Can AI Think?

Katja Maria Vogt

ValuesLab

Plan for today

- We turn to the question of whether AIs can think/are intelligent.
- Recap: The Belief Desire Framework
- Against the Belief Desire Framework
- The Predictive Brain Framework
- Brain and Mind
- Behavior and internal processes

The Belief Desire Framework

In Class 1, we worked with a long-standing distinction between two kinds of mental attitudes, going back to David Hume (1711-1776):

- Belief
- Desire

The Desire Belief Framework—Direction of Fit

Elizabeth Anscombe, two "directions of fit" (*Intention* 1957):

- Mind-to-world: In belief, we aim to bring the *mind to the world* = we aim to fit our minds to evidence about how the world is.

- World-to-mind: In desire, we aim to bring the *world to the mind* = make the world as we desire it to be.

  - Upshot: everything that's going on in the human mind is either belief-like ("cognitive") or desire-like ("conative").

Objections to the Belief Desire Framework

Objection 1 w.r.t. Belief Desire Framework:

- Is talk about beliefs and desires mere “folk psychology”?

- I.e., this may be how we talk when we make sense of ordinary situations; but it’s not a scientific approach to the mind.

- We need a more precise approach.

Objections to the Belief Desire Framework

- Objection 2: Why should we assume that everything that's going on in the human mind is either belief-like or desire-like?

- Alert: Later, we'll apply this to questions about AGI!

- Railton on the default mode of the human mind: "episodic and semantic memory, scene construction and the imaginative simulation of possible futures, counterfactual reasoning, inferring the mental states of others, self-referential processing, and ethical judgment." (2020, 57)

Railton, "Ethical Learning, Natural and Artificial" (2020)

Some default activities cut across the desire-belief divide:

- Memory: Analyzable as beliefs? Stored information? Affectively colored, shaped by how a person reconstructs her life, etc.

- Imagination: Involves beliefs about what is the case, counterfactual beliefs, etc. But also anticipatory pleasure/pain, etc.

Objections to Belief Desire Framework

- Objection 3: The Belief Desire Framework does not capture the predictive-probabilistic nature of what's going on in the brain.

- We'll pursue this line of thought by looking at Railton's argument against innatism.

- Innatism about X: X is an inborn ability/inborn knowledge; it doesn't seem possible that X was learned.

- Against innatism: Yes, it is possible that X was learned, namely, if we think of key processes as probabilistic.

Railton (2020)—Innatism

- Explanandum: how does it work that small children learn the language, even in the absence of explicit language instruction?

- This is a very significant accomplishment!

- Traditionally, psychologists thought language learning in infants is only explicable if there is are *innate* modules and *innate* grammar.

Railton (2020)—Probabilistic Learning

- Machine learning suggests a different model for human learning.
- Children have:
  - Very large amount of overheard language
  - Very large amount of fast, flexible computational capacity
- Children may learn the language through probabilistic means.

Railton (2020)—The Predictive Brain

According to Railton, we have evidence that infants

> "are beginning to form calibrated expectations about phonetic regularities, which are manifest in greater surprise at, and interest in, novel or anomalous sequences of phonemes—a characteristic feature of probabilistic learning." (2020, 51)

Probabilistic learning in children is

> "… a form of active experimentation, with the continuous formation of expectations on the basis of observed associations and continuous feedback from discrepancies between such expectations and actual outcomes." (2020, 51)

Helmholtz (1860), following Clark (2013)

- Perception is "a process of probabilistic, knowledge-driven inference."

- Sensory systems *infer* "sensory causes from their bodily effects." "… the brain does not build its current model of distal causes (its model of how the world is) simply by accumulating, from the bottom-up […] the brain tries to predict the current suite of cues from its best models of the possible causes." (p. 2)

- Desire is also explained probabilitistically, in ways that dissolve the belief-desire-distinction (see Appendix).

Railton (2020): Ethical Learning in Humans

Causal and social learning:

- "And no part of the infant's causal environment is more important for her than the *agents* in her life, so that causal and social learning are intimately linked, and intuitive psychology emerges alongside intuitive physics." (2020, 51)

- Theory of mind: learning to ascribe mental states to others.

Railton (2020): Rejecting a "Moral Module"

Innatists postulate a moral module. Railton argues against this:

> "… it was thought that there might be some region or regions of the brain specialized for ethical judgment. By contrast, the approach to ethical development sketched here would predict that the neural substrate of ethical judgment would involve regions or networks subserving general-purpose learning and judgment concerning a range of causal and theory-of-mind–related questions about situations, actions, outcomes, and agents." (2020, 57)

Predictive Brain Framework

- Suppose that what's going on in the mind doesn't fall neatly into two categories, belief and desire.

