A Humean challenge to predictive coding

Colin Klein

The Australian National University

colin.klein@anu.edu.au

Abstract: Predictive coding (PC) theories are attractive in part because they posit a single type of state that can play the roles standardly attributed to both beliefs and desires. Drawing on a well-known proof by Lewis in favor of Humeanism, I argue for a conditional claim: *if* you want PC to be a mechanistic story, *and* you want to be a thoroughgoing Bayesian, *then* you should be a Humean. Most predictive coders want the antecedent but deny the consequent, which is a problem. The story demonstrates that PC has a serious and underappreciated issue around learning what is valuable.

1. Predictive coding vs Humeanism

1.1 Humeanism

Humeanism, broadly speaking, is the thesis that belief and desire are distinct psychological entities with distinct but complimentary roles in causing action. Desires motivate you to make the world a certain way. Beliefs tell you the way the world actually is. Belief and desires work together to drive action. But beliefs and desires are fundamentally *orthogonal/* As Smith (1984,7) puts it, "For any belief and desire pair that we imagine, we can always imagine someone having the desire but lacking the belief, and vice-versa." Since there are no necessary connections between Belief and Desire, neither cab be reduced to the other. Our psychology thus always needs two kinds of states to explain action.

Humeanism has primarily been defended in the context of moral psychology. Moral beliefs provides an interesting test case: anti-Humeans argue think there's no gap between believing an action to be good and being motivated to do it. Humeans leave open the possibility of being unmoved by the mere belief that some action good (as, indeed, often seems to be the case).

Within philosophy of mind, Humeanism is often taken as a commonplace of folk psychology (Dennett 1987). That is not (just) tradition. There are good theoretical reasons to separate out beliefs and desires. Beliefs and desires are often taken to have different direction of fit, and so are responsive in different ways to the world. Beliefs and desires also seem to have different dynamics. Beliefs change on the basis of evidence and deliberation. Desires change by being satisfied, and perhaps by other (often obscure) routes. Our beliefs about our desires don't seem to affect our desires. I believe that eating crisps is bad for you. That, on its own, doesn't affect my desire for a pack of crisps (alas!). Conversely, functioning grown-ups don't let what they want affect their estimation of what is true: however much I desire that crisps be healthy, I know that this is no evidence in favor of them being healthy.

Distinguishing belief and desire also brings straightforward *explanatory* advantages that are difficult to get with just one entity. Because beliefs and desires are orthogonal, we can trivially explain why two people with the same beliefs act differently (their differing desires), and *mutis mutandis* for identical desires. Separating belief and desire give us combinatorial resources: the complex set of actions we observe gets explained by a relatively smaller, tidier set of beliefs and desires plus laws of combination.

As an example, consider the hoops that old-fashioned behaviorism had to jump through to try to reduce belief/desire talk down to talk of observable behavioral patterns (Putnam 1967).

What does it mean to say that someone is thirsty? That they are disposed to drink if there is water available. But that can't be right: they are disposed to drink if water is available *and* there is no lion between them and the water *and* there is not lemonade nearby *or* their past encounter with that particular lemonade has not produced the utterance "this lemonade is really far too sour" *and…* Even if you think you *could* get by just talking about behavior in this way, every explanatory law ends up sounding impossibly complex and *ad hoc.* Whereas by talking about the desire for water and how it interacts with other beliefs and desires, one gets a lot of explanation for cheap.

I've given only a crude sketch of Humeanism. The broad division is really between doxastic and motivational states, and beliefs and desires might not be the only members of those categories – one might want to include emotions or bodily sensations (Klein 2015) as among the intrinsically motivating states, for example. One might also want a theory with more mathematical sophistication. Decision theory, broadly construed, shows how to combine belief-like and desire-like states in optimal ways (Buchak 2016). When considering changes among belief-like states themselves, Bayesianism is an attractive formulation about how an agent should change their degree of belief, consistent with the laws of probability.

