Trends in Cognitive Sciences

CellPress
REVIEWS

Letter

Optimism and Pessimism in the Predictive Brain

Zekun Sun[1] and
Chaz Firestone[1,*]

Why don't prediction-error minimizers waste away in maximally predictable—but maximally boring—environments? Van de Cruys, Friston, and Clark [1] offer an elegant and even poetic answer: a bias toward 'optimism'.

The idea, echoed by Seth *et al*. [2] and a broader literature, is that agents tend to predict favorable outcomes: full bellies, stable blood-glucose levels, adequate hydration, and so on. When those predictions aren't met, prediction-error accumulates, which can be reduced by acting to *make* the predictions come true (so-called 'active inference'). For example, if you optimistically predict that you won't be hungry, but then you find yourself not having eaten in a while, you can reduce the ensuing prediction-error by taking matters into your own hands and eating. Optimistic predictions thus act like self-fulfilling prophecies, leading us toward flourishing states: after enough time in a dark, empty room, these predictions become falser and falser until there is no choice but to leave and fulfill them.

Does this response succeed? We're initially tempted to play some thought-experiment tennis and simply modify the original scenario. Instead of an empty room, we could add an IV drip with just the right electrolyte balance, a thermostat to tailor the temperature, and so on for other bodily functions. (Call it the 'Homeostatic Room'.) Since 'our expected states are determined by what it takes to maintain homeostasis' [3], this arrangement should be paradise for a surprise-minimizer. Yet, it seems unlikely that you'd stay.

But perhaps optimism extends even further than homeostatic states. If agents are optimistic not only about their blood-glucose and hydration levels but also which friends they will see, which hikes they will take, and more, then surprise-minimizers might leave even the Homeostatic Room and become well-adjusted members of society.

A deeper problem is lurking here. Predictive Processing (PP) considers beliefs and desires to be 'an elusive and degenerate duality' [1], and so aims to replace them with a single state: prediction. But appealing to optimism works against this aim, because what counts as 'optimistic' *depends* on one's desires. To be optimistic is, roughly, to predict that what you want to happen is what will actually happen—to believe that events will unfold in the way you desire. Bountiful feasts and beautiful hikes only get to be optimistic predictions because the relevent agents find those outcomes desirable. That's why optimism looks different for different people: an optimistic ascetic might predict a day of solitary meditation rather than a feast with friends; different outcomes count as optimistic for her, because she has different underlying desires. But no desires beyond 'minimize surprise' are permitted in PP's psychology; that, we thought, was the (central and radical) point. In that case, the only way optimism gets agents out of Dark Rooms and living their lives is by smuggling in desires after all; indeed, all the same desires we already knew the agents to have. (Nor will it do to define optimism relative to evolutionary considerations. We are not prisoners of our evolutionary drives, as the ascetic's case shows.) What work, then, has PP done?

PP as a 'theory of everything' is exciting because of the simplicity, power, and reach it advertises. It promises not only a tool to model some aspects of reward learning or visual processing [4], but also a worldview to 'radically reconceptualize who we are' [5]—'a single principle by which neural operations can account for perception, cognition, action, and even consciousness' [6]. If this single principle requires an optimistic bias for any behavior that is not transparently surprise-minimizing, it stands to lose this special appeal. The behaviorism analogy continues to be apt: when an initially elegant view—reinforcement all the way down—needs a supplementary list of self-reinforcers for every otherwise-unexplained behavior, it risks becoming false (are all those biases really in there?) or trivially true (Box 1).

Still, we acknowledge that this feeling is not universal; indeed, the replies raise similar objections against belief-desire psychology. But self-fulfilling predictions present additional difficulties.

The 'Pessimistic Prediction' Problem

We often make predictions that we hope won't come true. Suppose you are headed for a night on the town and, knowing yourself, you predict you will drink too much. Given this pessimism, you may try to *prevent* this prediction from being realized—e.g., by asking friends to monitor you, by bringing less cash, etc. A traditional account of such behaviors would invoke beliefs and desires: you believe you will overdrink, but you desire not to.

