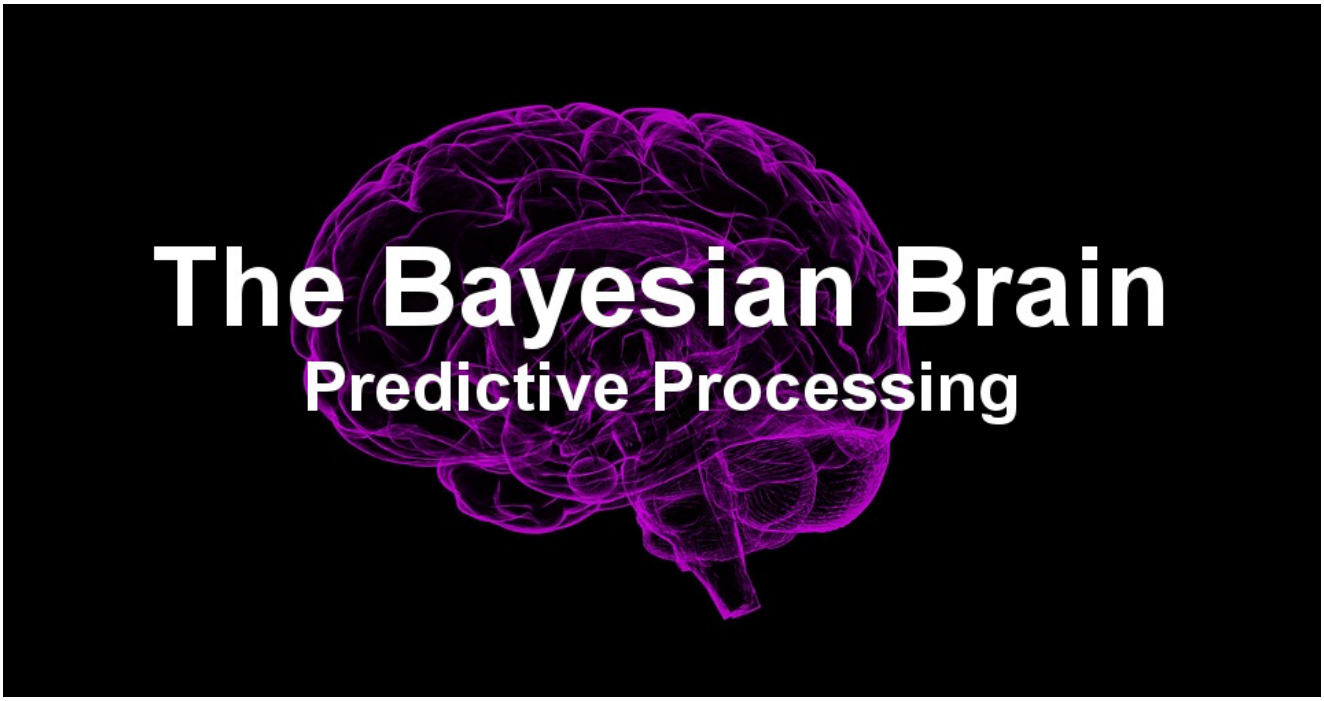
The Bayesian Brain: An Introduction to Predictive Processing

mindcoolness.com/blog/bayesian-brain-predictive-processing

Dom

July 28, 2018



The greatest theory of all time?

The more I learn about the Bayesian brain, the more it seems to me that the theory of predictive processing is about as important for neuroscience as the theory of evolution is for biology, and that Bayes' law is about as important for cognitive science as the Schrödinger equation is for physics.

That is quite an ambitious statement: if our brains really are Bayesian, which is to say that predictive processing is the fundamental principle of cognition, it would mean that *all* our sensing, feeling, thinking, and doing is a matter of making predictions.

In this blog post, I won't discuss the ample evidence we have for this theory,[1] but limit myself to presenting the idea itself. Though it is probably not the greatest theory of all time, it does have the potential to become the greatest theory in the history of cognitive science.

What is predictive processing?

During every moment of your life, your brain gathers statistics to adapt its model of the world, and this model's only job is to generate **predictions**. Your brain is a prediction machine. Just as the heart's main function is to pump blood through the body, so the brain's main function is to make predictions about the body. For example, your brain predicts incoming sensory data: what you're about to perceive from within (interoception) as from without (exteroception).