Humeanism and Bayesianism work well together. David Lewis, for example, defended a version of Humeanism couched in terms of credences and values. As he puts it:

> Desires are contingent. It is not contrary to reason—still less is it downright impossible!—to have peculiar and unusual desires, or to lack commonplace ones. It may be contrary to the laws of human nature, but those laws themselves are contingent regularities. Likewise there are no necessary connections between desire and belief. Any values can go with any credence. (Lewis 1996, 304)

The result combines the plausibility of traditional Humeanism with the power of Bayesianism. It's a nice package, and has worked well.

Yet within the heart of every philosopher is a tiny Ockham. He urges parsimony. Bayesianism has struck some as *such* a powerful theory that we might be able to get by with it alone. That is, with sufficient cleverness we might be able to do with credences only, rather than credences plus values.

This paper is about one such story, and how it fails.

1.2) Predictive Coding

Consider a simple homing missile. It has a sensor array which can track an infrared source. The steering fins are driven by the angle between the observed position of the source and the center of the field. This tends to bring the source back to the center, thereby eliminating deviations from its course to the target. Constant adjustments, combined with the missile's forward motion and a bit of luck, will lead the missile to impact.

We can fit the Humean story (if we are so inclined) to the missile's behavior. The missile has a *desire* to hit the target. It also has a *belief* about where the target currently is, and conditional on that some beliefs about the best actions to take to get to the target.

However, there's a wholly different way of explaining the homing missile, one closer to the cybernetic roots of its design (Ashby 1976). We could say that the missile *expects* or *predicts* that the heat source will be in the center of its array. The deviation results in a certain *prediction error*. It steers so as to *minimize* prediction error, which eventually brings it to its target. We can describe the whole of the missile's action in these terms as well: it expects to hit the target, and if it's going wrong it is surprised and tries to get rid of surprise by changing something. If it does well, the expected outcome obtains.

This second way of describing the missile, note, only appeals to *one* state-type (predictions) and *one* sort of process (error minimization). The state-type looks much closer to belief than to desire, and works with more complex cases just as well. A more sophisticated missile (e.g.) models the shape of its target too, we could picture it as starting with a guess and then updating its model in response to error. That looks a lot like updating credences in various potential models. So from this perspective, we can unify what looks like two different state-types into one, *contra* Humeanism.

This potential unification is at the heart of *predictive coding* approaches to cognition. Predictive Coding theories (PC) claim that the brain is a mechanism for updating models of the world via minimizing prediction error (Hohwy 2013, Clark 2013, 2015). In its most ambitious form, PC also claims that this is *all* that the brain does.[1]

A key part of predictive coding is *active inference*. Many models of motor control are fundamentally predictive, and use prediction error to guide skilled action; as Clark puts it, "Motor control is just more top-down sensory prediction." (Clark 2015, 21) We guide our actions in part by utilizing predictions about what will happen, and minimizing the mismatch.

[1] I will speak about predictive coding very broadly, to include (for example) accounts that talk about the Free Energy Principle (Friston 2010) or Active Inference (Kirchhoff 2018). While there are important differences within this family, all are committed to a single state type and Bayesian updating, and so will be vulnerable to the critique below.

This strategy generalizes. A mismatch between prediction and the world can be fixed either by updating beliefs or by changing the world. So in Clark's formulation:

> 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. Desires are here re-cast as predictions apt to be made true by action (Clark 2017, 115).

To act, then, you predict that you've *already* obtained the goal state, and then use the mismatch between that and the world to drive adaptive action. As Wiese puts it, "Loosely speaking, this entails a suspension of disbelief in the evidence for an absence of movement.... In other words, we attend away from evidence that we are not moving to enable our predictions to be fulfilled" (2017, 1240).

By doing so, we bring the world in line with our predictions. And, *contra* the Humean, we do so with just one thing.

1.3 What this needs to work

So far we have considered toy examples of a single desire and a single action, and set up the details the way we'd like. For PC to upend Humeanism, we'd have to show that this is possible for the complex set of beliefs and desires that we appear to have.