What about PP? Recall how prediction-error minimization and active inference jointly transform our predictions into self-fulfilling prophecies: your prediction that you will eat *compels* you to eat, because your failure to eat increases prediction-error that you must then act to minimize. While this mechanism works virtuously for optimistic predictions, it seems to invite disaster for pessimistic ones: the same mechanism should trigger a vicious feedback loop, whereby your overdrinking-prediction fulfills itself by making you drink all the more. Worse yet, your friends, anticipating your overdrinking, should now

Box 1. Is Predictive Processing Falsifiable?

Is the Predictive Processing (PP)/Free Energy Principle (FEP) view falsifiable? A common response is that the very question rests on a kind of conceptual mistake: 'FEP is a framework, not a testable hypothesis' [2]; "the notion that a 'framework' can have the attribute 'falsifiable' is a category error… falsifiable hypotheses are a hangover from classical inference" [7].

This strikes us as a little too clever. That PP/FEP is a framework, rather than a testable hypothesis, may well be true; but this does little to defuse broader concerns about its reach. One could simply reframe the question: for *which* mental phenomena is PP/FEP a useful and explanatory framework? Recall how all-encompassing the framework is meant to be: 'The free energy principle is extremely ambitious: it aims to explain *everything* about the mind' [3]. But as Seth *et al*. [2] and Clark [8] concede, much of what underlies human flourishing—art, adventure, charity, and more—strains the framework and its principles.

The analogy to behaviorism is helpful yet again. Reinforcement, too, is not a 'hypothesis' that can be 'falsified'. What can be falsified is the thesis that all of who and what we are is captured by that notion. So too for PP/FEP, hangovers aside.

What is the Motivation for PP?

We are nowhere near the first to discuss the Dark Room Problem, nor are we the first to worry that PP may be unfalsifiable [9]. Indeed, there are recent and powerful arguments that the most ambitious flavors of PP are not only falsifiable, but also *false* as theories of our own minds and brains [10,11]. It is notable, then, that despite claims by both replies that PP is 'testable' [1,2], they give no indication of what should strengthen or weaken our confidence in it. What should make us accept or reject PP?

The story of behaviorism again offers a guide. For all of behaviorism's troubles, Skinner had principled reasons for explaining behavior in terms of reinforcement: He believed that notions of internal mental states were unscientific, and feared 'the specter of teleology' raised by talk of intentions. Though the cognitive revolution showed how mental states can cause behaviors through the mechanisms of computation, we can still look back and appreciate that behaviorism was motivated (at least, until an alternative came along). If PP/FEP is merely a framework that redescribes psychological phenomena in predictive terms—rather than a falsifiable theory—then what is the motivation for embracing it? There are, in principle, many unfalsifiable ways to redescribe our minds; why this one?

minimize their prediction-error, perhaps encouraging you more than if you hadn't warned them. But this just isn't what happens: clinical self-sabotage notwithstanding, it is perfectly possible to make a prediction and then try to become *wrong* about it. More generally, pessimism rarely motivates self-fulfillment: anticipating a stock-market crash does not motivate you to bring the crash about; believing your favored politician will lose does not make you canvas for her opponent, etc.

Perhaps PP could reply by suggesting that you don't know your own predictions, even when they are made explicitly: you may think you predict the worst, but really you don't. Or maybe your prediction is disjunctive, even if it may not feel that way: "I'll overdrink, or I'll prevent it like so". But such replies introduce new costs of their own—such as an implausibly extreme self-ignorance, or a difficulty constraining and selecting among disjuncts—or they may not actually accommodate the full spectrum of cases (e.g., the stock market and election cases above, and pessimistic inaction more generally). What worked for optimism, then, seems to fail for pessimism.

Poetry, Rollercoasters, and Bucket-Lists

Optimistic predictions are not the only proposed solution to the Dark Room Problem. Seth *et al*. [2] offer another: leaving the room *does* increase prediction-error, but only over the short-term, because more surprise now enables better predictions later.