We used to think that perception is a simple feed-forward process. Say, you open your eyes and sensory data travels from your retina into your brain where it is fed forward through multiple levels of cortical processing, ultimately leading to a motor reaction. If your brain is Bayesian, however, it doesn't process sensory data like that. Instead, it uses **predictive processing** (also known as predictive coding)[2] to *predict* what your eyes will see *before* you get the actual data from the retina.

Your brain runs an internal model of the causal order of world that continually creates predictions about what you expect to perceive. These predictions are then matched with what you actually perceive, and the divergence between *predicted* sensory data and *actual* sensory data yields a **prediction error**. The better a prediction, the better the fit, and the less prediction error propagates up the hierarchy. Wait, what's a hierarchy?

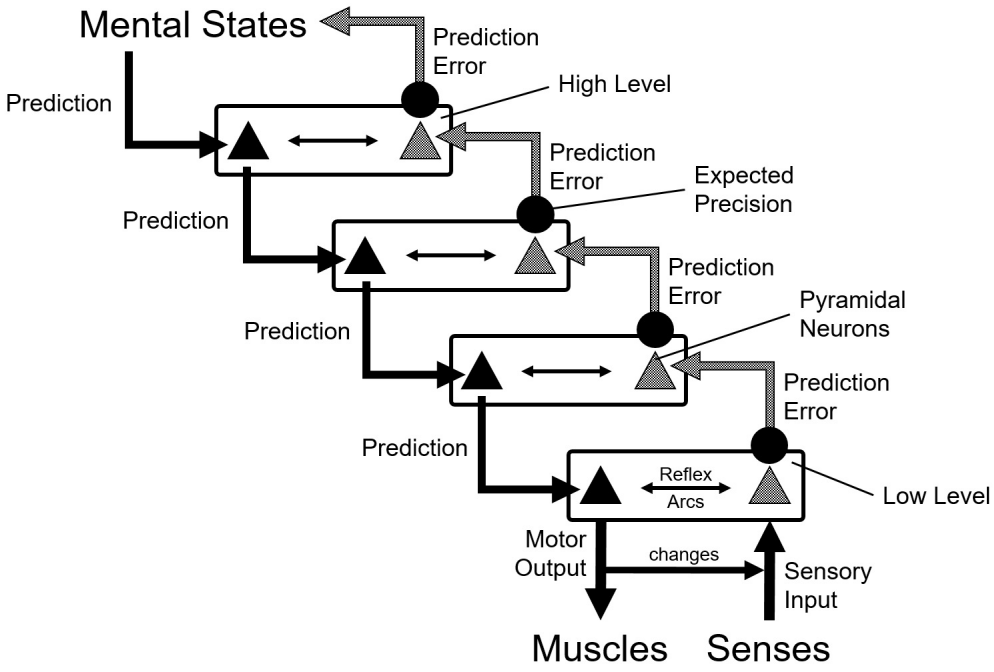
Your internal model (also called *generative model* because it generates predictions) is structured as a **bidirectional hierarchical cascade**:

- The model is a *cascade* because it involves multiple levels of processing, multiple cortical areas in the brain.
- The model is *hierarchical* because it comprises higher and lower processing layers: lower levels process simple data (e.g., sensory stimuli, affective signals, motor commands), higher level process categorizations (e.g., object recognition, emotion classification, action selection), and the highest levels process mental states (e.g., mental imagery, emotion experience, conscious goals, planning, reasoning).[3]
- The model is *bidirectional* because signals continually propagate in both directions: predictions move downward to pyramidal neurons of lower levels while prediction errors move upward to pyramidal neurons of higher levels.

Each processing layer predicts what's happening at the layer below and receives an error signal from that lower layer—a dynamic process that may or may not loop all the way down to incoming sensory data. A model is successful when the predictions it generates at every layer of the hierarchical cascade are accurate, producing only minimal prediction errors.

If your predictions don't fit the actual data, you get a high prediction error that updates your internal model—to reduce further discrepancies between expectation and evidence, between model and reality. Your brain hates unfulfilled expectations, so it structures its

model of the world and motivates action in such a way that more of its predictions come truer. Here's a schematic diagram of how this all works:



Schematic depiction of a bidirectional hierarchical cascade.

