Sub-Optimal as Optimal:

The Unfalsifiability of a Unified Theory of Bayes-Optimal Predictive Brain Function

Introduction

Inquiry into the role of probabilistic inference in brain processes is at least a 150 year old project, beginning with Helmholtz. Neuroscientific research has amassed enough structural and physiological data to suggest compelling possibilities for the actual biological instantiation of predictive processes. This has given rise to explanatory theories of various scope and complexity.

Problematically, some theories propose that the brain is unified as a "prediction machine" or "inference machine," or a "Bayesian brain."[1] The philosopher of cognitive science Andy Clark writes extensively about a "unified science of mind, brain, and action," (2013) made possible by the theoretical hierarchical Bayesian predictive coding (PC) framework. Many different terms exist to refer to this notion, so to simplify the discussion, this paper uses the term *unified theory*. This should be interpreted as the notion that the brain is a unified engine of hierarchical Bayesian predictive processing.

The unified theory (UT) of brain function is a shaky construction. Clark and Karl Friston claim that evidence for Bayes-optimal predictive coding and error-correction in *perception* can be extended to support the claim that action and higher cognitive functions operate by the same neuro-computational mechanisms (e.g. Clark, 2013; Friston, 2010). They have extended—to a precarious height—a theory of perceptual processing that was initially developed for machine-learning.[2] Regarding neural instantiation, we have only inconclusive indirect evidence.

The UT proposes that the brain is a Bayesian prediction machine that weighs incoming data with prior experience to make optimal inferences about the world. However, as this paper argues, the extraordinary complexity of our brain allows us to *internally generate evidential data*, which enables a Bayes-optimal PC explanation of sub-optimal psychology and behavior. Learned phobia, such as a fear of flying, is an example of this. Instead of this being a strength of the theory, the unconstrained explanatory power of the Bayesian predictive coding framework is an indication of its weakness. A scientific theory should make bold and specific predictions that allow for empirical observation to falsify it. This cannot yet be done with the UT, thus it is not yet scientific. Rather than concern ourselves with grand unification, our efforts should be toward garnering direct evidence *against* risky and testable theoretical

1 Hohwy (2013), Friston (2010), Clark (2013), respectively.
2 See "The Helmholtz Machine," P. Dayan et al. (1995).

predictions. This is the scientific methodology argued for by Karl Popper.[3]

This paper is organized as followings: the first section explains the basics of PC; the second section presents compelling indirect evidence for it and common criticisms of it; the third section demonstrates that there is a theory of unified brain function; the fourth section makes the case that phobia is an example of sub-optimal psychology that can be explained in terms of Bayes-optimal PC; the fifth section argues that the UT does not allow for falsification; the sixth section suggests ways that the UT can define testable theories of PC to guide neuroscientific research; the seventh section offers possible rebuttals to the arguments herein.

1. Predictive Coding

The PC framework goes by various names, including hierarchical predictive coding (Rao & Ballard, 1999), free-energy minimisation (Friston, 2007), prediction error minimization (Hohwy, 2013), and action-oriented predictive processing (Clark, 2013). PC encompasses many different versions of specific models of mental activity. If a model has the following components, then it is a PC model: hierarchical brain organization and bi-directional signal flow; predictive coding and error signals; internal generative models based on probability density distributions encoded by populations of neurons; and conditional probability, often Bayesian.

Neurophysiological research has revealed the brain to be functionally organized. Areas of closely related functions, such as those involved in a particular sensory modality, are arranged in hierarchies. Importantly, the flow of information through a hierarchy is bidirectional, meaning signaling flows upward and downward through the system (or, synonymously, forward and backward). At higher cortical levels, the hierarchical structure may be considered more horizontal than vertical, and signal flow may be multi-directional.

PC began as a theoretical response to the question of why there is so much downward signaling in perceptual systems. In the lateral geniculate nucleus of humans, for instance, approximately 80% of the incoming signals are from the primary visual cortex, the next higher level of the visual system (Bear et al., 2016). The PC explanation is that downward signals are prediction signals, whereas upward signals are either the incoming raw sensory data, or error signals produced when an upward-flowing signal meets with a downward-flowing prediction signal. Mismatch in the signals causes an error signal to propagate upward where it then instigates revision of the prediction. When error is minimized, there is minimal upward flow of information. According to PC, a percept is an optimized prediction about what is most likely being encountered in the world.