This means that PC theories must give a systematic, or at least non-*ad hoc*, way of translating between putative desires and some corresponding credences that can do the same work. The typical way this is done is via appeal to credences about the evolutionarily typical states of organisms (see e.g. Hohwy 2013, 85-6). I have argued (Klein 2018) that this strategy is a non-starter. Very crudely, there are a great number of evolutionarily advantageous states that are atypical for individuals (most male elephant seals never mate, but that is not an argument against it), and a great number of evolutionarily bad states that are typical for individuals (most fish end up eaten by bigger fish). This pattern holds across various ways of carving up the reference class for 'typical'.

Dubious evolutionary links are an optional feature of PC, however. PC is anti-Humean, and one might just appeal to various placeholder notions familiar from the anti-Humeanism literature. Perhaps all PC needs is, e.g., a translation scheme which takes us from "I desire X with degree P" to "I believe with credence P that X *is good / is valuable / is worth pursuing*". Cash out the details as seems fit. The upshot is that we get a special class of credences with the right sorts of *bona fides* to be inserted into the Bayesian story. In any case, PC needs *something* along those lines if it is to work. Our question, then, is whether one can find a suitable translation scheme between desires and the right sorts of beliefs.

1.4 The structure of the argument

I'll argue for no. The argument of the paper is conditional: *if* you want PC to be a mechanistic story, *and* you want to be a thoroughgoing Bayesian, *then* you should be a Humean. Most predictive coders want the antecedent and deny the consequent. That's a problem.

The conditional uses a few terms of art. By *mechanistic story*, I mean that the models PC presents are meant to be taken as roughly literal descriptions of the causal-mechanical processes that give rise to perception and action. Given that PC has grown out of cognitive neuroscience and empirically oriented philosophy of mind, this should not raise any eyebrows. Further, as Colombo and Hartmann (2015) argue, Bayesian cognitive theories (of which PC is an instance) are too weak unless read as constraining causal-mechanical models.

I take the requirement for a mechanistic story to be a low bar, but it does imply two important constraints. First, insofar as PC gives a story about *transitions* between states, that ought to imply a story about the causal mechanisms that underlie those transitions. Second, insofar as PC posits explanatory entities (like beliefs or models or whatnot), it is not allowed to posit infinitely many of them. Entities have to be instantiated, and there's only so much room in the skull.

Second, by a *thoroughgoing Bayesian*, I mean that insofar as your credences get updated, they get updated only by conditionalization on the available evidence. We can allow for some wiggle room (cognitive science is messy) and for complex, compartmentalized systems of belief and whatnot. The point is just that if you have some credences and you have to posit some additional, non-conditionalising, way to update them, you're no longer being a good Bayesian.

The conjunction of the two is clearly important to Predictive coders. Here's Hohwy motivating his project:

> In many ways, this broad line of reasoning is the impetus for this book: there is converging evidence that the brain is a Bayesian mechanism. This evidence comes from our conception of perception, from empirical studies of perception and cognition, from computational theory, from epistemology, and increasingly from neuroanatomy and neuroimaging. The best explanation of the occurrence of this evidence is that the brain is a Bayesian mechanism. (Hohwy 2013, 25).

Similarly, Clark presents PC as a mechanism for implementing ideal Bayesian updating to the best approximation we can muster (see esp. 2015 Appendix 1).

The link to Bayes is also important for the grand unifying ambitions of PC. For if PC is right, you can give a theory of mind with a single explanatory bit of ontology (models with credences) *and* a single transition rule (conditionalization). There may be a bit of complexity added by the connections between bits, but in general the grand, unifying ambitions of PC are supported precisely by its link to a relatively austere version of Bayesianism.

There is a very general argument, due to David Lewis (1988; 1996), that a Bayesian should be a Humean. The point of the paper, then, is to spell out this argument in a way that makes it clear that it is a problem for the predictive coder. I'll first go through and give an informal sketch of what motivates the argument. I will then turn to Lewis' more formal version and some of the secondary literature around it. The formalities are important, because they show that the issue is not (just) with particular formulations of PC — the project as a whole faces a serious challenge. Finally, I'll conclude with some reflections on learning, which is at the heart of the problem.