The black circles in the diagram (at the prediction error sources) depict something I haven't explained yet, namely **expected precisions**. Precisions determine the weight of prediction errors. If your brain expects that a certain prediction error won't be particularly reliable or important, it decreases its weight and thus the extent to which the error can update your internal model. For example, if it's dark and foggy outside, you can't expect the visual information you receive to be particularly reliable, so you add *that* expectation—in the form of expected precisions that modulate error signals—to your predictions in order to prevent your model from being unduly distorted by unreliable data. In other words, precisions are a measure of uncertainty that expresses your brain's *trust* in the expected sensory data and the ensuing the error signals. The brain wants prediction errors to have a high signal-to-noise ratio.

Expected precisions are formally equivalent to inverse variance and functionally equivalent to **attention**. When you pay a lot of attention to an object, you become confident that the information you get from it is relatively accurate. Say, you've been mindfully looking at a Ferrari under good lighting conditions, so you're pretty confident that it is red. The more attentive you are, the more weight you put on potential error signals. By contrast, if you're not really focused, your brain turns down the "error volume," i.e., the impact a prediction error can have on your model.

What makes the brain Bayesian?

The basic idea of **Bayesian probability** is that you update your beliefs in the light of new evidence. For instance, if you see a dog-like creature running towards you, your belief that it will attack you may be 34%; if you recognize that it's an unmuzzled pit bull, your belief that you'll be attacked may rise to 78%; if it starts barking, 92%; if another barking dog passes you from behind, the probability of your attack-belief being true drops again. That's a vague and incomplete example, but it illustrates how your certainty of a belief may change as you gain new evidence. Initial belief + new evidence = updated belief.

Bayes' law is the mathematical formalization of that idea: $P(B|E)=P(E|B)*P(B)/P(E)$, which calculates the conditional probability of your belief B being true, given evidence E. This is what the brain approximates[4] during predictive processing. In particular, your brain updates its statistical model of the world by integrating prediction errors in accordance with Bayes' theorem; hence the name *Bayesian brain*.[5]

With your model's prediction or *prior probability P(B)* and the lower-level data E within a broader hypothesis space *P(E)*, your brain learns about the *likelihood P(E|B)* of *E*, given hypothesis *B*. Applying Bayes' rule, this yields a *posterior probability P(B|E)*, which determines the prediction error.[6] The error signal, modulated by attention (an expected precision), then propagates back upward to the next higher level where it updates the model, correcting your brain's probabilistic map of reality, and on goes the hypothesis testing. Imagine, for instance, you want to cross a street:[7]

1. Based on a vast set of hypotheses *B* about the current situation, your brain computes a hierarchical cascade of predictions *P(B)*, including
    1. priors about how shapes, colors, and noises will change (for low-level perceptual inference),
    2. priors about you moving your eyes, head, and legs (for low-level active inference),
    3. priors about vehicles in motion and changing traffic lights (for high-level perceptual inference),[8] and
    4. a prior about you standing on the other side of the street (for high-level active inference).
2. Your brain receives sensory data E about what's actually happening on the street and in your body, including
    1. exteroceptive data representing shapes, colors, and noises,
    2. proprioceptive data representing eye, head, and leg movements,
    3. conceptual data representing vehicles and traffic lights, and
    4. agentic data representing your relation to the goal state.
3. Your model contextualizes that data to get *P(E)* and estimates the likelihoods *P(E|B)* of the data, given your hypotheses.

4. Your brain implicitly applies Bayes' theorem to estimate the posterior probabilities *P(B|E)* and determines prediction errors at the relevant processing layers by
    1. matching posteriors and priors regarding shapes, colors, and noises,
    2. matching posteriors and priors regarding your eye, head, and leg movements,
    3. matching posteriors and priors regarding vehicles and traffic lights, and
    4. matching the posterior and the prior regarding your goal state.
5. The prediction errors, weighed by expected precisions, propagate up the hierarchical cascade, level by level, and alter the corresponding set of hypotheses, thus updating the model to reduce future prediction errors; or they are actively reduced through motor commands to the muscles that move your eyes, head, and legs (more about that second option in a second).

After enough cycles of hypothesis testing and model updating during constant sensorimotor interaction with the world, you'll find yourself at the other side of the street (if all goes well). This complex predictive process can be described as your brain continually implementing Bayes' theorem.

