

# What do Predictive Coders Want?

Colin Klein  
Macquarie University  
Sydney NSW, 2109  
Australia  
colin.klein@mq.edu.au

## Abstract

The so-called “dark room problem” makes vivid the challenges that purely predictive models face in accounting for motivation. I argue that the problem is a serious one. Proposals for solving the dark room problem via predictive coding architectures are either empirically inadequate or computationally intractable. The Free Energy principle might avoid the problem, but only at the cost of setting itself up as a highly idealized model, which is then literally false to the world. I draw at least one optimistic conclusion, however. Real-world, real-time systems may embody motivational states in a variety of ways consistent with idealized principles like FEP, including ways that are intuitively embodied and extended. This may allow predictive coding theorists to reconcile their account with embodied principles, even if it ultimately undermines loftier ambitions.

## Keywords

Predictive Coding, Free Energy Principle, Homeostasis, Good Regulator Theorem, Extended Mind, Explanation

## Acknowledgements

Research on this work was funded by Australian Research Council Grant FT140100422.

For helpful discussions, thanks to Esther Klein, Julia Staffel, Wolfgang Schwartz, the ANU 2013 reading group on predictive coding, and participants at the 2015 CAVE “Predictive Coding, Delusions, and Agency” workshop at Macquarie University. For feedback on earlier drafts of this work, additional thanks to Peter Clutton, Jakob Hohwy, Max Coltheart, Michael Kirchhoff, Bryce Huebner, Luke Roelofs, Daniel Stoljar, two anonymous referees, the ANU Philosophy of Mind work in progress group, and an audience at the “Predictive Brains and Embodied, Enactive Cognition” workshop at the University of Wollongong.

# 1 Introduction

## 1.1 The dark room problem

Predictive coding (PC) models depict the nervous system as a machine for hierarchically minimizing the prediction error between internal models and sensorimotor input. Such models have successfully captured a variety of specific sensory and motor phenomena (Rao and Ballard, 1999; Huang and Rao, 2011; Clark, 2013; Hohwy, 2013). Inspired by those successes, some philosophers argue that predictive coding gives a quite general model of cognition—that is, that *everything* we do can ultimately be explained in terms of the minimization of prediction error (Friston, 2010; Clark, 2013; Hohwy, 2013; Clark, 2015).

Predictive coding draws further support from its formal links to Karl Friston’s Free Energy Principle (FEP). Proposed as an organizing principle of brain function, the FEP states that “...any self-organizing system that is at equilibrium with its environment must minimize its free energy” (Friston, 2010, 127).<sup>1</sup> Free energy is an information-theoretic concept closely related to entropy. Free energy, notes Friston “...bounds surprise, conceived as the difference between an organisms predictions about its sensory inputs (embodied in its models of the world) and the sensations it actually encounters” (Friston et al, 2012b, 1).

As I understand the two, PC is meant to be a story about how FEP is implemented. FEP says that organisms minimize free energy, and that they do so by minimizing the difference between their predictions about the world and the sensations they receive as input. PC postulates a mechanism by which such minimization takes place. Hence FEP gives general bounds on self-organizing systems, while PC gives a mechanism for implementing those systems. A heady mix, made even more attractive by Friston’s austere mathematical formulations of FEP.

According to both stories, the primary point of brains is to predict. You might think that this omits something important, though. There is more to life than just predicting what’s going to happen. We also *do* things. Prediction and action, in ordinary parlance, seem like importantly distinct processes. They have different effects, and they rely on different sorts of states. Predictions have to do with beliefs or credences: that is, we predict how the world *is* or *will be*. Action, by contrast, depends on preferences: we act to make the world the way *we desire it to be*. Credences and preferences can vary orthogonally, which is a further reason to think they are distinct things subserved by distinct neural states. Try to get by with only prediction, and you’ll end up just sitting there.

That is the nub of what has come to be known as the *Dark Room Problem*.<sup>2</sup> If I find myself in a dark, quiet room, I could predict very well what’s going to happen; indeed, there is a sense in which I could do no better at modeling my world. The world outside is complicated. Yet we don’t do this. Hence the challenge. As Andy Clark puts it:

---

<sup>1</sup>Note that ‘equilibrium’ can mean two things: thermodynamic equilibrium and local equilibrium with the environment. All organisms attempt to maintain local equilibrium with their environment in order that they may *avoid* pure thermodynamic equilibrium. Following Friston’s usage in this quote, I will use ‘equilibrium’ to mean local equilibrium with the environment.

<sup>2</sup>The original formulation was suggested by Mumford (1992) (though as an observation, not an objection): “In some sense, this is the state that the cortex is striving to achieve: perfect prediction of the world, like the oriental Nirvana... when nothing surprises you and new stimuli cause the merest ripple in your consciousness” (p247fn5).

How can a neural imperative to minimize prediction error by enslaving perception, action, and attention accommodate the obvious fact that animals don't simply seek a nice dark room and stay in it? Surely staying still inside a darkened room would afford easy and high-perfect prediction of our own unfolding neural states? (2013, 191)

Indeed, one of the things that we could model very accurately is that we'd die by continuing to stay in the room. Again, death is far more certain than anything that would happen were we to leave. Yet finding oneself in a dark room is not a death sentence. Something has gone wrong.

## 1.2 I'll be in my basement room

The dark room problem should be phrased with care. The difficult question is not "Why do organisms like us seek *food* rather than dark rooms?" There are plausible answers to that. Hohwy, for example, notes the advantages that accrue to humans to seeking complex, novel environments (2013, 175). Yet the dark room problem didn't arise because we were confused about human ethology. It stems instead from the concern that prediction is the sort of thing that can go on without being accompanied by any action whatsoever. The difficult question is really "Why do organisms like us *seek* anything at all, rather than just sitting quietly?" That is, the dark room ought to be read as presenting a problem about *motivation*, and how motivating states like desires can exist in a predictive coding framework.