2. The informal version

2.1 Setup

Any organism faces a trade-off between avoiding the bad and learning about the good. A new path might be more efficient or more dangerous. A new mushroom may lead to an awesome Saturday night or an agonizing death. While some basic actions may be hard-wired, a lot needs to be learned from experience. Worse, sometimes things change: the formerly good path becomes home to a hungry lion.

This learning process must take into account two distinct sorts of information: objective, non-relational information about the way the world actually is, and subjective, relational information about the value that different states have for you. Humeans, who separate credence and value can handle this naturally by positing two different sorts of learning processes, each fit for purpose in its relevant domain. So, for example, one might appeal to Bayesian conditionalization for the credence end of things, and reinforcement learning for the value end. Other combinations are possible. The point is just that the orthogonality of credences and values permits wholly distinct learning processes.

Predictive coding, on the other hand, must make do with just credences, and just conditionalization. That creates a fundamental tension: an adequate account of action doesn't permit learning, and vice-versa.

2.2 The necessary Stickiness of Desire

Take Clark's case of drinking water. Let's assume that the relevant proposition is something like

> **D:** When I am thirsty, I drink water

I have an appropriately high credence in **D**. If I am thirsty and I am not drinking, the mismatch between **D** and the world drives me to drink some water.

A few remarks about **D**. It is formulated conditionally because drinking is not unconditionally good. This gives a nice story about the cessation of action. Drinking eventually slakes my thirst, which makes **D** irrelevant.[2] **D** is formulated in terms of what *I* do, because it's only a mismatch between my own actions and the world that can drive my action. What makes it the case that I have a high credence in **D** in the first place might be (e.g.) evolutionary considerations about what things like me do, but my atypicality with respect to my conspecifics isn't enough to drive action. I've put the conditional in terms of actions, but the point should generalize to (e.g.) variational Bayes formulations that talk of internal control states rather than actions (Friston, Samothrakis, and Montague 2012). Finally, I've made **D** as simple as possible for exposition. More complication will not help.

D is meant to drive action. However, this comes with an important caveat. At first glance, if I am thirsty and not drinking, I could do two things to reduce the mismatch with **D**. I could drink water, *or* I could lower my credence in **D.**[3] That is, when I get thirsty, I could just decide that I am not the sort of thing that drinks when I'm thirsty; indeed, my current lack of drinking would seem to provide powerful ongoing evidence against **D.**

However, updating your beliefs is always going to be an easier solution than taking action. On the predictive coding story, changing your belief to "I don't drink when I'm thirsty (and will die of dehydration" isn't obviously wrong – that model is just as accurate, and probably more certain. This is the nub of what's sometimes called the 'dark room problem'. The literature around the dark room problem is confusing because it's often treated as this deep mystery — what reason does and organism have to avoid starving to death in a dark room? – that PC has somehow helped us solve. But properly understood, the Dark room problem is

[2] Here I assume that it is the underlying physiological change rather than the action which slakes thirst; see my (2015; 20ff) for discussion and defence.

[3] Strictly speaking, there are also two more options: I could revise my belief that I'm thirsty, or I could revise my belief that I am not drinking water. I take it that both of these would result in high-error situations and so aren't viable. By contrast, revising **D** is a way of minimizing prediction error, since it is only the conflict with **D** that gives rise to any error in the first place.

just a vivid illustration of the fact that, of two ways to make **D** true, we only ever do one of them. It's not clear, given PC, why that should be.

The only way I can see to avoid this bad outcome is if **D** somehow ends up *sticky*: that is, if it is effectively impossible to update **D** itself. Note that this means there must be a difference between predictions: some of them get updated (the belief-like ones) and some don't (the desire-like ones). In my earlier (2018) critique, I suggested that this breaks the fundamental simplicity of the PC model. Put that to one side. Also put aside worries about how the two states are reliably distinguished; presuppose for now whatever magic you'd like. Perhaps **D** gets arbitrarily high credence, or precision, or whatever you need to make it sticky. Assume whatever is necessary, so long as it is consistent with a mechanistic, thorough Bayesian story.