At each hierarchical level, populations of neurons encode a generative model. The higher the

3 See *The Logic of Scientific Discovery*, Popper, K. (1934).

level, the more general the model. Generative models statistically simulate observable data based on probability functions, thus neural populations encode conditional probability distributions that are shaped by experience. A prediction signal is a probabilistic inference based on the probability distribution of a generative model at a particular hierarchical level.

Hierarchical Bayes networks are often used to implement PC computation. In the Bayesian approach, Bayes' theorem[4] describes the process of weighing incoming data with prior experience. Bayes' theorem says that the probability that a particular hypothesis is true given the data (the posterior probability) is equal to the probability that one would see those exact same data if the hypothesis were true (the likelihood) times the probability that the hypothesis is true (the prior probability), divided by that same product for all other possible hypotheses that could explain the data, i.e. the sum of all other hypotheses given the data times the probability of each of the hypotheses.[5]

In the PC framework, hypotheses are considered predictions in the computational process of downward-flowing signals, and incoming data are the hypotheses for upward-flowing signals. The posterior probability at the upper level is the prior probability at the lower level. The prior probability and the likelihood of a hypothesis at each hierarchical level are derived from the generative model at that level. Arriving at an optimal state, such as a percept or belief, entails optimizing (maximizing) posterior probabilities.

2. Evidence and Criticism

There is compelling indirect evidence for PC. The evidence comes in various forms. Bayesian optimality may be explicitly implemented in a computer system that is designed to employ PC. The predictions made by such a system about a particular event—such as movement on a screen—may then be compared to predictions made by human subjects about the same event, which can be remarkably similar (e.g. Weiss et al., 2002). Indirect evidence for the biological plausibility of hierarchical PC has come from studies in which a PC system self-organizes to become structurally similar to a hierarchical system of the brain (e.g. Rao & Ballard, 1999). Alternatively, a mathematical model of possible predictive computation may be compared to experimental data; results from experiments on object-word acquisition in children are demonstrated to fit a Bayesian inference model (Xu & Tenenbaum, 2007b). Experiments using animal models have shown that the primary visual cortex (V1) shows less activity over a developmental period in which animals are trained to a particular type of visual stimulus (e.g. Berkes et al., 2011). This is considered indirect evidence of decreased surprise in V1, thus generative model optimization. Functional imaging studies in humans show decreased V1 activity when the onset of movement on a screen indicates its trajectory, i.e. when movement

4 $p(h_i|d) = p(d|h_i)p(h_i) \,/\, \sum_{h_j \in H} p(d|h_j)p(h_j)$

5 To avoid self-plagiarism: I used this sentence in my previous summary paper. It is my best attempt at a precise literal translation of the theorem.

is highly predictable (e.g. Alink et al., 2010). The reason for this might be that easily predicted movements require less predictive processing.[6]

There are criticisms of the PC framework. Regarding the Bayesian computation component, Marcus and Davis (2013) argue that in experiments aimed at revealing Bayesian inference, theory-confirming tasks are too often selected, and results are not being reported when tasks are not theory-confirming. More germane to the arguments of this paper is the issue of model selection, or the post hoc selection of prior probabilities and likelihoods. The priors and likelihoods of Bayesian models are crucial to the predictive success of the model, thus their selection can dramatically affect how well the model fits the behavior of test subjects. Bowers and Davis (2012) argue that "there are too many arbitrary ways that priors, likelihoods, utility functions, etc., can be altered in a Bayesian theory post hoc." Marcus and Davis echo this:

> Without independent data on subjects' priors, it is impossible to tell whether the Bayesian approach yields a good or a bad model, because the model's ultimate fit depends entirely on which priors subjects might actually represent.

This means that the models chosen might only be those that support the theory that human behavior is Bayes optimal, despite the fact that other similar but less supportive models could have been chosen.