To sharpen the point, consider what I take to be the obvious answer about the dark room. I in fact spend several hours a night laying quietly in a darkened room. What gets me moving? Usually, I'm hungry, and so I want to eat. If not for that, I might happily stay in bed for longer. But *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.) Similarly so for any other sensation I might feel.

Indeed, the dark room is just the most pressing way to illustrate the point. The problem is just clearest when prediction is trivial, but exactly the same issue arises for complex environments. Why couldn't I go about my day, predicting and modeling the world perfectly well, including the fact that I'm slowly starving to death? There is nothing conceptually impossible about an entity that merely monitors bad states: my car monitors its oil pressure to inform me when it is low. But it will continue to drive without oil up until it destroys itself.

Note that one cannot just say that cars do not perform homeostatic regulation and we do. That does not settle the issue: the dark room problem is an appeal for explanation about *why* and *how* we do homeostatic regulation, as it appears that PC/FEP does not explain this (and is compatible with the opposite). Nor will it help to appeal, as Hohwy (2013, 85–87) sometimes does, to minimizing *average* prediction error over longer timescales. The problem is that, for all we've said, the dark room creature has very low prediction error at any point—far lower than you or I do, given the complicated and hard-to-predict environments in which we find ourselves. Its average prediction error is thus lower than ours. Its total prediction error is lower too, because dying in a dark room gives you very little time in which to accumulate error.<sup>3</sup> Death

---

<sup>3</sup>Compare the old proverb: "If you give a man a fire, you warm him for a day. If you set a man on fire, you warm him for the rest of his life."

can't be held against the dark room creature: in the long run we are all dead. The challenge is thus to say why some of us stave it off for longer than others.

Prediction alone, even prediction of signals of danger, is not enough to get us to the adaptive actions that we in fact perform. That is, again, why most theories of action incorporate both belief and desire, or credence and preference, or *something* which promotes action rather than tracking the world. PC says it can do more with less. But that's only plausible if it can actually account for motivation.

### 1.3 What to expect

I have presented the dark room problem in a stark form to emphasize a point: strictly speaking, prediction and free-energy minimization *alone* are not enough to get adaptive action and motivation. That should be unsurprising: both predictive coding and free energy minimization are extremely general processes, broadly applicable to all kinds of systems. A homing torpedo attempts to minimize free energy by minimizing the difference between its current inputs and its expectation that the target be dead ahead. The better it does so, the *less* likely it is to survive the trip. That's because the homing missile 'expects' to ram its target, and minimization of surprisal with respect to that expectation results in its demise.

I emphasize this point because most presentations of FEP and PC pass over it quickly on the way to the standard solution. The mere fact that some system minimizes free energy (perhaps by implementing a predictive coding architecture) does not say anything at all about how it will act, because it does not say anything about what that system expects. The dark room problem merely makes this gap more dramatic.

I will turn to the proposed augmentations shortly. But before I do, it's worth emphasizing that these *are* augmentations. That will be important for evaluating the ultimate theoretical power of FEP and PC. Proponents of both are fond of emphasizing the simplicity of the basic story. So for example. Clark, singing the praises of PC, writes:

The sheer breadth of application is striking. Essentially the same models here account for a variety of superficially disparate effects spanning perception, action, and attention. Indeed, one way to think about the primary "added value" of these models is that they bring perception, action, and attention into a single unifying framework. (2013, 201)

Yet the degree of 'unification' involved must be evaluated with care—the basic story is straightforwardly inadequate, so it is the details of augmentation that must do the explanatory work, and thus provide the touchstone against which we evaluate the full theory.

Proponents owe us such a story. It will not do, for example, just to note that there must be *some* set of expectations that does the job: we don't explain how an organism stays alive by starting with the premise that it stays alive. The case of life is especially pressing in this regard. Every single organism that has ever lived either has died or will die comparatively soon. It is a remarkable trick to stay alive, and the trick is clearly a fragile one. The details matter.

With that in mind, here is the shape of the argument to come. In section 2, I begin by

considering the standard set of augmentations as they are presented in the predictive coding literature. I will argue that most don't work, and the ones that might work are mechanistically intractable. The predictive coder appears to be able to solve the dark room problem only at the expense of assuming an intractably complex set of innate expectations.

I then move, in section 3, to the more general formulation posed by FEP. I argue that there is a formulation of FEP which does not run into the problems faced by PC, but only because it is pitched at an extremely abstract level. Further, I claim that FEP so construed is literally false; it is best considered as an idealized model which must be altered in various ways in order to make it true to different systems.

Wet blankets thus thrown, I will conclude on a more upbeat note. Ambitious versions of FEP and PC appear to require internal representations of a fairly meaty sort. In section 4, I show that moving away from these ambitious versions leaves plenty of room for an embodied and extended account of the mind.

## 2 Prediction and expectation

### 2.1 Augmenting the standard story

Predictive coding alone cannot account for motivation: that is the thrust of the dark room problem. To account for motivation, the standard presentation of PC makes at least three additions. First, predictions need to be able to drive action. Building that in (in some form) is acceptable enough, the line goes, because prediction mismatch can be dealt with in two ways. Either the model can be updated to fit the sensory input (as in ordinary perceptual learning), or the sensory input can be changed to fit the model (by taking appropriate action). Either way of reducing prediction error is fine as far as the basic idea goes; in Clark's pithy phrase, "Motor control is just more top-down sensory prediction" (Clark 2015, 21). If we assume that motor control is also tightly linked to sensory input by the same sorts of predictive processes that govern model formulation, then we have the possibility for action.

