References

1. Gross, J.J. and John, O.P. (2003) Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *J. Pers. Soc. Psychol.* 85, 348–362
2. Buhle, J.T. *et al.* (2013) Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. *Cereb. Cortex* 24, 2981–2990
3. Burr, D.A. *et al.* Emotion dynamics across adulthood in everyday life: older adults are more emotionally stable and better at regulating desires. Emotion (in press).
4. Goldin, P.R. *et al.* (2008) The neural bases of emotion regulation: reappraisal and suppression of negative emotion. *Biol. Psychiatry* 63, 577–586
5. Shiota, M.N. and Levenson, R.W. (2009) Effects of aging on experimentally instructed detached reappraisal, positive reappraisal, and emotional behavior suppression. *Psychol. Aging* 24, 890–900
6. Winecoff, A. *et al.* (2011) Cognitive and neural contributors to emotion regulation in aging. *Soc. Cogn. Affect. Neurosci.* 6, 165–176
7. Blanke, E.S. *et al.* (2019) Mix it to fix it: emotion regulation variability in daily life. *Emotion* Published online February 4, 2019. https://doi.org/10.1037/emo0000566
8. Grommisch, G. *et al.* (2019) Modeling individual differences in emotion regulation repertoire in daily life with multilevel latent profile analysis. *Emotion* Published online September 2, 2019. https://doi.org/10.1037/emo0000669
9. Doré, B.P. *et al.* (2016) Toward a personalized science of emotion regulation. *Soc. Personal. Psychol. Compass* 10, 171–187

Forum

The Dark Room Problem

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

Predictive Processing theories hold that the mind's core aim is to minimize prediction-error about its experiences. But prediction-error minimization can be 'hacked', by placing oneself in highly predictable environments where nothing happens. Recent philosophical work suggests that this is a surprisingly serious challenge, highlighting the obstacles facing 'theories-of-everything' in psychology.

A dark, empty room presents few surprises. The information reaching the eyes is constant, uniform, and unremarkable; effective soundproofing could do the same for the ears. Add some creative seating, and the whole experience will be as dull and predictable as any experience could be.

Humans and other animals tend not to seek such experiences, and even find them aversive when endured for long enough: if several hours in such a room would be dreadfully boring, days or weeks would be unbearable. But according to a sweeping account of cognitive and neural functioning, this seemingly should not be the case. Predictive Processing theories (PP) hold that the mind's core aim is to 'minimize prediction-error' about its experiences – to be as least wrong as possible about what is happening [1–3]. This single principle is invoked to explain a vast array of behaviors and capacities, including attention, learning, memory, action, emotion, motivation, and more – a psychological theory-of-everything to 'unify these very diverse aspects of our mental lives under one principle' [1]. Yet, at first glance, PP seems committed to a bizarre hypothesis: that prediction-error minimizers – us, allegedly – should find their deepest motivations fulfilled by the most utterly boring experiences, since a sure way to minimize prediction-error is just to place oneself in a highly predictable environment (such as a dark, empty room where nothing much happens).

That prediction-error minimization might be short-circuited in this way is now known as the 'Dark Room Problem' [4]. Though it may sound fantastical, recent work in philosophy of cognitive science has amplified this challenge and highlighted its seriousness [5–7]. Why, according to PP, should anyone do anything other than idle in a predictable room? Here, we briefly review some answers to this question. We give special attention to one solution that recalls our field's oldest and most foundational disputes over all-encompassing theories of the mind and brain.

Some Intuitive Replies

Is the Dark Room Problem really a problem? You might think not. For example, enough time in the room would surely make you hungry or thirsty; wouldn't you leave to satisfy such needs?

Indeed, you would; but this observation only refocuses the original problem. If agents aim only to minimize prediction-error, then states should be avoided only insofar as they increase prediction-error. However, for someone idling in a dark room, hunger is highly predictable. As Klein notes [5], 'predicting hunger is not the same as being motivated by it. As I lay with my eyes shut, my cognitive system could predict perfectly well the progression of hunger signals. (It is not that complicated: I will get more and more hungry, and then die)'. The challenge posed by the Dark Room Problem is not to say why someone would leave; it is to say why prediction-error minimization should make someone leave – and it is not clear that it does, even to eat.

