This reflection on Sun and Firestone was originally written with the intention of publishing as a letter in TICS. It ended up too compact to be viable as a letter, but I think it says something important. I intend to flesh out the argument into a full paper in the future. Intentions like that being what they are, I archive this version for now.

Colin Klein
June 3, 2020

Predictive Processing: Avoiding the elephant in the room.

Colin Klein

The Australian National University
Correspondance: colin.klein@anu.edu.au

Keywords: Dark Room Problem, Predictive Processing, Motivation, Mechanisms, Cybernetics

Sun and Firestone's [1] excellent presentation of the Dark Room Problem highlights an important dilemma for Predictive Processing (PP). On canonical formulations of PP, the forward signal at any stage carries only prediction error. As Clark notes, this "...is the root of the attractive coding efficiencies that these models exhibit, since all that needs to be passed forward through the system is the error signal, which is what remains once predictions and driving signals have been matched" [2]. Yet the same austere commitments also give rise to the Dark Room Problem. PP's toolkit is small: no desires, no drives, no reward signals. As Sun and Firestone detail, it is hard to see how obvious facts about motivation can be recovered in any biologically plausible way.

The Dark Room is an especially vivid example, but motivation raises a whole set of issues for PP. Survival problems are often exquisitely time-sensitive. When a lion is approaching, getting an answer in the long run is no good. Yet speed is not necessarily a strength of error-controlled systems. Testing and updating hypotheses takes time. In an early discussion of cybernetics, MacKay [3] noted a key information-theoretic tradeoff for error-controlled regulation. The less bandwidth you devote to the error signal, the more cycles of testing and checking you will need to converge on an answer; the more bandwidth you devote to the error, the less you are distinguished from traditional models. Indeed, it is worth remembering that Conant and Ashby—no foes of prediction error!—cautioned for this reason that "Error controlled regulation is in fact a primitive and demonstrably inferior method of regulation," [4] suggesting that pre-emptive control evolves precisely to overcome these limitations.

To make the tradeoff vivid, suppose you are very wrong about the world. Upon opening your front door, you see an elephant. Your higher-level model of your living room must be revised. PP says the higher-level model only gets evidence about the mismatch between its prediction and the incoming sensory information. How *much* evidence—in the information-theoretic sense—does the higher level model get to work with? At one extreme, the

error signal carries a single bit of information. Anyone who has negotiated teatime with a fussy toddler (*Kiwi? No. Toastie? No. Vegemite? No...*) appreciates how long it can take for binary feedback to converge on a solution. At the other extreme, the prediction sent downwards is an n-dimensional vector representing a guess, and the 'error' signal is an n-dimensional vector representing the difference between the guess and the elephant. But an n-dimensional vector is (by hypothesis) enough specify the elephant on its own—why not just send that, rather than using twice the overall bandwidth for the same result? Realistic intermediate options—e.g. the Euclidean distance between guess and truth—use less bandwidth but do not provide a unique solution. Hence they require more cycles to rule out possibilities. There is no free lunch. The simplifications that PP introduces incur a debt, and it is not a given that it can be repaid.

That said, more modest, local PP models do succeed. I think MacKay's insight also helps show why they work when they do: sometimes other demands tip the balance in favor of PP. Feedback-regulated motor control using efferent copies shows that improved latency can sometimes be more important than conserving bandwidth [5]. The toddler teatime dance arises because evaluating solutions is often easier, computationally speaking, than generating them. Reinforcement learning via reward prediction errors arguably leverages this insight for efficient learning with low-bandwidth signals [6]. Rao and Ballard's model of predictive processing in V1 adds complexity, but allows us to learn an optimal basis for representing ecologically typical visual scenes [7]; the advantage there is not in single tasks but in learning.

So the tradeoffs can, sometimes, fall in PP's favor. Yet this cannot not be taken for granted. Science is the search for mechanisms [8]. Merely capturing some phenomenon, even with exquisite mathematical precision, does not show that you have found the mechanisms at work. Indeed, there is an instructive sequel to the dispute between Chomsky and Skinner. Chomsky later recognized that a theory of syntax cannot just be a formal theory of permissible syntactic transformations—it must also be constrained by considerations about the biological mechanisms that underlie linguistic competence [9]. Among the extensionally adequate theories of syntax, then, there are ones that are still not correct because they do not tell us how things actually work.

Similarly so with brains. Brains are tightly constrained by space, speed, and energetic demands [10]. Sun and Firestone have shown that PP arguably cannot give a satisfying story about how these demands are met in the case of motivation. Focus on these constraints, rather than mere extensional adequacy, will help PP move forward.

References

[1] Zekun Sun and Chaz Firestone. The dark room problem. *Trends in Cognitive Sciences*, 24:346–348, 2020.

[2] A. Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–253, 2013.

[3] DM MacKay. The epistemological problem for automata. In C.E. Shannon and J. McCarthy, editors, *Automata Studies*, volume 34, pages 235–252. Princeton University Press, Princeton, 1956.

[4] Roger C Conant and W Ross Ashby. Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97, 1970.

[5] R Chris Miall and Daniel M Wolpert. Forward models for physiological motor control. *Neural networks*, 9(8):1265–1279, 1996.

[6] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[7] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[8] C.F. Craver. *Explaining the brain*. Oxford University Press, New York, 2007.

[9] Howard Lasnik. The minimalist program in syntax. *Trends in cognitive sciences*, 6(10):432–437, 2002.

[10] Peter Sterling and Simon Laughlin. *Principles of neural design*. MIT Press, Cambridge, 2015.