The mere possibility for action is not yet enough, however. Without a further story, it is hard to see why the predictions that drive action wouldn't just be revised in light of recalcitrant experiences (Huebner, 2012). I'm hungry. I predict that I eat when I'm hungry, which should drive eating. But I'm not eating. Prediction error ought to drive me to revise my model. I could do so by abandoning the conditional. If I stop predicting that I eat when I'm hungry, I starve. A bad result.

Most ways of avoiding this do so by introducing a second mechanism which makes the desire-like predictions difficult or impossible to revise. Hence action becomes the only way to minimize prediction error. I'll adopt Hohwy's (2013) formulation, on which prediction error is always evaluated in the light of an independent estimate of the precision of the prediction. Both top-down models and bottom-up sensory states come with a precision—that is, an estimation of the variance of the predicted quantity. Precision can be used to give a kind of 'gain' on sensory signals which determines how seriously they ought to be taken (crudely, noisy sensory input loses against a precise model, and vice versa). Given such a mechanism, and the assumption that the relevant motivating states take a high enough precision to avoid revision in ordinary circumstances, desire-like predictions can become effectively impossible to revise. Further, since the

parameters used to estimate precision are themselves something which can be innate ((Hohwy, 2013, 64ff)

As it stands, the story still doesn't explain why any particular action is motivated rather than any other. We need to know why we do things that help us survive, rather than (as the homing torpedo does) things that help us perish. The third and final step involves postulating that organisms have an innate set of predictions that are determined by the states that are good or bad for them. Friston outlines the strategy in response to the Dark Room problem:

... when we enter a dark room, the first thing we do is switch on a light. This is because we expect the room to be brightly lit (or more exactly, we expect our bodily movements to bring this about). In other words, the state of a room being dark is surprising because we do not expect to occupy dark rooms. (Friston, 2013, 213)

Or consider Hohwy:

It is true we minimize prediction error and in this sense get rid of surprise. But this happens against the background of models of the world that do not predict high surprisal states, such as the prediction that we chronically inhabit a dark room (Hohwy, 2013, 87)

Or Seth, who claims that the principles behind PC mandate:

that organisms—in virtue of their survival—must avoid surprising states, where surprise is meant in an information theoretic sense as the negative log probability of the occurrence of an event... avoidance of atypical events (i.e., homeostatic regulation) necessitates a generative/predictive model of the causes of sensory inputs.(Seth, 2014b, 270-271)

As these expectations are crucial for survival, they must be at least partly innate. As Friston puts it: “This surprise depends upon (prior) expectations, but where do these prior beliefs come from? They come from evolution and experience, in the sense that if we did not have these prior beliefs, we would be drawn to dark rooms and die there” (2013, 213). These three steps together complete the standard picture. An organism gets out of the dark room because it expects to be doing something else. There is a mechanism which minimizes surprisal—that is, the difference between its current sensory state and its expectations—and that mechanism is blocked from merely changing the organisms expectations by the high precision put on the innate expectations. Hence it acts, and lives to fight another day.

Such is the story. Yet it will be successful only if we can spell out the expectations involved in the third step. That is nontrivial. I argue that once we try to make those expectations more precise, PC faces a dilemma. Either the relevant predictions are couched in terms of *states* of the world or over *actions* that the organism might take. Both are problematic: the former is empirically inadequate. The latter ends up too complex to be a plausible story about mechanisms, and is no longer distinguished from the sorts of belief-desire theories that PC was supposed to augment.

## 2.2 The soft tyranny of low expectations

Begin with predictions about states. That is, suppose that organisms have a number of predictions about states they might find themselves in, and seek to stay in likely states and avoid unlikely ones. These must be specified at the right fineness of grain. ‘Avoid death’ is fine counsel, but not yet a *policy* that I might follow. Conversely, the predictions cannot be too fine-grained: any of our states, described finely enough, is extremely improbable. The chance that anyone is sitting at *this* particular chair, drinking coffee made from *these* beans, at *this* time of morning is vanishingly small, despite its accuracy.

The trick will thus be to define these states in a useful way. Further, any successful strategy must identify states that are *good* for the organism with states that are *typical* for the organism (*mutis mutandis* for the bad). That’s a consequence of the double-edged nature of predictions. The very same entities must (on the PC account) be able to function both as claims about the world and as what motivates appropriate actions.

That is where the problem arises. There is no way to read ‘typical’ on which the typical and the good actually align.

For starters, ‘typical’ can’t be defined by relation to an organism’s *actual* life history. That is, I can’t predict that I will be in states that I have mostly been in before. Organisms live hard lives. I might have always been a little hungry. That makes hunger the typical state. It is still not one that I ought to seek out. Conversely, I might never have eaten my fill. That does not mean I should eschew the chance should it arise. Indeed, there are atypical states that lead to death, but which organisms are clearly motivated to pursue. Salmon spawn once and die. There is clearly some very strong drive that lies behind this behavior. That drive, trivially, cannot require the salmon’s future to resemble its past.

That is presumably why most PC theorists make the base broader, suggesting that we ought to make reference to the typical states of our conspecifics (which includes, I’ll assume, past conspecifics in evolutionarily similar circumstances). So for example, Hohwy writes:

[a] creature needs to be endowed with prior beliefs that tie it to its expected states. If it chronically expects to be in what are in effect its low surprisal states, then it will sample the world to minimize prediction error between those expectations and the state it actually finds itself in. . . These expectations are defining of the creature, because they tell us its expected states and thereby its phenotype. (Hohwy 2013, 85-6)

Predicting that one will be a typical conspecific has some obvious advantages. Individual salmon spawn but once in a lifetime, but many salmon spawned: what is atypical for an individual life might be typical when we zoom out to the collective history of the species.

