The predictive brain as a stubborn scientist

Daniel Yon[1*], Floris P. de Lange[2] & Clare Press[1]

1. Department of Psychological Sciences, Birkbeck, University of London, UK.
2. Donders Institute for Brain, Cognition and Behaviour, Radboud University, NL.

*Corresponding author: d.yon@bbk.ac.uk

Accepted as a Forum Article at *Trends in Cognitive Sciences* on Wednesday 10th October, 2018

Keywords: prediction, predictive coding, free energy, perception, action, cognition.

Abstract

Bayesian theories of perception have traditionally cast the brain as an idealised scientist, refining predictions about the outside world based on evidence sampled by the senses. However, recent predictive coding models include predictions that are resistant to change, and these stubborn predictions can be usefully incorporated into cognitive models.

There has been widespread interest in predictive coding (PC) models of cognitive functioning across the last decade [1]. Initial applications of these models to perception suggest that we infer the most likely state of the outside world by minimising prediction errors about its contents. More specifically, 'higher' neural areas predict the activity of 'lower' areas, and lower areas pass prediction error signals back up the hierarchy. Predictions are constantly updated based on these incoming error signals, and this iterative message-passing process generates a largely veridical model of the world. This bidirectional message-passing process – dubbed 'Perceptual Inference' (see Fig. 1) – likens perceptual processing to the scientific process. In the same way that an idealised scientist may develop hypotheses about the outside world, compare these to collected evidence and adjust their ideas accordingly, perceptual systems generate hypotheses about the extracranial world, compare these to evidence provided by the senses and use the discrepancy to refine their beliefs. The top-down predictions in these schemes provide an explanation for a range of neuroscientific phenomena, such as the finding that units in low-level sensory regions (e.g., primary visual cortex) can respond to implied rather than actual properties of the sensory input (e.g., illusory contours in a Kanisza triangle [2]).

The popularity of PC within the perceptual domain has spurred recent enthusiastic claims that the twin concepts of prediction and prediction error may provide a unifying basis for perception, cognition and action [1,3]. These PC models have therefore been applied to a range of topics in the cognitive and clinical sciences, including language [4], theory of mind [5], self-recognition [6], schizophrenia and depression [7]. Such accounts emphasize how the machinery of PC explains the flexibility of perception, action and cognition in a constantly changing world. However, these models that apply the PC concepts have given little attention to a core assumption of PC models – that not all

predictions are flexible. Namely, the brain deploys certain stubborn predictions (see Box 1 for discussion of the PC meaning of 'prediction') that are resistant to evidence-based updating.

For example, PC accounts hypothesise that actions are driven by strong sensory predictions about the intended state of one's body. In this process, known as 'Active Inference' (see Fig. 1), agents do not update their predictions based on ascending sensory signals, but instead engage reflexes that ensure the descending prediction comes true [1]. For example, if I would like my hand to grasp a cup (intended state) rather than remain immobile (current state), the prediction error generated by the mismatch between predicted and current states is resolved through reflexes that reconfigure the body in line with the predicted (intended) state. A conceptually identical scheme is thought to underlie homeostatic control of visceral body states (e.g., such that the body remains at the predicted temperature of ~37°C , [8]). A key postulate in these models is that for a top-down prediction to change the state of the world by driving action, it must be resistant to revision by sensory evidence. In computational terms, these neural predictions are assigned high 'precision', which is equivalent to ignoring sensory input (prediction errors) that could update them. The possibility that certain predictions are evidence-resistant recasts the brain as sometimes operating like a 'stubborn' scientist, possessing some hypotheses that evidence cannot change [9]. This process is necessary for Active Inference. It is not necessary for Perceptual Inference, but stubborn predictions are also possible in Perceptual Inference, where predictions are not updated on the basis of evidence yet do not result in action [2].

Incorporating stubborn predictions into cognitive theories could explain some phenomena that currently elude accounts which emphasise the 'flexible' nature of

predictive coding. For example, in computational neuropsychiatry it is frequently suggested that disorders which arise through aberrant predictions could be treated through behavioural and psychotherapeutic interventions that provide patients with the opportunity to learn the 'right thing' [10]. However, this approach may need careful consideration if psychopathologies arise due to aberrations in stubborn predictions, which are by definition resistant to learning. For instance, schizophrenia is frequently associated with passivity experiences or delusions of control, whereby patients report feeling that their movements are in fact caused by an external force [7]. If these delusions arise due to an aberration in stubborn predictions concerning whether sensory events temporally contingent upon one's actions are caused by them, these predictions may not be changed purely through behavioural learning interventions. Similarly, if depressive symptoms arise due to atypical predictions about the controllability of the external world [7] and representations of agency emerge through stubborn prediction mechanisms that control these states through Active Inference, these beliefs are unlikely to be updated simply by providing new sensory evidence.

