Controlled Optimism: Reply to Sun and Firestone on the Dark Room Problem

Sander Van de Cruys[1*],

Karl J. Friston[2],

Andy Clark[3]

[1]Laboratory of experimental psychology, University of Leuven (KU Leuven)

[2]Wellcome Centre for Human Neuroimaging, University College London

[3]University of Sussex and Macquarie University

*Correspondence: sander.vandecruys@kuleuven.be (S. Van de Cruys).

Keywords: motivation, predictive processing, dark room problem

Word count: main text: 867; box: 199

Sun & Firestone [1] argue that the Dark Room Problem poses an important challenge to the ambitions of predictive processing accounts – specifically, they worry that a standard response threatens the story with triviality, asserting merely that prediction-driven agents avoid dark, food-free corners because they 'predict that they will not stay in them'.

In response, we wish to highlight the principled role of 'optimistic predictions'. Our predictive models can [must] be optimistically biased, in that the distribution of expected states is realised, when we act upon the world. For example, the model may include interoceptive expectations on adequate glucose levels throughout a famine. These allostatic predictions are deep-set, ingrained in low-level structural mechanisms (and underwrite foraging for food). But optimistic predictions occur at higher-levels and in more flexible ways too. For example, placebo response flows from confident expectations of relief, whose effects reach all the way down to spinal cord responses [2]. Crucially, such effects vary when experimenters manipulate subjects' confidence, altering the precision with which they predict relief [3]. Yet expect too much and rapid disappointment (and downgrading of precision over future expectations) follows. Effective predictive processing agents must form *optimistic yet sufficiently realistic* expectations about their own future states and behaviors. Technically, posterior beliefs are the optimal mixture of prior optimism and sensory evidence.

As James observed [4], the evidence as to whether a belief is true often only becomes accessible after first adopting the belief without evidence. One empirical consequence of this is the Dunning-Kruger effect, the finding that most people consider their competence on a complex task as above average. Clinical depression seems to be associated with both greater realism and diminished motivation [5] suggesting an adaptively valuable trade-off between epistemic accuracy and optimistic prediction. While not classically rational, it is the optimistic slant [6] that leads us, as aspirational – and curious – beings, out of the dark corners of the world. This is not simply because we think we are curious – we are curious because we think; i.e., form beliefs about the epistemic and conative consequences of our actions.

The prediction error minimization principle is simultaneously epistemic and conative (motivational). Indeed, there is no essential difference between goals or desires and beliefs or predictions on this account [7]. Sun and Firestone (following Klein [8]) worry that this yields a new puzzle, about when to resolve error by altering the world and when to resolve error by altering expectations. This is largely solved by the special poise of precise proprioceptive predictions to engage bodily action [7]. More generally, expectations about when to resolve error by update versus action themselves form part of the generative model.

How explanatory is PP? Suppose we ask, using Sun & Firestone's example, "Why do we donate to charity?". Classical psychological explanations would resort to

positing different kinds of goals here (e.g. maintaining a good reputation). Though useful in daily life, such explanations also tend towards circularity, and fail to resolve the hard question of teleology. ("Why do we do charitable things? Because we want to act charitably."). But PP has the power to go further, to 'dissect teleology'. This is because the inferential tools of PP provide a (Bayesian) method for learning our goals and preferences, and for learning when to be driven to action by optimistic predictions and when to update our beliefs instead.

Both the idiosyncrasies in—and the contextual fluidity of—our goals and choices are often overlooked in classical motivation theories. These tend to assume a set of intuitively plausible, allegedly fixed and universal goals (such as autonomy, status, etc). However, few of our actual goals are absolute. Goals, like expectations, vary in their imperviousness to evidence. Their direction of fit, whether mind-to-world (i.e belief-like) or world-to-mind (i.e. goal-like), is a constant negotiation, rather than set in stone. This negotiation crucially depends on estimated confidence (*precisions*) in the attainability of expected states using one's actions (*control states* in PP speak) [see Box 1]. For example, a child that sees herself as a math whiz (goal), who discovers they are not that good at math, updates her prior expectations, making something goal-like into something belief-like. Often, we infer who we are—our values and goals—from our (inter)actions, decisions, and resulting observations, rather than the other way around [9]. The point here is not to belittle the importance of goals in driving

behavior, but rather to emphasize the bidirectional influences needed to really explain behavior.