(Note that while the relevant predictions track facts about conspecifics, they must be *about* my own states. That is, I predict that I drink when thirsty, though I do so because my conspecifics typically drink when thirsty. That’s because it must be the mismatch with my own state that drives appropriate action, and my own state provides very little evidence about the typical conspecific.)

Alas, even reference to the typical conspecific won’t do either. Evolutionarily typical states

need not be the most adaptive ones. Most acorns rot (Griffiths and Gray, 1994). Most juvenile fish are eaten before adulthood (Dahlberg, 1979). Conversely the adaptive actions may not be typical of the species. Among southern elephant seals, most males never mate (Fabiani et al, 2004). It would be disastrous if each took that as reason to avoid it. Again, there is a wedge between what is typical and what is good, even when we broaden the reference class to include conspecifics.

Many of our conspecifics may be failures. They are thus a poor guide to what we ought to do. That suggests a narrower reference class. Perhaps it is only the *successful* conspecifics that count. Even if most of my conspecifics never mated, the successful ones obviously did. They are the ones I ought to emulate. (Dale Carnegie philosophy of mind: expecting to be successful is the best way to end up successful.) We might cash out success in a variety of ways—perhaps it is only the conspecifics that reproduced, or that had especially high fitness, that ought to be counted.

However it goes, exactly the same problems recur. On the one hand, there might be situations that are extremely good, but rare even for successful conspecifics. Few successful humans have won the lottery. That’s no reason to burn a winning ticket. Conversely, even successful conspecifics might have lived in harsh environments with unfortunate features. Most humans, even reproductively successful humans, have been illiterate. They were successful because they adapted to their environment without the ability to read. But that’s no reason to avoid literacy now.

In addition to the same old issues, reference to successful agents—especially if highly idealized—introduces two more subtle problems. First, there is variation among conspecifics. Successful conspecifics might thus be better placed than you. If the most successful seals were larger and so able to win fights for mates, then weakling seals would do worse to imitate them. Conversely, I might know more than the successful conspecific. Most of my ancestors would be struck dumb with amazement were they to see a flashlight. I shouldn’t do the same if the lights go out.

Second, there will be situations that successful agents entirely avoid. If so, then predictions defined relative to the states of actual agents will be undefined. But we have motivations in bad states as well as good. So (as a crude example), what did the successful agents do when they found themselves naked in the snow? Suppose this has just never happened: successful agents were and are always smart enough to bring a coat. But then the successful agents offer no guide to what we should do when we find ourselves in that situation.

Note that the problem is not merely that being naked in the snow is surprising. It is. It is rather that getting out of the snow by going inside is equally surprising, because it never happens among the successful. Yet as practical problems go, this is not a particularly complex one: if you’re naked in the snow, go inside. That is good advice even if no successful agent has ever had occasion to take it.

While the example is contrived, the problem is more general. So long as any PC account of motivation fixes predictions based on frequent or likely states of an organism, it will be unable to give counsel in entirely novel situations. Yet being able to recognize novel situations and avoid them (for threats) or exploit them (for opportunities) is one of the hallmarks of adaptive behavior.

## 2.3 Predicting policies

Spelling out the adaptive actions in terms of states didn't work. An important reason why it failed was because the presumed repertoire of actions was so poor. The only actions available to organisms were to avoid the unlikely states and seek the likely ones. But once we specified the states at any level of precision, we discovered that there were unlikely but good states, and likely but bad ones. Mere 'approach' and 'avoid' imperatives might be enough for simple organisms, but they can't be enough for any moderately complex one. If I'm hungry, I can't do any old thing: I need to take a complex series of actions in order to find food.

A natural response to this challenge is to move to predictions about the *activities* of the successful conspecific: that is, what the successful conspecific *would have done* given your situation. As Friston, Samothrakis, and Montague put it, the relevant prior beliefs "are not about states of the world but transitions among states (i.e., a policy)" (Friston et al, 2012a, 524). The proper counsel to the hungry is not simply that they should avoid being hungry, but to look for food.

In general, then, let's assume that organisms have predictions about the optimal actions to take when they face challenges. Note here that the prediction error mismatch would no longer be with the state of being hungry, but rather with predictions like "I look for food when I'm hungry."<sup>4</sup> As with states, the precision on this prediction will have to be kept arbitrarily high, lest I infer from my lack of action that I am wrong about what I do.

This brief sketch would obviously need further refinement. However it is spelled out, though, PC faces a different, equally bad, problem.

We need predictions about action in order to account for motivation. But those must be in *addition* to the ordinary predictive models ("I am hungry," "This room is dark") that underly perceptual inference. We act, but we also represent the world. That means PC has now carved off two types of state. One has content about states of the world, needn't be innate (except in broad outlines), tracks states of the world, and changes in response to perceptual evidence. The other has content about future actions, must be fairly fine-grained and innate, drives action, and is immune to revision through perceptual evidence. All of which is to say that *we've just reconstructed the classic differences between beliefs and desires*. There are now states with a mind-to-world direction of fit and states with a world-to-mind direction of fit (Anscombe, 1957). Further, we've had to build in those differences by hand. That makes Clark's promise of bringing "perception, action, and attention into a single unifying framework" (2013, 201) seem like a bit of a bait-and-switch. There is a single type of process, true, but the complexity has simply been pushed out into the parameterization.<sup>5</sup>

Further, framing the motivating states in terms of actions reveals how complicated the PC story would have to be. Defining prediction error in terms of states was at least simple: being too cold

---

<sup>4</sup>That would solve another problem for prediction in terms of states. On most versions of PC, "...perceptual content is the predictions of the currently best hypothesis about the world" (Hohwy, 2013, 48). As Wolfgang Schwartz points out (personal communication), if we took this literally it would mean that we act because we hallucinate the goal state and thereby move towards it. That is absurd. An action-based view, by contrast, lets the organism veridically model both the problem and progress towards the solution.