A less dismissible reply might invoke curiosity and exploration. Even if leaving a dark room increases short-term prediction error, perhaps doing so could reduce long-term prediction-error (e.g., if exploring the outside world can further hone your prediction skills). However, even this intuition underestimates the Dark Room Problem's insight. As Clark acknowledges [7], not all motivations that drive us from dark rooms reduce to instrumentally valuable exploration, even over the long-term. Humans are endlessly creative: we dance, ride rollercoasters, donate to charity, and read poetry; we even seek surprise itself in certain aesthetic pursuits, purposefully entering unpredictable states just for the thrill of it. In its most ambitious flavors, PP aims to explain every psychological state we have: 'perception and action and everything mental in between' [1]. However, even if some behaviors reduce long-term prediction-error, it is not clear that all behaviors serve that purpose, and so they seem unexplained by PP.

Predict Yourself

What replies remain? Though there are still others, one reply in particular stands out both for its ingenuity and its precarity. Friston [2] suggests that the Dark Room Problem is off-track from the start, because it mistakenly assumes that prediction-error in dark rooms is low; instead, 'the state of a room being dark is surprising, because we do not expect to occupy dark rooms'. In other words, agents have predictions not only about the world, but also about themselves – including perhaps, 'I don't linger in dark rooms'. In that case, occupying a dark room generates high prediction-error after all, and so exiting the room reduces it (by fulfilling the prediction that you would leave). Indeed, such predictions could be innate, and even 'stubborn': hypotheses that are 'resistant to evidence-based updating' [8].

Unlike others, this reply really does 'solve' the Dark Room Problem, at least in allowing prediction-error minimization to recommend leaving. But might its local success expose a more global risk for PP's broader project? Our purpose here is not quite to decide that question. Instead, whether or not this reply ultimately succeeds (Box 1), we think it raises one of the deepest issues a theory can face in the first place – and one our field has confronted before.

Box 1. The Collapse of Belief and Desire under PP

How does prediction lead to action? Traditionally, actions are explained by pairs of states: beliefs and desires. Suppose you drink some water to quench your thirst: Neither your desire alone ('I want to quench my thirst') nor your belief alone ('Drinking water will quench my thirst') explains your drinking; but the two states, together, do.

PP, however, proposes one state – prediction – for this role. We predict that we will drink water, and then we alter our environment to make the prediction true. As Clark puts it [11]:

> My desire to drink a glass of water now is cast as a prediction that I am drinking a glass of water now – a prediction that will yield streams of error signals that may be resolved by bringing the drinking about, thus making the world conform to my prediction.

However, this scenario poses a puzzle. Mismatches between predictions and the environment can always be resolved in two different ways: (i) changing one's environment to match the prediction (e.g., drinking), or (ii) changing one's prediction to match the environment (e.g., predicting that you will not drink after all). Since each resolves the prediction-environment mismatch, the puzzle is why thirsty prediction-error minimizers should choose (i) over (ii). Indeed, updating predictions may well be easier than finding water and drinking it; so why act at all?

Unrevisable Predictions?

One solution could be that some predictions are simply unrevisable: 'hypotheses that evidence cannot change' [8]. What kind of unrevisable prediction could help here? 'I won't die of thirst' is not quite specific enough; 'don't die' is not exactly a policy one can follow. But perhaps 'I drink water when I'm thirsty' could work. If 'I drink water when I'm thirsty' is unrevisable, then choice (ii) is out; drinking is the only option.

However, Klein [5,12] offers powerful reasons to doubt that such predictions could be unrevisable in the right way for PP. Suppose, for example, that you learn that the water around you is unsafe to drink, or that today is a religious fast, or that you are staging a hunger strike. People in such circumstances forego drinking – precisely, it seems, by revising their predictions about their behavior when thirsty.