These criticisms point to an issue with the Bayesian framework. It is an issue of constraints, or lack thereof. The posterior probabilities that result from incoming data can differ extremely if the prior probabilities or likelihoods are different. To make convincing Bayesian models—models that seem to produce the same posteriors that people do—inductive constraints are necessary. However, without knowing the internal constraints in a particular person, or in humans in general, we *have* to make them up. This does not mean that the Bayesian framework is inappropriate, but it does mean that we need to acknowledge the weakness of a general theory that lacks the ability to make precise predictions without post hoc manipulation. The need for manipulation can be diminished if we can determine the neural implementation of the various aspects of the Bayesian PC framework. For instance, determining which constraints are learned and which are innate at a particular hierarchical level would help to guide research in the right direction.

3. Unified Theory

Clark describes a unifying framework called the "hierarchical prediction machine approach," though as of 2013 he prefers the name "action-oriented predictive processing." In a critical response to Clark's 2013 paper, Anderson and Chemero (2013) somewhat derogatorily dubbed his unifying attempt the "Grand Unified Theory (GUT) of Brain Function." To avoid

6 For a longer list of examples, see Clark (2013).

the derogatory undertone, this paper uses "unified theory" instead.

Before further discussing the weakness of the UT, it is necessary to further reveal the existence of a UT of brain function. Clark's UT is based on Friston's work and ideas from computational neuroscience. Clark (2013) writes:

> Recent work by Friston…generalizes this basic "hierarchical predictive processing" model to include action. According to what I shall now dub "action-oriented predictive processing," perception and action both follow the same deep "logic" and are even implemented using the same computational strategies. A fundamental attraction of these accounts thus lies in their ability to offer a deeply unified account of perception, cognition, and action.

This demonstrates that PC is no longer only relegated to sensory systems, but is now also "generalized" to include motor and cognitive systems. Regarding action, Clark claims that motor commands enact predictions about what movement the body will make next. In Friston's (2003) words:

> In motor systems error signals self-suppress, not through neurally mediated effects, but by eliciting movements that change bottom-up proprioceptive and sensory input. This unifying perspective on perception and action suggests that action is both perceived and caused by its perception.

Regarding cognition, Clark is an incrementalist. He proposes that "you do indeed get full-blown, human cognition by gradually adding 'bells and whistles' to basic (embodied, embedded) strategies relating to the present at hand" (2014). In his 2013 paper, he writes:

> Importantly…hierarchical predictive processing models now bring "bottom-up" insights from cognitive neuroscience into increasingly productive contact with those powerful computational mechanisms of learning and inference, in a unifying framework able (as Griffiths et al. correctly stress[7]) to accommodate a very wide variety of surface representational forms.

His stance is that we may be able to explain *all* brain functions by merging machine learning strategies like self-organizing neural networks with generative Bayesian models of rationality and inductive inference, and then demonstrate how they are neurally implemented. He argues that Friston has achieved the theoretical framework for this, and that a wide range of studies have provided indirect evidence for the tenability of such a UT of brain function across the full spectrum of human mental activity.

4. Internally-Generated Data

The UT's Bayesian PC framework rests on the notion that all brain processes continually

7 Griffiths and his frequent collaborators, including Tenenbaum (mentioned above), primarily work on computational Bayesian models of higher cognitive functions.

converge upon Bayes-optimal predictions, thus higher cognitive acts are also at least approximately Bayes optimal. Furthermore, our predictive processing becomes more accurate through repeated exposure to the statistical regularities of the environment. This is what shapes the probability density functions of the generative models that underpin our predictions. Therefore, a mature adult who is fully acquainted with the likelihood of a particular event occurring should usually be able to make accurate predictions about it. The extent to which we make accurate predictions when we have enough experiential evidence to do so is the extent to which we are considered rational, at least colloquially speaking; "you should have known better" is a common admonishment. Irrational behavior may be considered sub-optimal, in that rational behavior optimizes our chances of success in most circumstances but irrational behavior does not.

To ground this with a familiar example, consider the case of being afraid to fly on a plane. Most adults are aware that planes occasionally crash, but that car crashes are much more common. Therefore, we should feel more assured of our safety as a plane passenger than a car passenger. To remind each other of this fact, it is often relayed that "you are much more likely to die in a car crash than in a plane crash." Some people are even aware of the measured statistical likelihood of dying in a plane crash versus a car crash. Despite all of this, some adults have a fear of flying—they know a plane crash is unlikely, but they are afraid of it anyway. People who are too afraid to fly across the country might choose to drive instead—a much riskier decision, and arguably a much less rational one.