<sup>5</sup>Compare: universal Turing machines can also be made very simple in the sense of having relatively few states and symbols. Further, such simple machines are completely universal in the sense that the same process does whatever can be done. But that comes at the cost of considerable complexity in coding the inputs (Minsky, 1967, §14.8). As such, the simplicity and universality of minimal Turing machines does not constitute an argument in their favor as a plausible architectures of mind. (There are, of course, many other considerations against TM architectures.)

is bad, so ceasing to be cold is good. That wasn't enough to produce adaptive action, because it couldn't tell the cold organism the right thing to do. Yet once we start building in adaptive actions as part of the predictions, we realize just how complex they would have to be. In the ordinary course of things, we have to choose between many actions at a time. Actions thus have to be ranked against one another. If I am hungry *and* I need to plan a lecture, I must decide which comes first—and that in turn will depend on many different idiosyncratic considerations about context. Many actions can be satisfied in a variety of ways, and the correct action is usually context-sensitive.

This challenge is formidable for a system that has only one type of primitive with which to work. Note, for example, we can't simply pack the adaptive actions into a series of conditional predictions like "When  $(C_1 \wedge C_2 \wedge C_3 \dots)$ , I will  $(A_1 \vee A_2 \vee A_3 \dots)$ ." There are astronomically many different predictions of that sort that even simple agents need to satisfy. Ordinary belief-desire models can avail themselves of the standard combinatorial resources of computational theories to try to sort out these problems. It is not at all clear how hierarchy plus prediction error could do the job. (Indeed, in this regard PC seems to face the exactly the same set of problems as did traditional behaviorism (Putnam, 1967/1991).) Further, the vast majority of these must be innate if the organism is to have any chance at survival—for remember, without these additional expectations in place, neither FEP nor PC can solve the dark room problem.

Nor is it clear how we could come to learn about changes in the most adaptive ways of acting. I am transferring through LAX. I notice a sign saying that a new walkway connects two terminals, and I thereby learn that I could walk that way to catch my flight. How does that work? Must we say that my conditional expectations are defined over every *possible* combination of contexts? But it's a bit of a stretch to think that my innate endowment of expectations includes facts about construction at LAX. Conversely, suppose I find that my old favorite walkway at LAX has closed down. Recall that my previous expectation—'When I have a tight connection, I will go down *this* corridor' must have been given an arbitrarily high precision (otherwise, upon being late and standing in front of the corridor I might just as well have revised my model about how I acted rather than taking the path, and so missed my flight). But then if that expectation is un-revisable, then it is un-revisable even when that same action starts becoming maladaptive.

Sufficient cleverness may let PC theorists avoid these difficulties. an empirically adequate story, though, seems likely to be very, very complex. That is the deep issue. Again, the bookkeeping with one state gets complicated, while a system with more primitives (such as a belief-desire model) can more easily keep track of shifting needs, goals, and facts about the world. It is therefore not obvious that PC built along these lines can solve the dark room problem. If it does, it may well closely resemble more traditional theories about motivation.

### 3 Free energy and explanation

The preceding story was about a proposed mechanism, predictive coding, which was meant to implement the process described by FEP. FEP itself can be read as more abstract, however: as a very general story which is meant to describe and explicate whatever particular mechanism happens to move us. On this reading, to talk about an organisms' expectations of the world is not to propose that there are specific, concrete things which play a causal role in driving behavior. Rather, talk about minimization of free energy and an organisms' expectations is meant to be something like a description of how whole organisms behave. On such a story, what makes it true that I expect to go inside when it is cold may well be something like a

traditional desire: the point is not to describe mechanisms but rather the overall dynamic of a system.

So conceived, FEP would still be vulnerable to the arguments in 2.2, as those concerned the general problem of cashing out expectations in terms of states. FEP might well be immune to the arguments in 2.3, however, for those concerned unattractive features of a specific way of implementing expectations. There is no contradiction in saying that people behave *as if* they have an astronomical list of fine-grained conditional expectations, so long as we are clear that they don't behave thus because each of those expectations are individually embodied in the nervous system.

Even considered abstractly, however, I think that we ought to be wary of the FEP. For the dark room still raises its head: we ultimately want abstract principles to describe and constrain causal stories. It is not obvious that FEP actually does so.

It will not do to suggest, as Friston sometimes does, that FEP makes survival so obvious as to be something like a tautology. So for example, he claims that the FEP

means a Dark-Room agent can only exist if there are embodied agents that can survive indefinitely in dark rooms (e.g., caves). In short, Dark-Room agents can only exist if they can exist. The tautology here is deliberate, it appeals to exactly the same tautology in natural selection (Why am I here? — because I have adaptive fitness: Why do I have adaptive fitness? — because I am here)... In fact, adaptive fitness and (negative) free energy are considered by some to be the same thing. (Friston et al, 2012b, 2)

Similarly, just before discussing the good regulator theorem (about which more shortly), Hohwy writes that "...in a circular sounding way, the idea also asserts that the fact that we tend to find these creatures in certain states and not others explains why they have the expectations they have" (Hohwy, 2013, 86).

Appeal to apparent tautologies should trouble you. For whatever tautologies do, they don't explain why things happen. At best, they give us *reason to believe* that something is the case. But philosophy of science has moved away from epistemic conceptions of explanation and towards ontic ones (Woodward, 2003; Craver, 2007). Good explanations detail a causal story, and it is not obvious that FEP does so.