Moreover, adding caveats ('I drink water when I'm thirsty, unless the water is unsafe, or I'm fasting, or…') may not help. First, it would be unclear how the caveats get there to begin with; after all, the predictions are meant to be unrevisable (and there is no innate knowledge of fasting rituals). Second, such ungainly predictions risk losing their status as explanations altogether. After enough additions, they become mere post-hoc descriptions of how we act — not the mechanism by which we act.

Self-Prediction vs Self-Reinforcement

Over a half-century ago, B.F. Skinner and Noam Chomsky sparred over whether a different core principle – reinforcement – could account for all of human behavior. Among many arguments Chomsky raised against behaviorism, an underappreciated one was to catalog ordinary activities that seem not to arise from reinforcement. Children and adults, Chomsky noted, do things like talk to themselves when nobody is around, make music in private, or imitate the sounds of cars and airplanes – none of which is typically a 'rewarded' behavior. So why, on behaviorism, do we do such things? Skinner's answer was 'self-reinforcement': we talk to ourselves because it feels rewarding to do so, such that we are the reinforcers of our own behavior.

Chomsky replied, rightly, that appeals to self-reinforcement actually undermine behaviorist explanations, because they are either (i) false (is talking to oneself really 'rewarding'?), or (ii) trivially true – a panacea that could explain any behavior imaginable. And mechanisms that can explain anything ultimately explain nothing, because they become empty or unfalsifiable: 'When we read that a person plays what music he likes, says what he likes, thinks what he likes, reads what books he likes, etc., because he finds it reinforcing…the term "reinforcement" has no explanatory force' [9].

We worry that 'self-prediction' shares this property with self-reinforcement, and so risks a similar dilemma for PP. Either the self-prediction account is: (i) false (do we really predict our own room-lingering tendencies?), or (ii) trivially true, accommodating any possible behavior. *Why do we dance?* Because we predict we won't stay still. *Why do we donate to charity?* Because we predict we will do good deeds. *Why do we seek others?* Because 'the brain has a prior which says "brains don't like to be alone"' [10]. In some moods, these answers will land as deep and profound truths about the mind; but in others they will simply be nonexplanations that lead one to repeat one's question. If leaving a dark room is explained by predictions that we will not linger in dark rooms, then no behavior could be inconsistent with PP; anything we do could be explained by predictions that we wouldn't not do it.

PP has made valuable and lasting contributions to our understanding of cognition, and the Dark Room Problem will not undo this progress. But history shows how psychological theories-of-everything can be undone precisely by their totalizing ambitions. However the Dark Room Problem is overcome, PP must take care to avoid the same fate.

Acknowledgments

For helpful discussion and/or comments on earlier drafts, the authors thank Jonathan Flombaum, Jorge Morales, Ian Phillips, Brad Wyble, and members of the Johns Hopkins Perception & Mind Laboratory.

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

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

https://doi.org/10.1016/j.tics.2020.02.006

References

1. Hohwy, J. (2013) *The Predictive Mind*, Oxford University Press, Oxford
2. Friston, K. (2013) Active inference and free energy. *Behav. Brain Sci.* 36, 212–213
3. Clark, A. (2016) *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford University Press, New York
4. Friston, K. *et al.* (2012) Free-energy minimization and the dark-room problem. *Front. Psychol.* 3, 130
5. Klein, C. (2018) What do predictive coders want? *Synthese* 195, 2541–2557
6. Sims, A. (2016) A problem of scope for the free energy principle as a theory of cognition. *Philos. Psychol.* 29, 967–980
7. Clark, A. (2018) A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenol. Cogn. Sci.* 17, 521–534
8. Yon, D. *et al.* (2019) The predictive brain as a stubborn scientist. *Trends Cogn. Sci.* 23, 6–8
9. Chomsky, N. (1959) Review of *Verbal Behavior* by B.F. Skinner. *Language* 35, 26–58
10. Allen, M. and Friston, K.J. (2018) From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese* 195, 2459–2482
11. Clark, A. (2017) Predictions, precision, and agentive attention. *Conscious. Cogn.* 56, 115–119
12. Klein, C. A Humean challenge to prediction coding. In *The Science and Philosophy of Predictive Processing* (Gouveia, S., Mendonça, D. and Curado, M., eds), Bloomsbury Press (in press)