Stubborn predictions could also be incorporated into models of typical cognitive functioning that are couched in PC frameworks. For example, one recent account [6] suggested that the hierarchical, belief-refining machinery of PC provides an ideal basis for understanding how the brain generates deep multimodal representations required for self-recognition. These models can accommodate the flexibility of self-representations (e.g. where a rubber hand is incorporated into one's body) through the notion that predictions about what constitutes 'my body' can be revised on the basis of co-occurring visual, tactile and proprioceptive signals. However, updating representations about which sensory inputs belong to 'my body' may depend upon the stubborn prediction that I only possess one body, with the inputs belonging to it being

spatially and temporally coincident. If one allowed these high-level predictions to change in the face of sensory evidence (e.g., when seeing four of one's own hands while standing next to a mirror) some peculiar representations of the self would likely emerge. Some other basic perceptual beliefs may also be resistant to change, e.g., it may be difficult to change the expectation that light comes from above or that falling objects will accelerate at a rate specified by gravity [2]. These represent situations where the predictions have generally always been true, both for the individual's ancestors and in their own learning environment.

When incorporating these predictions, it is important to consider the likely multiple and interacting causes for stubbornness. Some predictions are stubborn because they are necessary for survival (e.g., adaptive body temperature). These predictions will likely have been established phylogenetically, often through changes to neural structure (e.g., relatively few bottom-up projections in relevant neural regions) and will be impossible to change. Other predictions may become stubborn through ontogenetic processes (see Box 1), which could provide a principled explanation of 'sensitive periods' in development. Understanding the cause of stubbornness is especially important when attempting to alter such predictions. For example, drugs and psychotherapy frequently have synergistic effects in treating a variety of conditions. Given that many disorders are associated with aberrant neuromodulatory systems, and that PC proposes that neuromodulators control the relative weights given to top-down predictions and bottom-up evidence [10], pharmacological treatments are perhaps best conceptualised as interventions on the flexibility of beliefs (or equivalently, on the weighting of evidence) and indeed may be necessary if some stubborn predictions are ever to be altered.

In conclusion, the predictive brain may often function as a stubborn, rather than idealised scientist, failing to update predictions on the basis of sensory evidence. This element of PC frameworks has been largely overlooked within the cognitive sciences but incorporating stubborn predictions into cognitive models couched in hierarchical PC can aid their explanation of cognitive function in both health and disease.

Box 1: The meaning of 'prediction' in predictive coding and the cognitive sciences

Stubborn PC predictions may have received little attention by many cognitive scientists partly due to different understandings of 'prediction' between disciplines. Cognitive scientists typically equate predictions with expectations that reflect the sampled statistics of our environment [2], and these expectations will therefore tend to be flexible. In contrast, any descending cortical signal can constitute a prediction in PC. This reasoning reflects the fact that PC typically construes predictions as any mechanism that provides information about the likely distribution of environmental states, even if this distributional information is not known to the animal [1]. For example, the fact that line detectors in the visual system are connected to shape detectors can be thought of as a stubborn structural prediction that certain arrangements of lines (shapes) are more likely than others. Some stubborn predictions may be acquired through genetic 'priors' [11] embodying information about the kind of world we inhabit (e.g., adaptive body temperature). However other stubborn predictions may emerge via learning. For instance, learning about strong predictive relationships between events can 'block' or 'overshadow' learning about other relationships and the weight we give to new evidence in updating our beliefs declines when we estimate we are in a stable world [12].

References

[1] Friston, K. J. (2010) The free-energy principle: a unified brain theory? Nat. Rev. Neurosci., 11, 127–138.

[2] de Lange, F.P. et al. (2018) How do expectations shape perception? Trends Cogn. Sci. DOI: 10.1016/j.tics.2018.06.002

[3] Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci, 36, 181–204.

[4] Blank, H. and Davis, M. H. (2016) Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. PLoS Biol., 14, p. e1002577.

[5] Koster-Hale, J. and Saxe, R. (2013) Theory of mind: a neural prediction problem, Neuron, 79, 836–848.

[6] Apps, M. A. J. and Tsakiris, M. (2014) The free-energy self: a predictive coding account of self-recognition. Neurosci Biobehav Rev, 41, 85–97.

[7] Adams, R. A. et al. (2016) Computational psychiatry: towards a mathematically informed understanding of mental illness, J Neurol Neurosurg Psychiatry, 87, 53–63.

[8] Seth, A. K. and Friston, K. J. (2016) Active interoceptive inference and the emotional brain. Philos. Trans. R. Soc. Lond., B, Biol. Sci., 371, 19.

[9] Bruineberg, J. et al. (2018) The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. Synthese, 195, 2417-2444.

[10] Haker, H. et al. (2016) Can Bayesian theories of autism spectrum disorder help improve clinical practice? Front Psychiatry, 7, 107.