2.3 A sticky problem

Now comes the problem: If **D** can't be updated, ***D*** *can't be updated*. That means my experiences can't actually change what I value. That seems wrong.

Suppose the water in my area becomes contaminated—enough to make me a little sick, and that previously unattractive pineapple juice now becomes a better option to quench my thirst. It is unclear how this fact alone would even come to bear on **D**, but ignore that.

D is sticky, which means that it is resistant to evidence. So whatever evidence is supposed to be a valid reason to change **D** won't do the job. Indeed, note that the evidence that the water is bad—which is ultimately supposed to bear on **D**—is surely going to be somewhat imprecise and inconsistent. In terms of evidence against **D, then,** it will be much weaker than the extremely strong sensory evidence I have that I am not drinking water when I am not drinking water. Which means that if **D** is sticky enough to drive action, it's too sticky to revise.

There are two tempting ways to get out of this, neither of which PC can actually endorse. First, you could think that **D** is just too simple: that in fact, the right analogue is something like "When I am thirsty *and* I do not have previous evidence that the water is bad *and* …. I drink." Now, that's a super weird response. It wriggles out of the problem by denying that you actually learn anything about what's good (you always know all the relevant conditionals, and all you ever learn is which antecedent to apply.) But even if you're one of the rare philosophers of cognitive science who have a fondness for the Platonic Doctrine of Recollection, it still won't do. That ellipses hides a lot. The possible combinations of states that count for or against drinking are effectively infinite, which means in turn that the number of distinct specific states in which I drink will be infinite. So while complicating **D** might result in a good *description* of how I act, it can't specify the *mechanism* by which I act. But that's what PC needed.

Second, it's tempting to think of more elaborate schemes about how to update **D**. Why not (say) just add a rule that lets you update **D** in case you get evidence that the water is bad? The problem is generated because of the inflexibility of the updating rule, after all. But remember, the inflexibility is a feature, not a bug. Add a rule that doesn't involve updating by conditionalization, and you're no longer a thorough Bayesian. But that's what PC needed.

So it's not that there aren't solutions. It's just that the obvious solutions involve either giving up on PC being a mechanistic story or else giving up on the thoroughgoing Bayesianism. That secures the conditional argument

2.4 A further perspective

Before moving to Lewis, I want to offer a further perspective on the informal argument. One of the standard arguments in favor of Humeanism is supposed to be that beliefs (and the like) just aren't the sort of thing that can motivate. That is, there is a *synchronic* problem getting things with a belief-like direction of fit to do the job they need to do. That's arguably a piece of folk psychology, and PC abandons it in the course of showing how a belief-like state can be made to do the right kind of work. I think that's fine: good models can trump folk psychology.

But what the above shows is that there is a much stickier *diachronic* problem that the anti-Humean faces, one that bites PC especially hard. It's not just that credences and values appear to do different things. They also require being updated in different, often orthogonal, ways. This is related to direction of fit worries: one should update a belief when it becomes false, and a value when it becomes bad for you. Yet it is a distinct, and arguably more difficult, problem to solve. It is especially hard to solve if, like predictive coders, the attraction of your theory rests on the unificatory power of a single, austere, learning rule.

Finally, and again, it's worth reiterating that the Humean has a lot easier time here. For the Humean already has two distinct state-types, which means they can easily appeal to two distinct rules. Neither the synchronic nor the diachronic problem get much purchase; what the

Humean loses in ontological parsimony, they more than gain back in simplicity of the overall dynamics.

3 Lewis' general argument.

3.1 The setup

The above argument was linked to a particular way of cashing out things like **D**. PC is a broad tent, and perhaps you favor some other way of linking up value and credence.