Friston's parallel with natural selection is telling in this regard. Despite his claim, very few biologists or philosophers consider natural selection to embody a tautology (claims of tautology are usually associated with the *enemies* of evolutionary theory). First, natural selection says that *heritable* characters are responsible for at least some reproductive success, and this explains their prevalence in subsequent generations. Second, some heritable traits do more than correlate with survival: they are also part of the *mechanism* for survival. They thus confer a propensity for survival both in the organism and in the subsequent generations that inherit the trait. Both responses, note, assume that what is important for explanation are the particular traits that causally contribute to survival in present and future generations (Sterelny and Griffiths, 1999, 84-5). It is the causal story that explains adaptation, and that causal story goes in one direction.

The right direction of explanation must go from minimizing free energy to survival. Yet insofar as FEP implies a causal story about that direction of explanation, it appears to be wrong. On

the one hand, minimizing free energy cannot be sufficient for survival. *Nothing* is sufficient for survival—to repeat a theme that recurs, everything dies. However we build in facts about the organism, there are obviously situations where we minimize free energy and yet perish. Consider, for example, Hohwy’s example of the unfortunate mobster hurtling towards the bottom of the sea, cement overshoes dragging him inexorably downward ((Hohwy, 2013) 85-86, embellished). Hohwy suggests that the best our mobster can do to minimize surprisal (and therefore free energy) is to use perceptual inference to at least accurately model his sorry state. But of course, such accurate modeling—and so minimizing surprisal as best the agent can—cannot actually prevent his death.

Conversely, minimizing free energy cannot be necessary for survival either. I think this fact is often obscured by the contrast cases authors choose when they explicate FEP: the options are either being happy and healthy or else hurtling towards the bottom of the sea. But there is a large grey area between the two: life mostly requires getting by well enough, most of the time. Yet FEP places an austere set of constraints on organisms: they must minimize free energy, and so resist change, in some way that approximates optimality. We know humans aren’t optimal, though. We can’t be. We die. We die in situations that other humans successfully navigate. Every day, people step off the curb without looking, drunkenly taunt crocodiles, work on live wires, and text while changing lanes. Sometimes that works out, and sometimes it doesn’t. Even when the stakes are lower, we’re often less than robust.

Natural selection requires differential fitness between organisms: that is, some have to be doing *better* than others within a niche.<sup>6</sup> Friston, confusingly, even seems to agree with this, claiming that within a niche “...some creatures (models) are more optimal than others” (Friston et al, 2012b, 6). But ‘optimal’ is not a comparative; if there is variation, some creatures *aren’t* optimal. I doubt, then, that any living creature comes close to minimizing free energy: most of the time, we just keep it low enough to avoid dying, and we do not do so robustly.

Indeed, I think that the problems with PC ultimately derive from its links to FEP. The focus on *minimizing* prediction error, along with the corporal bottleneck that means we can only make one action at a time, is precisely what makes it so difficult to spell out the relevant predictions in terms of states. The absolute nature of minimization leaves no room for intermediate, partial satisfaction of desires. In any given situation, there can only be one right thing to do. If that is eating an apple, then eating half an apple is as bad as eating a chunk of hot lava. Neither minimize free energy, and so neither can be the right solution. The FEP picture thus reminds me of the Stoic position that good does not admit of degree, and hence that

...just as a drowning man is no more able to breathe if he be not far from the surface of the water, so that he might at any moment emerge, than if he were actually at the bottom already, and just as a puppy on the point of opening its eyes is no less blind than one just born, similarly a man that has made some progress towards the state of virtue is none the less in misery than he that has made no progress at all. (Cicero *De Finibus* III.14.48, Loeb translation)

That is why expectations had to be as numerous as situations for action: when there is no room

---

<sup>6</sup>An anonymous reviewer suggested that Friston simply uses ‘existence’ and ‘being well-adapted’ as synonyms. Perhaps so, but they are not synonyms, precisely because there must be variation in fitness to drive natural selection. Further, a change in environment can make an existing organism poorly adapted (by, for example, rendering it sterile) without changing whether or not it continues to exist. Finally, evolution and adaptation requires organisms to do more than exist—they must also reproduce. Replication is fundamental to Darwinian theory (Godfrey-Smith, 2009). Yet FEP says very little, as far as I know, about that more specific requirement.

for half-measures, every possible context must be addressed.

I find absolutism as unappealing in philosophy of mind as it is in ethics. As many authors have stressed, the problems of living have to be solved in real-time, with finite resources, in uncooperative and rapidly changing environments (Feldman, 2013, 27). We have good reason to think that evolution should prefer systems that are good enough, rather than systems that would be optimal in the long run (Simon, 1996). FEP may tell us what the *best* organisms look like, and departure from optimality may even (in many cases) increase the chance of death. But again, we constantly face the chance of death, and yet persist long enough to count as survival.

None of this is to suggest that considerations of ideal or optimal systems have no place in theory-building. They do. But there are idealizations and idealizations. FEP strikes me as closer to what McMullin (1985) called a Galilean idealization: literally false, but with some understanding gained via over-simplification. Such idealizations can often be elaborated to be true of particular systems, and those elaborated models—which often look very different from the original model—can have considerable explanatory power (Klein, 2008).<sup>7</sup> But if this is the case, then it is worth keeping in mind that FEP is a starting point from which one might develop explanations, and that its defense would ultimately rest on the empirical adequacy of detailed models which spring from it. Simplicity does not count in its favor, for FEP is simple in the way that friction-free planes and infinite populations of bunnies are simple: that is, a deliberate simplification, which buys scientific fruitfulness at the cost of literal truth.