In the psychology of heuristics and biases,[8] irrational fears are often a case of the availability heuristic: if a memory is salient, such as memories of news stories about frightening plane crashes, then the likelihood of an event occurring might be deemed far higher than the actual statistical likelihood. Though this explanation alone would not satisfy a behavioral neuroscientist, it does seem to describe the thought process that leads to an irrational fear. What it does not explain are cases when no amount of evidence can correct the bias. An irrational fear that cannot be corrected by statistical evidence is a phobia.

Regarding learned phobia, behavioral neuroscience studies have shown that experiences of pain and stress can condition fear responses, and that the amygdala is a structure consistently involved in mediating emotional response across species, particularly fear. Explaining learned phobia requires describing the formation and strengthening processes of the neural circuitry that connects the amygdala and sympathetic nervous system to the parts of the brain involved in memory and cognitive assessment.

What is curious about the case of a person who learns an irrational fear of flying is that they may never have had a more negative experience of flying than hearing about cases of a plane crash. The Bayesian PC explanation for this can be formulated as follows. A fear of flying, not

8 See *Judgement Under Uncertainty: Heuristics and Biases*, by Kahneman, D., Slovic, P., and Tversky, A. (1982).

allayed by awareness of the statistically minimal likelihood of a crash, is caused by *internally-generated evidential data*. The repeated mental process of imagining fear-inducing scenarios has the same effect that the actual experience of those scenarios would have. This means that the probability density distributions, which should correspond well to the real-world, have been distorted by overwhelming internally-generated data.

In the language of Bayes theorem, the probability that a person will think that a plane will crash given that they are imagining the plane crashing (the posterior probability) is proportional to the probability that the person is imagining the plane crashing given that the plane will crash (the likelihood) times the probability that the plane will crash (the prior probability). The posterior is passed down (or horizontally, in the case of higher levels) to be the prior in the lower-level computation, but if this computation constantly results in an error signal, then the posterior probability will continually increase until it reaches a sustained cognitive state of certainty that the plane will crash.

In the case of phobia, a perpetual error signal results from the internally-generated sensory data that a plane crash is certain. When a prediction that the plane will *not* crash meets the internally-generated data saying that it *will* crash, an error signal is produced, which then adjusts the probability density distribution at the higher level, which adjusts the generative model, which results in revised predictions, thus higher posterior probabilities. Therefore, phobia is a positive feedback loop in circuitry involving neural predictive coding populations in the amygdala and the higher cortical areas responsible for cognitive assessment. This would be a plausible explanation for perpetually incorrect belief formation using the Bayesian PC framework of the UT.[9]

5. Unfalsifiable Theory

The above Bayesian PC rationale for phobia might be criticized by proponents of the UT, but it would not be criticized for being attempted. By claiming that the brain is a hierarchical prediction machine or a Bayesian inference engine, we are encouraged to use the same basic rational to explain any brain functions, even apparently sub-optimal cases like mental illness. In the case of mental illness, Friston has done just this (albeit without much depth). He writes:

> The basic message here is that a fundamental failing of predictive coding mechanisms may underpin many neuropsychiatric disorders, particularly those that involve complicated or difficult Bayesian inference problems that predictive coding tries to solve. If this is the case, one might expect empirical evidence for failures of predictive coding at all levels of the hierarchy… (Friston, 2012).

In the above account of phobia, the idea of internally-generated evidential data is compliant

9 For the case of delusions, see "Unraveling the mind," Gerrans, P. (2013).

with the loose constraints of the UT, yet very problematically allows for the explanation of *sub-optimal* psychology in a supposedly near-*optimal* neuro-computational system. Therefore, what may seem like falsifying evidence—sub-optimal psychology in an optimal system—is actually evidence that can be absorbed by the theory, or by clever adjustments to the theory. To put it another way (and to reiterate points made above), post hoc or "arbitrary" (Bowers & Davis, 2012) selection of likelihoods and priors in a Bayesian model render the model unscientific: if it can always be adjusted to explain or avoid contrary evidence, then it cannot be falsified.