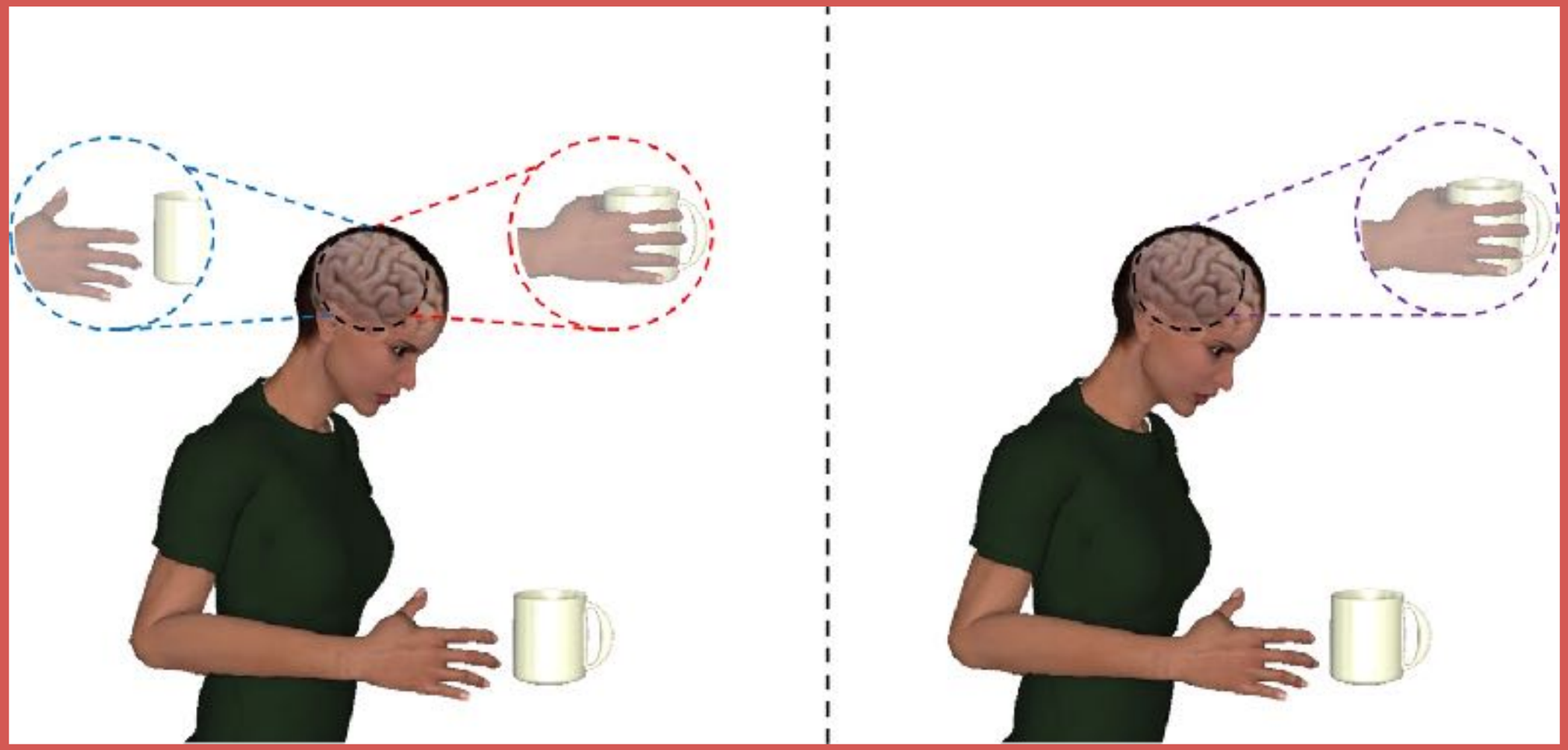
- Framework that explains what's going on the mind in terms of only one kind of state: Predictive Processing Framework.

- The brain is a probabilistic prediction engine (Yon et al, "Beliefs and Desires in the Predictive Brain," 2020):

  - Belief-like representations of which states of the world are most probable.

  - Desire-like representations of which states of the world are most valuable.

Desire Belief vs. Predictive Mind, Yon et al (2020, p.2)



Left: Desire (red) and belief (blue). Right: Prediction.

Yon et al, "Beliefs and Desires in the Predictive Brain" (2020)

- Example: "the hungry rat presses the lever because it expects itself to press, since it expects not to be hungry in the future." (p. 2)

- How would this work for more complex desires?
  - desire for world peace,
  - desire to fast for religious reasons,
  - desires one finds oneself with but tries not to have, etc.

Yon et al, "Beliefs and Desires in the Predictive Brain" (2020)

- The brain predicts something.
- New input.
- There's a mismatch between prediction and input.