## 4 Conclusion: good regulation requires getting out of your head

The preceding has been pessimistic about both FEP and PC’s prospects as Grand Unified Theory of the mind. I hope to end on a more upbeat note. Many have noted—either with pride (Hohwy, 2013) or horror (Anderson and Chemero, 2013)—that the ambitious forms of predictive coding appear to be incompatible with the embodied, extended turn in philosophy of mind. Not everyone accepts this conclusion. Clark (2013), notably, denies it. But when PC is pitched in the heroic register, it does seem hard to resist the conclusion that internal models *alone* are sufficient to account for behavior.

I have no particular stake in the embodied/extended cognition debates. It strikes me, however, that several of the considerations advanced above might leave more room for embodied and extended models.

To begin, consider a principle sometimes invoked by proponents of PC: Conant and Ashby’s Good Regulator Theorem (GRT) (1970). Conant and Ashby showed that any system that is an optimal regulator—that is, a system that can reliably maintain itself in a set of ‘good’ states—must model its environment. The GRT is a cybernetic principle and, as Seth persuasively argues, PC/FEP is the modern inheritor of the cybernetic tradition (Seth, 2014a). Friston explicitly links GRT and FEP, claiming that

---

<sup>7</sup>General principles about optimal function may also offer guidance on the boundaries and constraints on such organisms. These constraints are likely to be very broad, and so very easy to satisfy. Friston, for example, notes that “the free-energy bound on surprise tells us that adaptive agents must perform some sort of recognition or perceptual inference” (Friston et al, 2012b, 6). Undoubtedly true, but not particularly controversial. Leibniz was probably the last major philosopher to deny that adaptive action required causal contact with the external world. Few were convinced.

Under the free-energy principle, the agent will become an optimal (if approximate) model of its environment. This is because, mathematically, surprise is also the negative log-evidence for the model entailed by the agent. This means minimizing surprise maximizes the evidence for the agent (model). Put simply, the agent becomes a model of the environment in which it is immersed. This is exactly consistent with the Good Regulator theorem of Conant and Ashby (Friston et al, 2012b, 2)

That is exciting news for philosophers, Hohwy argues, because it appears to put strong constraints on our representational capacities (Hohwy, 2013, 86). Since we’re alive and kicking, then we must (via the GRT) represent the world pretty accurately.

GRT is a strange bedfellow for predictive coders. Indeed, GRT makes the dark room problem especially glaring. The ‘good’ states of the GRT need not be states *of the regulator* (Conant and Ashby, 1970, 90). A thermostat keeping a tank of oil viscous has its optimality measured against states of the tank, not states of the thermostat. So the GRT does not even pretend to guarantee the persistence of the regulator. Cyberneticists like Ashby were also quite aware of this, and spent a lot of time detailing mechanisms and distinguishing classes of regulatory systems that were specific to organisms (Ashby, 1956).

More importantly, Conant and Ashby note that models which incorporate feedback (as PC does) can’t be optimal regulators. Feedback systems are always a step behind. Hence “Error-controlled regulation is in fact a primitive and demonstrably inferior method of regulation,” and optimal systems must use completely feedforward predictors (Conant and Ashby, 1970, 92). The fact that there is error to correct at all shows that we are not good regulators in the GRT sense. Similarly, minimizing free energy is, on the PC account, a process that occurs over unspecified but often relatively long time periods. But then even in cases where we *do* minimize free energy, that’s often only after a stretch of time in which we do not.

Models that incorporate feedback thus cannot be optimal regulators in Conant and Ashby’s sense. Since it is pretty clear that we do incorporate feedback to guide our actions, we cannot be optimal regulators either. There is an upshot to all of this. Although feedback-driven regulation can’t be optimal, the requirements it implies are correspondingly less demanding. Optimal modeling requires having an internal feedforward model that bears some appropriate iso- or homomorphism with the environment (Conant and Ashby, 1970, 96). It is hard to see how such models could avoid being strongly internalist. Models with feedback have fewer constraints: at their best they require only a one-to-one correspondence between the set of regulator states and the set of world-states that require distinct actions (Conant and Ashby, 1970, 96).

Fans of embodied cognition should be pleased. Nothing constrains the states that satisfy this weaker requirement to be states of the *brain*. Ashby is explicit about this, writing that passive devices like the “the tree’s bark, the seal’s coat of blubber, and the human skull” are all crucial bits of regulatory networks that screen off environmental variation (Conant and Ashby, 1970, 201). Formally speaking, these play exactly the same regulatory role as the more straightforwardly representational modeling that the brain does.

That means that much of the ‘modeling’ done by regulators like us can be done by the body itself (Seth, 2014a). I thus agree with Seth, claim about the link between homeostatic models and predictive coding, “. . . it is left open whether such models need be explicitly encoded in control structures or can remain implicit in an agent’s phenotype” (2014b, 271). This seems especially obvious for basic homeostatic processes. When we look at what matters for keeping our bodies

alive, we find numerous processes that are not brain-bound. Our ability to regulate plasma osmolarity or renal creatine levels (say) is crucial for survival. Yet many of the mechanisms that underly this ability are bodily and peripheral rather than neural and central.

Even homeostatic processes with a significant central component often depend on peripheral processing. Consider pain perception. Pain perception is part of a sophisticated homeostatic system that keeps our bodies more or less intact (Klein, 2015). Yet a substantial portion of pain processing is done peripherally and spinally, and this peripheral processing is crucial for adaptive behavior (Melzack and Wall, 1965). Indeed, it has been one of the general trends of pain science in the past century to emphasize the importance of broad, distributed processing that crosses the central-peripheral boundary (Melzack, 1999).