[11] Allen, M. and Friston, K.J. (2018) From cognitivism to autopoiesis: towards a computational framework for the embodied mind. Synthese, 195, 2459–2482.

[12] Behrens, T E. J. et al. (2007) Learning the value of information in an uncertain world, Nat. Neurosci., 10, 1214–1221.

Figure and legend

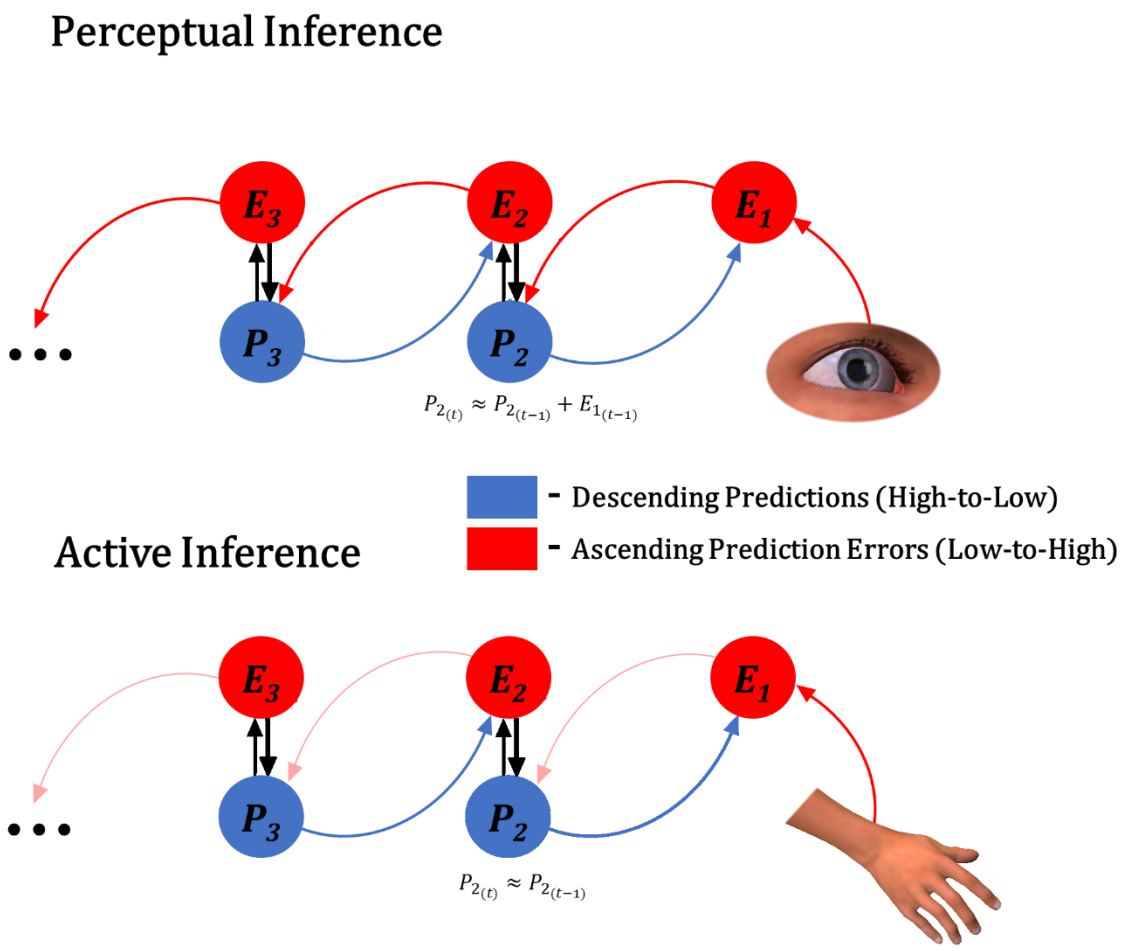

Fig 1: How information flows through the cortical hierarchy in predictive coding. In Perceptual Inference, sensory information (e.g. from the eyes) is conveyed up the hierarchy by prediction error units (red) to adjust prediction signals (blue). Activity in the prediction units is adjusted based on signals from the error units to minimise prediction error (i.e., prediction activity at timepoint *t* reflects predictions and error signals from timepoint *t*-1). Minimising prediction error generates veridical representations of the world. In Active Inference, the prediction error is instead reduced by peripheral reflexes (e.g. that move the hand) to change states of the body and the world in line with predictions. This process involves assigning greater weight to top-down

predictions (saturated blue arrows), which is equivalent to reducing the weight given to incoming sensory evidence (unsaturated red arrows). Therefore, predictions are resistant to revision through sensory evidence (i.e., activity at timepoint *t* is similar to *t-1*, providing the intended state remains the same) and are therefore 'stubborn'.