.

Minimal Idea 2: Belief as a Propositional Attitude

A couple of examples for propositional attitudes:

1. Sara hopes that there is a class on AI in the Fall.

2. Sara believes that there is a class on AI in the Fall.

3. Sara knows that there is a class on AI in the Fall.

Cognizer—attitude—*that P*.

P [proposition] is, in this case, *there is a class on AI in the Fall*.

To hope, to believe, to know, and so on, are attitudes to P.

Minimal Idea 3: Truth and Falsity

Belief is a non-factive propositional attitude.

That is, if Sara believes that P, it can either be the case that P (be a fact that P) or not be the case that P (not be a fact that P).

Here's another way of putting this: If Sara believes that P, P can be either true or false.

Minimal Idea 4: Truth as the Aim of Belief

Truth is the *aim* of belief.

When you form a belief, you aim to hold something to be true that really is true.

Cf. Williams (1973)

Do AIs Have Beliefs?

Do AIs have beliefs?

VOTE 2:

Ascriptionism (some version of dispositionalism or interpretationism)

Traditionalism (beliefs are distinctively human mental states that relate to complex attitudes and skills, such as sincerity/insincerity, assertion/lying, etc.)

Intentions

We didn’t only ask about beliefs. We also asked whether to ascribe *intentions* to the cat.

Intention, Ends, Responsibility, Etc.

In ascribing intentions to some entity—a human being, a cat, an AI—we talk about it as an *agent*. Here are a few related notions:

- agents
- action
- intention
- ends
- responsibility

A Third Option?

- Is there a third option?

- Some philosophers work toward what one might call moderate ascriptionism. Roughly, the thought is that there's something compelling about ascriptionism, but it doesn't tell the full story. We don't want to forget about the mind!

A Third Option?

Cibralic and Mattingly (2021) defend a third option w.r.t. responsibility.

- In the traditionalist framework, we can't ascribe responsibility to AI, because we can't ascribe representations.
- But we *want* to be able to ascribe responsibility. Why? More on the next slide.
- They propose a minimalist account of representation for AI.

The Motivations of Ascriptionism

Why do we *want* to ascribe representations (and perhaps, more generally, mental states)?

- So far, we discussed something like *inference to the best explanation*: the best way to make sense of a given behavior is to ascribe mental states.

- Cibralic and Mattingly introduce another kind of motivation: we want to be able to distinguish between the responsibility of those who built the AI, and the specific individual outputs, which (in some sense that is TBD) the AI is responsible for.

Responsibility Gap

Cibralic and Mattingly's motivation responds to the so-called responsibility gap (Andreas Matthias, 2004):

- Those who build an AI are responsible for its overall design.
- But they cannot predict a specific output at a given occasion.
- Hence, there is something that someone else should be responsible for.
- That someone else might be the AI.

Moderate Ascriptionism?

The motivation for *wanting* a moderate ascriptionism seems strong:

- We may not want the reductionism of not caring at all about the mind; this rules out strict versions of ascriptionism.
- It is tempting to ascribe mental states to AI. But it may only be a metaphor, not an explanation.
- We may need the Cibralic/Mattingly distinction between what the AI designer is responsible for versus specific outputs "by" the AI.

Does this get us all the way to responsibility? More on the next slide.

However…

Does this get us all the way to *responsibility*? Why not say that:

- The AI *causes* the outputs.
- In some sense, *no one* is responsible for the specific outputs, because AI isn't an agent who is suitably held responsible.
- We can't just assign responsibility to X because we need someone to blame…

Take-Away and Questions

- When we ascribe beliefs, intentions, etc., to AIs, is that just for ease of expression, a façon de parler?
- Do we have philosophical reasons, grounded in what we take AIs to be, to ascribe mental states to them?
- What about contexts where something goes wrong? Do we need to be able to ascribe responsibility to AIs?
- Or do we only need causal and mathematical explanations?
- We looked at belief and intention. There's a host of similar issues. For example: can an LLM *speak*?

Readings

- Beba Cibralic and James Mattingly, “Machine agency and representation” (2021)
- Eric Schwitzgebel, “Belief,” The Stanford Encyclopedia of Philosophy (Spring 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/spr2024/entries/belief/>
- Murray Shanahan, “Talking about Large Language Models” (2023)
- Alan M. Turing, “Computing Machinery and Intelligence” (1950)
- Bernard Williams, “Deciding to Believe” (1973)