There is a very general argument by Lewis (1988,1996) that appears to weigh against *any* attempt to reduce values to credences. This has received surprisingly little attention in the Predictive coding literature (Fazelpour, Ransom, and Mole (2017, footnote 11) is the only mention I've found). That's a pity. I think that, properly understood, Lewis' argument shows that the informal argument in section 2 will hold regardless of how you cash out the particulars.

I'll start with Lewis's argument, then look at some responses to tie it all together. I follow the presentation in (Lewis 1996). First, let's suppose each individual can be described as having two functions: *V*, which takes propositions to their value to the individual; and *C,* which takes propositions to an individual's credence in the proposition. The Humean thinks that this description holds because there are distinct desires (that ground *V)* and beliefs (that ground *C*).

The job for the non-Humean is to show a principled way to translate from *V* to *C.* That is, for any proposition *A*, we want a mapping function that takes us from *V(A)* to a belief that has the right motivational role. We did that in an ad hoc way above when we went from "I desire to drink water when I'm thirsty" to "When I'm thirsty, I drink water." But that translation could take any number of forms, so long as it is consistent – we might posit a high credence in "Drinking water is good" or "Drinking water is the thing to do" or whatever. I'll cash this out as "*A* is good" below, but that's shorthand for a bunch of different possibilities.

Using Lewis' terms, PC must posit a *halo function* that maps any proposition *A* onto a corresponding A^0 such that $V(A) = C(A^0)$. What Lewis terms the *Desire-as-Belief* (DAB) thesis is just the claim that some such function exists. This is not yet to give a mechanism, but merely to posit a systematic mapping; conversely, the Humean *denial* of DAB is the claim that there is no such systematic mapping.

Now, this is something of an odd setup by many lights. DAB as stated collapses goodness into two states—good or bad, halo-on or halo-off—and altering this to allow for degrees of goodness and badness requires a more complex account (See Hájek 2015 for an excellent discussion). There's also a question about how this should apply to actual reasoners. People have intransitive preferences: that is, they can prefer A to B, B to C, and C to A (Tversky 1969). Yet one cannot have an intransitive credence function.

While these may be difficulties for Lewis' account considered quite generally, I submit that they are *also* issues for PC. That is, it is also unclear quite how the PC framework deals with degrees of value, given that most of the framework is spelled out in terms of the *optimal* action for the agent in question (see, e.g., Friston, Thornton, and Clark 2012). I think DAB is actually a relatively good picture of how PC envisions the translation from value to credence; let's assume, for the sake of argument, that it is fit for purpose.

3.2 The argument

As a logical matter, DAB is equivalent to the conjunction of two other propositions:

Desire as conditional belief (DACB): $V(A) = C(A^0| A)$

Independence *(IND):* $C(A^0| A) = C(A^0)$

DACB says, roughly, that your values should link up to the credences that something is good conditional on you actually having that thing. Put that to one side; IND is where the action will be. IND says that your belief that *A* is good should be independent of whether *A* is actually the case. That seems like a good general rule of thumb. IND is also necessary for the predictive coder. It is the mismatch between A^0 and *A* that drives action, so that mismatch needs to be preserved in the face of evidence that what is good does not (yet) obtain.

Yet as Lewis notes, IND does not generally hold. Indeed, there are clear counterexamples. Suppose *A* and $A^0 > 0$, and that I learn that $\sim(A \ \& \ A^0)$. Then $C(A^0| A) = 0$, but $C(A^0) > 0$. IND fails.

To make the case concrete, suppose I've heard that sit-ups are good for you (and I want to do anything that's good for me). I don't really know what I'm doing, but I dutifully attempt some. I pull my back. I conclude that either sit-ups aren't good for you, or else I wasn't actually doing sit-ups. In that case, the value I place on sit-ups conditional on me having done them is zero: I tried, and I got hurt. But I am unsure whether I really did sit-ups, and so I'm unsure of whether sit-ups are as bad as they seem. So my value on sit-ups *tout court* remains nonzero. Which means that the credences doing value-like work are not independent of the credences that are supposed to drive action, which is what the predictive coder needs.