Our piece discussed this response: if Dark Rooms are only local minima within a broader predictive landscape, then prediction-error minimization can indeed recommend leaving, even without optimistic biases. Still, this reply, like PP itself, leaves such an approach worryingly underconstrained. Over what timeframe do minds like ours minimize expected free energy? (A day? A lifetime? The arc of our species?) Seth *et al*. don't say, invoking only 'long temporal horizons', 'extended sequences', and 'the future'—vague and flexible notions that could bend to accommodate nearly any result.

More importantly, leaving the room is only a first step. Humans are, or appear to be, motivated by so much more than error-minimization: we play music, help strangers, read poetry, have children, build churches, climb mountains, and smell flowers. We do so even when these experiences are overly familiar and so of little predictive benefit (e.g., replaying a favorite song, or rereading a cherished poem). And we do so even when they are unlikely to further any long-term predictive agenda (e.g., the terminally ill patient who finally takes that exciting bucket-list trip). Can it really be that these activities, so essential to human flourishing, arise for their long-term predictive utility?

Seth *et al*. sometimes seem to admit not: 'will this approach explain rollercoaster-riding and poetry-reading? In the details, perhaps not'. But they elsewhere dismiss such behaviors as 'rare', instead emphasizing how PP explains 'epistemic actions such as eye movements', or even 'an agent's beliefs about the world' (a reference to their model of *E. coli* chemotaxis). We worry that this has things backwards. Saccadic planning and flagellar control are surely important processes; indeed, we study (one of) them too [12]. But they are just not the critical phenomena under discussion. Here we are moved by Clark, who elsewhere acknowledges that humans' drive for self-actualization presents 'the most genuinely challenging incarnation of the Darkened Room worry' [8]. That seems exactly right. Personal growth, aesthetic

experience, moral worth, and so much else about us are not rare or peculiar distractions—they anchor the full and meaningful lives we seek. Perhaps, then, one's feelings about PP as a 'theory of everything' will turn on one's feelings about humanity itself. This would be a credit to PP: any theory that motivates such deep and enduring questions is one worth our attention.

Acknowledgments

For helpful discussion and/or comments on earlier drafts, we thank Ned Block, Steven Gross, Ben Hayden, Colin Klein, Ian Phillips, Dan Williams, and members of the Perception and Mind Laboratory.

[1]Department of Psychological and Brain Sciences, Johns Hopkins University, 3400 N Charles St, Baltimore, MD 21218, USA

*Correspondence:
chaz@jhu.edu (C. Firestone).

https://doi.org/10.1016/j.tics.2020.06.001

References

1. Van de Cruys, S. *et al.* Controlled optimism: reply to Sun and Firestone on the Dark Room Problem. *Trends Cogn. Sci.* (in press)
2. Seth, A.K. *et al.* Curious inferences: reply to Sun and Firestone on the Dark Room Problem. *Trends Cogn. Sci.* (in press)
3. Hohwy, J. (2015) The neural organ explains the mind. In *Open MIND* (Metzinger, T.K. and Windt, J.M., eds), pp. 1–23, MIND Group
4. Rao, R.P. and Ballard, D.H. (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87
5. Hohwy, J. (2013) *The Predictive Mind*, Oxford University Press
6. Seth, A.K. (2015) The cybernetic Bayesian brain. In *Open MIND* (Metzinger, T.K. and Windt, J.M., eds), pp. 1–24, MIND Group
7. Friston, K. *et al.* (2018) Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston. *ALIUS Bull.* 2, 17–43
8. Clark, A. (2018) A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenol. Cogn. Sci.* 17, 521–534
9. Kogo, N. and Trengove, C. (2015) Is predictive coding theory articulated enough to be testable? *Front. Comput. Neurosci.* 9, 2960
10. Kwisthout, J. and van Rooij, I. (2019) Computational resource demands of a predictive Bayesian brain. *Comput. Brain Behav.* 3, 174–188
11. Williams, D. (2018) Predictive coding and thought. *Synthese* 76, 695–727
12. Sun, Z.K. *et al.* (2016) Experimental pain induces attentional bias that is modified by enhanced motivation: An eye tracking study. *Eur. J. Pain* 20, 1266–1277