"… mental states once thought to be crucial in explaining behaviour —such as goals, drives and desires—are reduced to predictions."

- no essential difference between desires and beliefs
- desired outcomes = those that an agent believes it will obtain

Companions in Guilt

Imagine the following exchange:

- “The LLM is a next token predictor, based on a probability calculus. Hence, it doesn’t have beliefs, desires, etc.”
- “The human brain is also a probabilistic next token predictor!”

If this is plausible, what follows?

Mind? Brain? Psyche? Soul?

When we compare AIs and humans, what are we talking about?

- The brain? Roughly, the brain is an organ, studied in empirical fields including neuroscience, physiology, medicine, etc.

- The mind? Researchers disagree on whether the mind can be fully explained by the empirical sciences.

- Psychological states? Researchers disagree on whether psychology is a purely empirical field.

- Soul? Religion and theology discuss the soul, in ways that self-consciously depart from the sciences.

The Simple Solution

- The Predictive Processing Framework is about the brain.
- The Belief Desire Framework is about the mind.
- By not being about the same, both frameworks are compatible!

VOTE: Are the frameworks compatible? Yes—No—Abstain.

The Simple Solution Applied to AI

Is there a distinction w.r.t. LLMs that is analogous to brain vs. mind?

- Internal process: probabilistic next token prediction (PNTP).

- Behavior/outputs: look like expressions of beliefs/etc.

- Upshot: that LLMs are PNTPs does not resolve the question of whether they have beliefs, desires, etc.

Alan Turing, “Computing Machinery and Intelligence” (1950)

- “Can machines think?” Turing considers the question not sharp enough.
- Let’s replace it by another, one that is “expressed in relatively unambiguous terms.” But: there’s no revised question!
- Instead Turing introduces the Imitation Game. Today: Turing Test.
- Imagine a person, a judge or interrogator, who sits in a room. There are two other rooms. In one of them, there’s a machine; in the third room there’s another person. The judge poses questions, and assesses the replies from the machine and the person.If the judge cannot discern which answers come from the machine, the machine passes the Turing Test.

Alan Turing, “Computing Machinery and Intelligence” (1950)

“May not machines carry out something which ought to be described as thinking but which is very different from what a man does?” (435)

- What a machine does

- How we talk about it

Alan Turing, "Computing Machinery and Intelligence" (1950)

- Concern: is it a mistake to infer from how we *talk about a machine* to what it *really does*?

- Contrary to this impression, Turing says: If a "machine can be constructed to play the imitation game satisfactorily […] we need not be troubled by this objection." (435)

- Why not?! Why not say that the machine succeeds as imitating thought, rather than say, that it thinks?

Turing's Legacy

What type of position is, in today's terms, close to Turing's view?

- Ascriptionism? If this is the best explanation of X's behavior, let's go ahead and ascribe intelligence to X.

- Computationalism? There's no difference between mind and the internal process; "mental processes are computational processes; the brain, qua computer, is the seat of cognition." (Shapiro and Spaulding, "The Embodied Mind," SEP, 2025)

Ned Block, “Psychologism and Behaviorism” (1981)

- Ned Block (1981) rejects the conflation of, in his terms,
  - The internal process in a machine
  - The machine’s behavior
- The internal-vs-behavior distinction provides two options, which Block calls Psychologism and Behaviorism.

Internal Process vs Behavior—Psychologism

Psychologism (a version of what we called Traditionalism):

- Output: the AI produces outputs that make it look intelligent.

- Internal: the AI doesn’t think/isn’t intelligent.

Paper Topic 2

Explain and discuss the following quote from Turing: “May not machines carry out something which ought to be described as thinking but which is very different from what a man does?” If a “machine can be constructed to play the imitation game satisfactorily […] we need not be troubled by this objection.” (1950: 435)

You can discuss this either specifically w.r.t. Turing, or relate it to other ideas. For example, does Turing’s position anticipate today’s ascriptionism? Is Block’s rejection of Turing’s position compelling?

Readings

- Ned Block, “Psychologism and Behaviorism,” *The Philosophical Review* 90.1 (1981): 5-43.

- Andy Clark, “Whatever next? Predictive brains, situated agents, and the future of cognitive science,” *Frontiers in Theoretical and Philosophical Psychology*, CUP 2013.

- Daniel Yon, Cecilia Heyes, Clare Press, “Beliefs and desires in the predictive brain,” *Nature Communications* 2020 [cited as Yon et al]

- Lawrence Shapiro and Shannon Spaulding, “Embodied Cognition,” *The Stanford Encyclopedia of Philosophy* (Summer 2025 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/sum2025/entries/embodied-cognition/>.

- Peter Railton, “Ethical Learning, Natural and Artificial,” in ed. Matthew Liao, *Ethics of Artificial Intelligence*, OUP 2020.