Box 1: Creating and dissolving goals by precision tuning

How do goals and motivations emerge under predictive processing? This translates directly into asking how certain predictions accrue high precision. Some deep-set homeostatic predictions come with innate, high precision. Often precision is accrued (through Bayesian learning) by repeating an action, as for habits. Other times beliefs take on precision by social learning, as when we come to adopt the goals and expectations from parents or peers. There are no fixed hierarchies here: In the case of the hunger striker or the martyr, social or ideological expectations have been conferred (temporary) higher precision than interoceptive ones. Finally, goal adoption is facilitated by positive rates of error minimisation (doing better than expected at minimising error), and expectations of positive slopes of error reduction act as yet another force causing us to prefer rich environments over darkened rooms [10]. Overall, PP equips agents with an initially unwarranted level of confidence (precision) on the attainability of favorable outcomes, necessary to disclose the very conditions of their (fallible) realization. Computationally, this is expressed in the view of planning-as-inference, which firmly assumes the observation (fulfillment) of the desired outcomes, and from there infers which behaviors (policies) would 'generate' those favorable outcomes with the greatest certainty [11].

The PP story makes highly testable claims. Once we define an organism and niche, designating initial bodily (including neural) structure and capabilities, and thereby seeding them with initial expectations about the kinds of state they should inhabit—given their viability conditions—everything else must follow from the attempt to minimize long-term prediction error. At that point, the story delivers specific predictions, including predictions about how they will act, how they will resolve apparent belief/desire conflicts, and how they will update their expectations given new sensory evidence. It is traditional accounts that then seem underconstrained and hard to falsify, since they posit a elusive and degenerate duality, in the form of beliefs and desires—interacting in some unspecified way with bedrock systemic structure.

Acknowledgments

AC is supported by an ERC Advanced Grant (XSPECT—DLV—692739), and KJF by a Wellcome Principal Research Fellowship (Ref: 088130/Z/09/Z). SVdC is supported by the Methusalem program by the Flemish Government (METH/14/02), awarded to Johan Wagemans.

References

1 Sun, Z. and Firestone, C. (2020) The Dark Room Problem. *Trends Cogn. Sci.* DOI: 10.1016/j.tics.2020.02.006

2 Eippert, F., Finsterbusch, J., Bingel, U., and Büchel, C (2009) Direct Evidence for

Spinal Cord Involvement in Placebo Analgesia *Science* 1 : 404

3 Büchel, C., Geuter, S., Sprenger, C., & Eippert, F., et al. Placebo analgesia: A predictive coding perspective. *Neuron*, 81(6): 2014 p. 1223–1239

4 James, W. (1896) *The will to believe and other essays in popular philosophy*, Longmans, Green, and Company.

5 Taylor, S.E. and Brown, J.D. (1988) Illusion and well-being: a social psychological perspective on mental health. *Psychol. Bull.* 103, 193–210

6 Sharot, T. *et al.* (2011) How unrealistic optimism is maintained in the face of reality. *Nat. Neurosci.* 14, 1475–1479

7 Clark, A. (2020) Beyond Desire? Agency, Choice, and the Predictive Mind. *Australas. J. Philos.* 98, 1–15

8 Klein, C. (2018) What do predictive coders want? *Synthese* 195, 2541–2557

9 Srivastava, N. and Schrater, P. (2015) Learning What to Want: Context-Sensitive Preference Learning. *PLoS One* 10, e0141129

10 Van de Cruys, S. (2017) Affective Value in the Predictive Mind. In *Philosophy and Predictive Processing* (Metzinger, T. K. and Wiese, W., eds), MIND Group

11 Pezzulo, G. *et al.* (2015) Active Inference, homeostatic regulation and adaptive behavioural control. *Prog. Neurobiol.* 134, 17–35