What does this explain?

The scope of the Bayesian brain hypothesis is extremely ambitious. It's meant to be a unifying framework for *all* neural, cognitive, and psychological phenomena. If true, predictive processing explains, at a computational level, *everything* about the brain and mind —for reasons we shall see soon. But let's first take a look at how the framework of predictive processing applies to various cognitive phenomena:

1. **Perception** is the prediction of sensory inputs and the inferential[9] process of minimizing prediction error by changing the internal predictive model. This is **perceptual inference**: reducing prediction error by updating your model so that sensory inputs match with prior expectations. Through perception, you make your model more similar to the world.
    1. *Imagination* may be a side-function of the brain's predictive machinery based on high-level predictions that don't travel all the way down the hierarchical cascade and aren't matched with current sensory data, for example, by radically lowering the precision weighting on low-level error signals.
    2. *Delusions* and *hallucinations* may stem from alterations in the ability to integrate incoming data with perceptual predictions (see Griffin & Fletcher 2017).

2. **Action** is proprioceptive[10] prediction and the inferential process of minimizing prediction error by changing sensory inputs. This is **active inference** (also called *predictive control*): reducing prediction error by moving the body so that sensory inputs match with prior expectations. Through motor action, you make the world conform to your model. As the predicted proprioceptive states are not yet actual, actions change the world to make them so (which means that all your actions are essentially self-fulfilling prophecies).
    1. *Reflex arcs* operate at the lowest layer of the active inferential process and work to fulfil proprioceptive predictions until the expected sensory input is obtained. Unlike with perceptual inference, the model parameters are not updated, but kept stable.
    2. *Intentional behavior* is a high-level prediction about an abstract future goal state. Once a concrete opportunity to act arises, it entrains multilevel cascades of lower-level predictions to change the world in a way that makes the high-level prediction come true (i.e., achieve the goal state), thereby reducing long-term prediction error. The achievement of conscious goals is a good example of predictive processing operating at large spatiotemporal scales, for it may take minutes, months, or years to achieve a goal, and it may require one, say, to travel overseas).
3. **Emotion** is interoceptive[11] prediction and the active inferential process of minimizing prediction error by triggering physiological changes (see Seth 2013).
    1. *Drive* is the active-inferential process of minimizing interoceptive prediction error through high-level action. For example, when the brain detects low blood sugar levels through interoceptive inference and the resulting prediction errors aren't resolved through autonomic control (e.g., metabolize body fat), you will feel driven to minimize error signals through voluntary action (e.g., eat sugary food).
    2. *Mental disorder* may be the inability to reduce interoceptive prediction errors.
4. **Attention** is expected precision optimization; it modulates the weight of prediction errors. *Shifts in attention* result from the hyperprior that the same sensorimotor hypotheses (lower-level priors) should not be retained for too long because the world is an ever-changing place that will kill you if you don't adapt (which you can't if you never shift your attention, if you always give the same weight to the errors of related predictions).
5. **Learning** is the updating of your internal model based on prediction errors so that your predictions gradually improve. The better your predictions about the causal, probabilistic structure of the world, the more effectively you can engage with it. This is why teenagers move more fluently through space than toddlers, why you're less clumsy at playing a sport or a musical instrument after having practiced the relevant motor skills, and why it's important to pursue truth—enhanced effectiveness due to better predictions.

6. **Memory** consists of the learned parameters of your internal model, whereas its non-acquired parameters would be the innate knowledge evolution has genetically built into your nervous system. Both parts determine your brain's predictions.
7. **Self-awareness** is the inferential process of minimizing prediction error by changing your internal self-model, i.e., the model that generates predictions about what's most likely to be "you" (see Braun et al. 2018).
    1. *Agency* arises from a good enough fit between your self-model's predictions and exteroceptive input.[12] If the prediction error is too high, you feel like you're not in control of your actions.
    2. *Ownership* arises from a good enough fit between your self-model's predictions and proprioceptive input. If the prediction error is too high, you might experience that one of your limbs doesn't belong to you, or you might have an out-of-body experience.
8. **Belief** is a hyperprior; a systemic prior with a high degree of abstraction; a high-level prediction that entails general knowledge about the world. Some examples:
    1. *Physical beliefs*. You expect things to change over time, you expect heavy objects to fall fast, and you expect that you can't move left and right at the same time.
    2. *Physiological beliefs*. You expect to see something when you open your eyes and you expect fire to hurt and burn your skin.
    3. *Psychological beliefs.* You expect a great performance to make you feel proud and you expect to feel regret when you shy away from a challenge.
    4. *Social beliefs*. You expect happy people to smile, offensive words to trigger a reaction, and power to corrupt.
    5. *Cultural beliefs*. You expect cars to slow down as they approach stop signs, special offers to be highlighted in stores, and handshakes to not last an hour.