DAB says that there's a mapping from (apparent) values to credences. That's the halo-function. That mapping needs to be specified independently of what's actually the case, so that value-like credences can drive actions. But then updating by conditionalization will, in many cases, break that mapping. So contra DAB, we can't map values to credences after all.

3.3 A refinement.

There has been considerable discussion of Lewis' result. As Hájek and Pettit (2004) note, an important class of responses can be captured in terms of quantifier scope. Making the quantifiers more explicit, Hájek and Pettit contrast two versions of DAB:

> Lewisian DAB: There is a halo function such that for any pair of C and V and any proposition A, $V(A) = C(A^0)$

In other words, there is a single, fixed halo function that holds across shifts in credence. This is arguably how PC actually sets things up (or, at least, this is the most plausible way to read predictive coders' frequent appeal to an organism's evolutionary history). Hájek and Pettit agree that Lewisian DAB is untenable. However, they note that by re-ordering the quantifiers, we get the much more plausible:

> Indexical DAB: For any pair of C and V and any proposition A, there is a halo function such that $V(A) = C(A^0)$

Indexical DAB is much easier to satisfy, because it allows A-halo to vary as a function of circumstance. Intuitively, this allows for the possibility that an agent may (e.g.) learn that they were wrong about what is good (2004, 83). The most plausible way of cashing this out in the ethical case involves 'indexicalist' formulations, on which the halo function tracks something indexed to the agent. Similar responses involve separating (e.g.) evaluative and non-evaluative propositions (Bradley and List 2008).

Lewis himself seems to think that letting the halo-function vary would be a cheap victory. Hájek (2015) notes that this is only if the halo-function is allowed to vary arbitrarily, so long as there is an interpretation on which the halo over a proposition is "genuinely earned" (p440). I take this to mean that the halo function and its dynamics needs to reflect something about how we think the corresponding predicate (like 'is good') behaves.

But here we come to the crux of the problem for PC. The debate around Lewis is primarily about *whether* an appropriate mapping from *V* to *C* can be found. PC needs something more: it needs to tell us *how* that updating can take place. The lesson of either form of DAB, I take it, is that however this works, conditionalization is out of the question. So if there's a mechanistic story, it's not a thoroughly Bayesian one: something other than conditionalization has to be in place to keep the halo function happy.

But that's just to say that's just to say that the predictive coder can't have a mechanistic story that's also fully Bayesian about change in credences – unless, of course they want to posit some other states (like desires) that underwrite the remapping of the halo function across updates in credences. Which is, again, to secure our conditional.

4. Conclusion: Change matters

I suggested at the end of section 2 that the deep problem for the predictive coder has to do with diachronic change in values. If you want to give a mechanistic story (and PC should), and you don't want to give up on thorough Bayesianism about credence change (and doing so would make PC unattractive), then you need to posit other processes to keep this running. You need to do this, note, *even* if you think that action in the particular case is driven entirely by mismatch with credence-y things. In section 2, I argued that this mismatch only works if you can keep the value-like credences fixed, which means you need some other mechanism for them to change when they need to change. In Section 3, I gave Lewis' argument that no matter how you set this up, you can't get by with just conditionalization on one's credences. Something has to give. And that, you might think, is a good place to return to desires, or values, or something old-fashioned – not because you need them for action, but because you need them to learn which actions are best.

This problems with PC have remained obscure, I think, because of the belief that certain modelling results show that the PC framework can operate without a formal value function. So, for example, in criticizing Ransom et al. (2017), Clark claims that objectors who focus on the initiation of action fail "…to recognize the true scope of the formal demonstration that any set of behaviors prescribed by reward, cost, or utility functions can be prescribed by an apt set of systemic beliefs or priors." (2017, 117). Yet the 'apt' part does more work than it might seem, and more work than it should.