In addition to looking to the body, we might also look outward to the world. I argued that humans are not optimal regulators, as evidenced by the fact that humans regularly make fatal mistakes. Perhaps you worried that this was oversold, as fatal mistakes cannot be *very* common. The frequency is a bit besides the point: failures are enough to show that we are not optimal. But it is true that modern humans do not tend to die for stupid reasons.

The ‘modern’ part is crucial, though. Life used to be more precarious. Erik Larson (2004) recounts, for example, that in 1890s Chicago trains killed an average of two pedestrians *a day*. Our improved survival rates are not due to increased vigilance. Rather, much of the job of modeling fatal contingencies has been *externalized*. I confidently navigate Central Station, in a hurry, while listening to music. I do so because I predict that there will be signs, warnings, and walls that protect me from potentially dangerous situations. My complex motivational state—my desire to avoid getting killed by a train—can be replaced by easier-to-implement policies of avoiding clearly-marked dangers. That works well: I’ve never once been hit by a train.

Not all motivational contingencies can be offloaded into the environment in that way. Many homeostatic drives are too basic, and some desires are simply too complex to offload. As fans of extended cognition are wont to emphasize, however, even complex desires can often be broken down into a series of simpler, signposted steps. This is why (for example) surgeons and airline pilots drastically improve their performance by offloading important regulatory steps onto external checklists (Gawande, 2010).

We are not optimal regulators. But insofar as we are regulators, part of our regulatory capacity is subserved by both our bodies and by structures that we have placed into the world. Fans of predictive coding are fond of the gnomonic claims that “phenotypes are predictors (models) of their low surprisal states” (Hohwy, 2013, 86). Given what I’ve said, we can make sense of this by noting that ‘phenotype’ might in fact cover the whole body and perhaps the local niche as well.

I am pessimistic about the ability of predictive coding models to handle the basic motivational states. I intend that to be a productive pessimism. For I think that backing away from grand ambitions might actually open up space to explore more body- and world-based senses of representation.

## References

- Anderson ML, Chemero T (2013) The problem with brain GUTs: Conflation of different senses of ‘prediction’ threatens metaphysical disaster. *Behavioral and Brain Sciences* 36(3):204–205
- Anscombe GEM (1957) *Intention*. Harvard University Press, Cambridge
- Ashby WR (1956) *An introduction to cybernetics*. Chapman & Hail Ltd., London
- Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36(3):181–253
- Clark A (2015) *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press
- Conant RC, Ashby WR (1970) Every good regulator of a system must be a model of that system. *International journal of systems science* 1(2):89–97
- Craver C (2007) *Explaining the brain*. Oxford University Press, New York
- Dahlberg MD (1979) A review of survival rates of fish eggs and larvae in relation to impact assessments. *Marine Fisheries Review* 41(3):1–12
- Fabiani A, Galimberti F, Sanvito S, Hoelzel AR (2004) Extreme polygyny among southern elephant seals on Sea Lion Island, Falkland Islands. *Behavioral Ecology* 15(6):961–969
- Feldman J (2013) Tuning your priors to the world. *Topics in cognitive science* 5(1):13–34
- Friston K (2010) The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11(2):127–138
- Friston K (2013) Active inference and free energy. *Behavioral and Brain Sciences* 36(3):212–213
- Friston K, Samothrakis S, Montague R (2012a) Active inference and agency: optimal control without cost functions. *Biological cybernetics* 106(8-9):523–541
- Friston K, Thornton C, Clark A (2012b) Free-energy minimization and the dark-room problem. *Frontiers in psychology* 3:1–7
- Gawande A (2010) *The checklist manifesto: how to get things right*. Henry Holt and company, New York
- Godfrey-Smith P (2009) *Darwinian populations and natural selection*. Oxford University Press, New York
- Griffiths PE, Gray RD (1994) Developmental systems and evolutionary explanation. *The Journal of Philosophy* 91(6):277–304
- Hohwy J (2013) *The predictive mind*. Oxford University Press, New York
- Huang Y, Rao RP (2011) Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(5):580–593
- Huebner B (2012) Surprisal and valuation in the predictive brain. *Frontiers in psychology* 3(415):1–2
- Klein C (2008) An ideal solution to disputes about multiply realized kinds. *Philosophical Studies* 140(2):161–177

- Klein C (2015) *What the Body Commands: The Imperative Theory of Pain*. MIT Press, Cambridge, MA
- Larson E (2004) *The Devil in the White City*. Vintage Books, New York
- McMullin E (1985) Galilean idealization. *Studies in the History and Philosophy of Science* 16(3):247–273
- Melzack R (1999) From the gate to the neuromatrix. *Pain* 82:S121–S126
- Melzack R, Wall P (1965) Pain mechanisms: a new theory. *Science* 150(699):971–979
- Minsky ML (1967) *Computation: finite and infinite machines*. Prentice-Hall, Inc., Englewood Cliffs, NJ
- Mumford D (1992) On the computational architecture of the neocortex. *Biological cybernetics* 66(3):241–251
- Putnam H (1967/1991) The nature of mental states. In: Rosenthal DM (ed) *The Nature of Mind*, Oxford University Press, New York, pp 197–210
- Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* 2(1):79–87
- Seth AK (2014a) The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In: Metzinger TK, Windt JM (eds) *Open Mind*, MIND Group, Frankfurt am Main
- Seth AK (2014b) Response to Gu and FitzGerald: Interoceptive inference: from decision-making to organism integrity. *Trends in cognitive sciences* 18(6):270–271
- Simon HA (1996) *The Sciences of the Artificial*, 3rd edn. MIT Press, Cambridge
- Sterelny K, Griffiths PE (1999) *Sex and Death*. University of Chicago Press, Chicago
- Woodward J (2003) *Making Things Happen*. Oxford University Press, New York