Notice how the Bayesian brain also manifests in some of our deepest **values**:

- We value *truth* because accurate beliefs about the world allow us to make good predictions.[13]
- We value *honesty* and *authenticity* because we know what we can expect from honest, authentic people, which improves our prediction making in social contexts that involves them.
- We value *simplicity* because simple beliefs enable us to generate high-level predictions quickly.
- We value *wisdom* because reflected life experience equips us with relatively reliable hyperpriors that minimize long-term prediction error.

At the same time, the predictive nature of human cognition explains why we tend to get enslaved by habits, feel drawn to **comfort zones**, and shy away from uncertainty. Beyond habits, comforts, and certainties, our priors and hyperpriors are less reliable, which leads us to make worse predictions and get higher prediction errors. But high prediction errors are

precisely what the brain works so hard to prevent. The effort of radically updating our models to accommodate novel, unfamiliar situations can be so high that the brain simply decides to trigger feelings of fear, anxiety, or discomfort to motivate us to stick to what we can more reliably predict, i.e., our habitual behavior patterns in comfort zones of soothing regularity.

If we want to leave or expand our comfort zones, we must convince our brains that the immediately resulting high prediction errors are worth it at a broader timescale. Accordingly, we can conceptualize willpower as a hyperprior—namely, as the abstract systemic prediction that, in certain contexts, higher prediction errors in the short term will lead to lower prediction errors in the long term. **Self-control** may thus minimize overall prediction error. But how can the average prediction error over time even matter to a brain that constantly has to deal with immediate errors in any given moment? The answer is to be found in…

The free energy principle

Due to the second law of thermodynamics, the majority of your body's possible states are death and dysfunction. You're lucky that you're alive and able to read this. Sooner or later, **entropy** will catch up with each of us, and we will enter the physiologically unexpected yet physically probable and inevitable state of death. Yet this is precisely what we evolved to avoid! Evolution "designed" us to struggle for survival, to fight chaos, to resist entropy. The human body evolved to maintain itself within expected states that are easily predictable.

According to the **free energy principle**, we achieve such maintenance by suppressing our *free energy*, the information-theoretic equivalent of overall (long-term average) prediction error. Everything we do (and everything any living creature does) is, on average and over time, done to minimize free energy, which corresponds to the brain's job of minimizing prediction error.

And all this to stay within expected, relatively stable states… Why? Because we can survive only within a certain range of physiological states. Too hot and we die, too cold and we die, too much oxygen and we die, too little oxygen and we die, et cetera. Too much or too little of anything will kill us, hence the fundamental physiological principle of **homeostasis**.

To make a long story short, the following principles are hypothesized to be equivalent, though operating at different levels of description:

- cognitive level: minimize prediction error, maximize model fit
- information-theoretic level: suppress free energy,[14] reduce surprisal[15]
- physiological level: maintain homeostasis
- physical level: resist entropy
- biological level: survive

The upshot here is that predictive processing captures what the brain contributes to the body's evolutionary fitness.[16] According to Friston (2009), the free energy principle also explains all structural and functional aspects of the brain, including its anatomy, connectivity, synaptic physiology, electrophysiology, and psychophysiology. I must concede, however, that there's still much about free energy I don't understand, and so much math I haven't gone through yet, that it would be preposterous for me to keep writing as if I actually knew what I'm talking about here.

Five objections

Sensory deprivation

Problem. If all our brains want is low prediction errors, shouldn't the winning strategy be to motivate us to lock ourselves up in a dark, silent room? What better way to make good predictions about expected sensory input than depriving the body of sensory stimulation? In total darkness, we can't get high prediction errors for expecting to see black, black, and black.