The modelling results that Clark refers to (such as Friston, Samothrakis, and Montague 2012) typically start by fixing the value function, and then showing that optimal policies with respect to the value function can be learned. These are usual quite sparse problems as well: in the case of Friston, Samothrakis, and Montague, for example, they consider a mountain-car problem that has a single, fixed goal and in which the agent has already "learned the constraints afforded by the world it operates in." (2012, 533).

Here's one way to interpret such claims: if you give me an initial credence function *C* and a non-updating value function *V*, along with complete flexibility about how to set the actual credences (interpreted as a lack of constraints on the halo function), I can find a total set of credences *C** that allows an agent performing active inference to act as if they were behaving in accordance with the initial *C* and *V*.

Yet what the above has shown is that even if we can do this, and even if we are convinced that this is not mere description but picks mechanisms, we're *still* missing a story about how the valued propositions change. The mountain car doesn't have to deliberate about what's good in life: it has to get to its spot, and that's it. Keeping that fixed, learning about what *means* will get it to its end is unproblematic.

Our lives are not so simple. We must learn about what is good, and we must re-learn what is good changes when the world changes. A very natural way to model this is to separate out learning about the world and learning about value, and to treat these as distinct but co-equal processes that fruitfully interact. PC can remain an interesting and valuable part of the story about how beliefs are updated if you'd like—there's just more work to do too.

Humeanism is not without its flaws. However, positing two independent state types—with different evolutionary demands, different directions of fit, and different combinatorial resources—brings along a host of explanatory resources. Whatever marks of simplicity Humeanism loses by having two state-types rather than one, it gains back and more on the simplicity and empirical plausibility of the resulting explanations it can give. [4]

[4] Thanks to Peter Clutton, Ben Henke, Julia Haas, and Ross Pain for helpful comments on a previous draft, and to an audience at Macquarie University for useful feedback in the early stages of the project. This paper was supported by Australian Research Council grant FT140100422.

References

Ashby, W. R. (1976). *Design for a Brain; the Origin of Adaptive Behavior.* Chapman and Hall.

Bradley, R. and List, C. (2008). Desire-as-belief revisited. *Analysis*, 69(1):31– 37.

Buchak, L (2016) Decision Theory. In *Oxford Handbook of Probability and Philosophy* eds. Christopher Hitchcock and Alan Hájek. Oxford, Oxford University Press: 789-814.

Clark, A. (2017). Predictions, precision, and agentive attention. *Consciousness and cognition*, 56:115–119.

Dennett, D. (1987). *The Intentional Stance*. MIT Press, Cambridge.

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., Samothrakis, S., and Montague, R. (2012). Active inference and agency: optimal control without cost functions. *Biological cybernetics*, 106(8-9):523–541.

Friston, K., Thornton, C., and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in psychology*, 3:1–7.

Hájek, A. and Pettit, P. (2004) Desire Beyond Belief. *The Australasian Journal of Philosophy* 82(1):77-92.

Hájek, A. (2015) On the plurality of Lewis's triviality results. In *A Companion to David Lewis.* Ed. Barry Loewer and Jonathan Schaffer. New York: John Wiley & Sons, 425-445.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press, New York.

Kirchoff, M. et al. (2018) The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The royal society interface* 15(138): 20170792

Klein C (2015) *What the Body Commands: The Imperative Theory of Pain.* Cambridge: MIT Press.

Klein, C. (2018). What do predictive coders want? *Synthese*, 195(6):2541–2557.

Lewis, D. (1988). Desire as belief. *Mind*, 97(387):323–332.

Lewis, D. (1996). Desire as belief II. *Mind*, 105(418):303–313.

Ransom, M., Fazelpour, S., and Mole, C. (2017). Attention in the predictive mind. *Consciousness and cognition*, 47:99–112.

Smith, M. (1984). *The Moral Problem*. Cambridge: Blackwell.

Tverskey, A. (1969) Intransitivity of Preferences. *Psychological review* 76(1): 31-48.

Wiese, W. (2017). Action is enabled by systematic misrepresentations. *Erkenntnis*, 82(6):1233–1252.