Do AIs have beliefs? Do they have intentions?

Katja Maria Vogt

ValuesLab

Plan for today

- Why beliefs, rather than consciousness, intelligence, and so on?
- Example: Do animals have beliefs?
- What *are* beliefs?
- Do AIs have beliefs?
- What are intentions? Do AIs have intentions?

Two lines of inquiry

- You already studied Turing's "imitation game" aka Turing Test.
- Turing's initial question was "can machines think?"
- This question inspired extended debate about *intelligence*.
- It also inspired discussion of other highly regarded capacities in the human mind, first and foremost *consciousness*.
- But what is intelligence? What is consciousness?
- Let's start in a more pedestrian fashion… (we can always return!)

Beliefs and Intentions

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?

Let’s start with beliefs, and postpone ends/intention.

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.

What *Are* Beliefs?

Whether we think that AIs have beliefs depends in part on what beliefs *are*.

Let's start with three minimal ideas:

- Belief is holding true.

- Belief is a propositional attitude, that is, an attitude to a proposition or a state of affairs.

- Belief aims at the truth.

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.

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.

Belief is Non-Factive

Another look at examples:

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.

4. Sara sees that there is a class on AI in the Fall.

3 and 4 are *factive*. If Sara knows/sees that there is a class on AI in the Fall, it is a fact that there is a class on AI in the Fall.

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.

Simplified: Beliefs can be true or false.

Why is this a simplification? Strictly speaking, propositions—not beliefs—are true or false.

Truth as the Aim of Belief

If you believe that there's a class on AI in the Fall, you want to get right whether there's a class on AI in the Fall.

Truth is the *aim* of belief.

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

The Options

Types of positions on the question of whether AIs have beliefs:

- Traditionalist: Beliefs are mental/psychological states that are distinctively human.

- Ascriptionist: If a behavior is best explained by ascribing beliefs to X, X has beliefs. Forget about the mind!

- 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!

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…”

Bernard Williams (1973)

In a classic paper, Bernard Williams (1973) ascribes the following features to belief:

1. Belief aims at the truth.

2. The most straightforward expression of belief is assertion.

3. The assertion that P is neither a necessary nor a sufficient condition for the belief that P.

4. Factual beliefs can be based on evidence. Even if they are, they have a causal history.

5. Belief is an explanatory notion. We can explain what someone does by invoking their beliefs.

Bernard Williams on B-States

Williams imagines a machine that provides answers to prompts

- the machine can only have a lesser kind of belief, which he calls B-states;
- the machine produces assertions;
- but it can’t be insincere;
- it is a distinctive feature of human belief that belief and assertion can come apart.

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?

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)

Do AIs Have Beliefs?

Do AIs have beliefs?

VOTE 2:

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

Ascriptionism (some version of dispositionalism or interpretationism)

Intentions

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

Traditionalism and Ascriptionism about Intention

Do AIs have intentions? Let's apply the two frameworks:

Traditionalism: Intentions are distinctively human mental states.

Ascriptionism: X has intentions if its behavior is best explained by ascribing intentions to it.

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?

Cibralic and Mattingly (2021) defend a third option, in between traditionalism and ascriptionism, 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, and only about behavior; this rules out strict versions of ascriptionism.

- It is tempting to ascribe mental states to AI, because this looks like the best explanation. But it may only be a metaphor, not an explanation. Still, perhaps some dimension of this can be saved?

- 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?

- Do we have philosophical reasons, grounded in what we take AIs to be, not to ascribe mental states to them?

- What about contexts where something goes wrong? Do we need to be able to ascribe responsibility to AIs? Perhaps even, to blame them? 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)