Answer. Even though this strategy may be effective (and relaxing) in the short term, the demands of the world and our bodies cannot be avoided for long. The brain, being a bodily organ, never ceases to generate interoceptive prediction cascades, no matter the degree of exteroceptive deprivation. Soon we'll feel bored, hungry, and driven to leave the room again, which will put our bodies in increasingly unexpected states. Locked in a dark, silent room for a long period, we can't maintain homeostasis. In Bayesian terms, we have a hyperprior that occupying the same state for too long will increase long-term prediction error. To always stay within a range of expected states, it's usually more efficient to engage with a stimulus-rich environment (except, of course, when we need to sleep).

Human rationality

Problem. Humans are notoriously bad at probabilistic reasoning, at thinking rationally about probabilities.[17] So if mental processes can barely handle Bayesian inference, how can we assume that the brain, which produces the mind, is Bayesian?

Answer. The Bayesian brain is not inhabited by some homunculus doing complicated math. Its probabilistic inferences are not explicit, but implicit. Brain cells fire in a way that "naturally" approximates Bayes' law, just like ants move in a way that "naturally" approximates Gaussian and Pareto distributions without there being the need for a chief ant strategist who calculates the optimal path for his colony. What does "naturally" mean here? In the case of ants: pheromone secretion. In the case of neurons: Bayesian sampling.

The Bayesian brain is a sampler, not a calculator: it samples information from a local landscape of probabilities (its environment). Over time, action and perception, associated with a practically infinite number of predictions, yield an effectively infinite number of samples from the unimaginably complex probability distributions that constitute reality. This enormous amount of data makes the brain (asymptotically) conform to the laws of probability, thus mechanistically realizing Bayes' theorem. Of course, the brain doesn't store all those samples; it just uses them to update its internal model of the world before it draws new samples from its current local environment.

The human mind, by contrast, having only a finite number of samples, must rely on cognitive shortcuts (heuristics) to generate predictions. Although this has proven evolutionarily adaptive, it now leads us to be cognitively biased and statistically irrational. In *Bayesian Brain without Probabilities*, Sanborn & Chater (2016) explain how sampling produces various reasoning biases. Remember, however, that laboratory studies on cognitive biases are conducted in extremely scarce environments. When people make everyday cognitive judgments in rich, familiar, realistic contexts, their reasoning can be much closer to Bayes optimality than with random judgments in a lab context where sampling is radically limited (see Griffiths & Tenenbaum 2006 and Maguire et al. 2018).

Moreover, consider how the *confirmation bias* (our tendency to focus on information that confirms our preconceptions) follows directly from the Bayesian brain hypothesis. After all, preconceptions are vital for the internal model our brain uses to make predictions. What we notice and focus on is highly determined by our preconceptions, by what we already know. And since predictive processing makes us search for whatever is most likely to verify our predictions, it will generally favor confirmatory evidence, rather than surprising evidence, as a strategy to minimize prediction error.

Offline cognition

Problem. The brain does so much more than perceive the world and command motor actions. Not everything we do is directly related to perception and action. We plan holidays, make financial decisions, reflect on philosophical arguments, and ruminate about events that happened years ago. Such cognitive activities happen "offline," i.e., without immediate real-world interaction. How can predictive processing account for that?

Answer. Offline cognition is only a form of high-level prediction making that doesn't necessarily propagate predictions all the way down to the lowest layers of the hierarchical cascade (or suppresses low-level model updates by radically lowering respective precision weights)—although it might at some point, in which case the timescales are simply much larger than those of "online" cognition, i.e., cognition concerned with immediate sensorimotor interactions in the present environment.

Phenomenology

Problem. The world doesn't look like a complex set of intertwined probability density distributions. It looks rich and colorful and usually unambiguous, often beautiful. So how can we say that cognition is nothing but predictive processing? Is this not overly reductionistic?

Answer. Predictive processing is a theory about how the brain encodes information about the world, not how people experience it. Phenomenology is simply on a higher level of description than cognitive science, similar to how physics could (on an even lower level and to little avail) describe cognitive processes in terms of molecules, atoms, and electrons. A more interesting question would be how much of psychology can be reduced to predictive processing.

