The Predictive Brain

Katja Maria Vogt

ValuesLab

Imagine the Following Conversation…

- W.r.t. AIs, there's often a black-box problem: we don't know how the AI arrives at a particular output.
- In other words: a lack of interpretability and transparency.
- Wait a minute. Isn't this the same in the human brain?
- Early modern philosophers such as Descartes argued that the human mind is introspectively transparent. Today, people tend to disagree.

Imagine the Following Conversation…

- But still, there's a huge difference!
- LLMs are next token predictors, based on a probability calculus.
- That's totally different from the human mind!
- Wait a minute. The human brain is also a next token predictor…
- This is the upshot of an influential framework, called predictive processing. The brain is a probabilistic prediction engine.

Plan for Today

We examine approaches in neuroscience with a view to the question of how our thinking compares to AI:

- Desire Belief Framework
- Predictive Processing Framework

The Desire Belief Framework

A long-standing framework goes back to David Hume (1711-1776).

- There are two kinds of mental attitudes:
  - Desire
  - Belief

The Desire Belief Framework

Perhaps the tradition even goes back to Aristotle (384–322 BCE).

- There are two kinds of activities/faculties in the mind:
  - Desire
  - Reason

The Desire Belief Framework—Direction of Fit

Elizabeth Anscombe is influential in distinguishing between two "directions of fit" (*Intention* 1957):

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

- 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.

The Belief Desire Framework—Advantages and Objections

Advantage: Talk about desires and beliefs is intuitive.

Objection 1: We're not talking at the level of how the brain works.

Response 1: Perhaps OK? We're talking about the mind, not the brain.

The Belief Desire Framework—Advantages and Objections

Objection 2: Is it true that our mental activity is dichotomous?

Response 2: Why give up on the “two directions” premise?

Beyond the Belief Desire Framework

Pursuing Objection 1: What goes on in the human mind?

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)

Default activities cut across the desire-belief divide. Examples:

- 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.

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)

Railton (2020)—The Predictive Brain

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)

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.

> "… 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)

What Follows?

- Railton argues that language-and-ethical learning are probabilistic.
- Does this mean that the mind is, in general, a “prediction engine”?

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)

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

Belief and desire in neuroscience:

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

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

Recall the objection to the Desire Belief Framework that it makes the mind seem dichotomous.

- Is there a framework that explains what's going on the mind in terms of only *one* kind of state?

- Yes! The predictive processing framework.

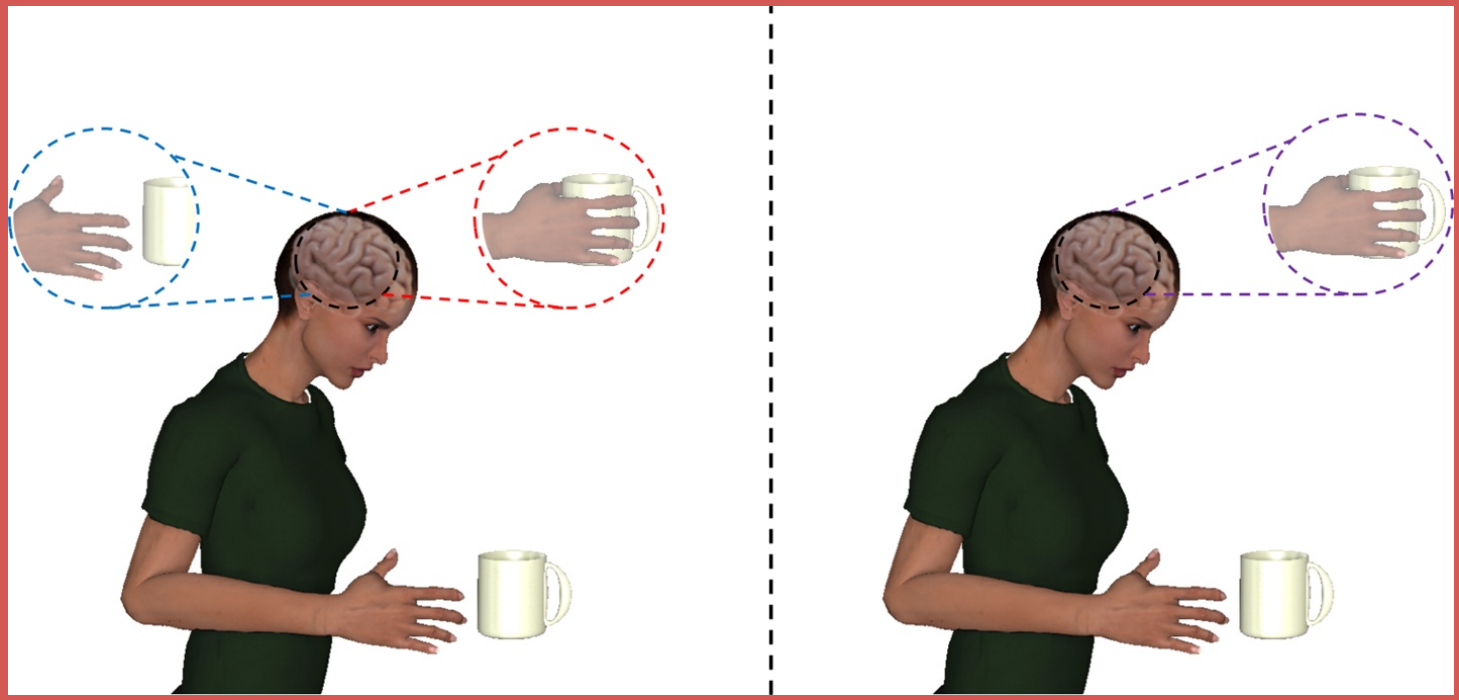
Yon et al, "Beliefs and Desires in the Predictive Brain" (2020)

- 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

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.

Upshots

Suppose:

- The mind is a prediction engine.
- Human beings don't really have desires and beliefs.
- What about the traditionalist argument that we can't ascribe beliefs/ etc to AIs, because these are distinctively human mental states?

Stepping Back—the Mind isn't a Computer!

- The Predictive Processing Framework can seem to conceive of the brain too much as if it were a computer.
- One field that offers push-back: the Embodied Mind framework.

Shapiro and Spaulding, "The Embodied Mind," SEP (2025)

- Computationalism: "mental processes are computational processes; the brain, qua computer, is the seat of cognition."

- The Embodied Mind framework rejects computationalism.

  "… the body or the body's interactions with the environment constitute or contribute to cognition in ways that require a new framework for its investigation. Mental processes are not, or not only, computational processes…"

Do AIs Have Beliefs?

Do AIs have beliefs?

VOTE 3:

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)

Predictive Processing Framework (neither AIs nor humans have beliefs; talk about belief is metaphorical either way)

Readings

- 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.