Falsifiability

Problem. If predictive processing can explain everything about the brain and mind, does this not expose its triviality? Even more, researchers might retroactively tweak their prior assumptions to make any experimental result fit a Bayesian interpretation *post hoc*. This would mean that the theory effectively explains *nothing*. What keeps the Bayesian brain hypothesis from being trivial and unfalsifiable?

Answer. Evolution, too, is an extremely ambitious theory that aims to explain everything about life and biological systems with only a few basic tools. Naturally, this invites a lot of post-hoc theorizing, especially in the field of evolutionary psychology where just-so stories are often isolated from alternative interpretations. But this doesn't automatically invalidate the theory. All it means is that we must think harder to come up with viable alternative hypotheses and stipulate priors based on independent evidence. "For example," writes Hohwy (2015, p. 14), "there is independent evidence that we expect light to come more or less from above [...], that objects move fairly slowly [...], and that we expect others to look at us." Such kinds of evidence should be integrated with neuroscientific studies.

Predictive processing and the free energy principle can be falsified. If descending prediction signals were found not to carry expected precisions, this would falsify the theory. Or it might turn out that some brain areas are not best described in terms of prediction error signaling. Lastly, the discovery of an organism that doesn't at all act to stay within expected states (to maintain homeostasis) would be a wholesale falsification, though highly unlikely.

Conclusion

Predictive processing provides a framework for understanding all areas of neuroscience and cognitive science at a computational level. Although the Bayesian brain theory is still in its fledgling stage, confirmatory evidence is flowing in on a weekly basis from a vast range of

different fields. And although it offers a highly integrative and ambitious account of the brain and human cognition, it does leave much unspecified. For specification, we must also engage with the paradigms of evolution (as cognition has an evolutionary history), embodied embeddedness (as cognition is embedded in physical environments), and sociocultural situatedness (as cognition is situated in social and cultural contexts). Only if we cover and combine all levels of analysis can we truly understand how humans work.

Read more

- Philosophy and Predictive Processing
- How the Brain Makes Emotions
- The Bayesian Brain: Placebo Effects Explained

References

Braun N, Debener S, Spychala N, Bongartz E, Sörös P, Müller HHO, Philipsen A (2018). The Senses of Agency and Ownership: A Review, *Frontiers in Psychology*, Vol. 9(535).

Buckley CL, Kim CS, McGregor S, Seth AK (2017). The free energy principle for action and perception: A mathematical review, *Journal of Mathematical Psychology*, Vol. 81, pp. 55-79.

Clark A (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science, *Behavioral and Brain Sciences*, Vol. 36(3), pp. 181-204.

Friston KJ, Stephan Ke (2007). Free-energy and the brain, *Synthese*, Vol. 159(3), pp. 417-458.

Friston KJ (2009). The free-energy principle: a rough guide to the brain?, *Trends in Cognitive Sciences*, Vol. 13(7), pp. 293-301.

Griffin JD, Fletcher PC (2017). Predictive Processing, Source Monitoring, and Psychosis, *Annual Review of Clinical Psychology*, Vol. 13, pp. 265-289.

Griffiths TL, Tenenbaum JB (2006). Optimal predictions in everyday cognition, *Psychological Science*, Vol. 17(9), pp. 767-773.

Harkness DL & Keshava A (2017). Moving from the What to the How and Where – Bayesian Models and Predictive Processing, *Philosophy and Predictive Processing*, 16.

Hohwy J (2015). The Neural Organ Explains the Mind, *Open MIND* by Metzinger T & Windt JM (Eds.), 19(T).

Maguire P, Moser P, Maguire R, Keane MT (2018). Why the Conjunction Effect Is Rarely a Fallacy: How Learning Influences Uncertainty and the Conjunction Rule, *Frontiers in Psychology*, Vol. 9(1011).

Metzinger T, Wiese W (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing, *Philosophy and Predictive Processing*, 1.

Sanborn AN, Chater N (2016). Bayesian Brains without Probabilities, *Trends in Cognitive Science*, Vol. 20(12), pp. 883-893.

Seth AK (2013). Interoceptive inference, emotion, and the embodied self, *Trends in Cognitive Science*, Vol. 17(11), pp. 565-573.

Thornton C (2016). Predictive processing simplified: The infotropic machine, *Brain and Cognition*, Vol. 112, pp. 13-24.