As it stands, the UT is reminiscent of Freudianism in its heyday: it seems that any function or condition can be explained by the Bayesian PC framework. This, as Popper argued, is not a strength. For the UT to become more scientific, it must be clear what its specific predictions are, what evidence would falsify those predictions, and what experiments might garner that evidence.

To further strengthen the claim that the UT is not falsifiable, consider the excellent argument that Spratling (2013) makes in response to Clark's 2013 paper. Spratling points out that more than one set of PC neural mechanisms fit the indirect evidence we have for the general framework. He writes:

> …claims…that prediction neurons correspond to pyramidal cells in the deep layers of the cortex, while error-detecting neurons correspond to pyramidal cells in superficial cortical layers, are not predictions of PC in general, but predictions of one specific implementation of PC. These claims, therefore, do not constitute falsifiable predictions of PC (if they did then the idea that PC operates in the retina…could be rejected, due to the lack of cortical pyramidal cells in retinal circuitry!). Indeed, it is highly doubtful that these claims even constitute falsifiable predictions of the standard implementation of PC.

This argument opens up many avenues of criticism. Not only are Friston and Clark's claims about the different encoding roles for deep versus superficial pyramidal cells in the cortex not a prediction that allows us to falsify the "standard implementation of PC" (Spratling's term for the UT), it reminds us that predictions are not being made about the numerous other types of neurons (and glial cells) in the brain, or for the cytoarchitectural differences that define Brodmann's areas, or differences in the cerebellum, midbrain, and brain stem—or more importantly, how all of this complexity is actually unified by the same coding framework. The UT should clearly state what we should expect to be the different roles of these features, how we should determine if they in fact fulfill those roles, and how evidence that those roles are not fulfilled falsifies the UT.

Though it may be true that the brain can be fully explained at a mesoscopic level by a relatively simple rationale, it is not scientifically fruitful to practice applying that rationale if we cannot demonstrate the ways it might fail to explain brain processes. As Popper (1935)

writes:

> Bold ideas, unjustified anticipations, and speculative thought, are our only means for interpreting nature: our only organon, our only instrument, for grasping her. And we must hazard them to win our prize. Those among us who are unwilling to expose their ideas to the hazard of refutation do not take part in the scientific game.

6. Testable Theory

Though indirect evidence should certainly not be discounted as insufficient for science, neuroscience has the means[10] to start systematically garnering direct evidence for the neural mechanisms of Bayesian PC. Nevertheless, according to Clark (2013) and Enger and Summerfield (2009, 2013), there have been few studies to this end. While the UT does propose that there should be separate populations of neurons encoding prediction and error signals, and though it waves vaguely to deep versus superficial pyramidal cells in the cortex, Spratling points out that this degree of prediction specificity may not be enough to guide scientific inquiry toward potentially falsifying direct neural evidence. And this is not the only prediction lacking. The following are a few other examples of the types of predictions a scientific UT of brain function should make.

One very helpful set of predictions would be what exactly the markers for a predictive neural system are. For instance, how do we discern what is definitely *not* a system employing predictive processing (or more specifically, Bayesian computation, error minimization, etc.)? The assumption seems to be that all *mammals* employ PC mechanisms, but the argument has not been made that simpler animals do not. Given the ethical prohibition of invasive testing on humans, the technological limitations for gathering sufficient evidence through non-invasive means, and the financial constraints on research, it behooves us to determine if a prediction can be tested in a very simple animal model, and how simple the animal can be. If we can conclude that all extant neural systems are hierarchical Bayesian PC systems, then we should go straight to the simplest neural systems for experimental purposes. For example, *C. elegans* might be an ideal candidate given that its entire 302-neuron nervous system has been mapped (as well as its complete genome), but not if it is far too simple of a system to allow for PC experimentation. Unfortunately, the UT lacks a testable prediction regarding this basic question of how to distinguish between a non-PC system and a PC system.

It is also crucial to know how we should parse the brain into hierarchies for testing purposes. In the human visual system, this seems obvious, at least at lower levels. For more complexly integrated neocortical areas such as the frontal lobe, it is not clear what the UT would define as a hierarchical level. Predictive estimator populations are proposed to be separated into distinct hierarchal units, but to test the theory we need to first define where exactly those

10 For an overview of relevant emerging technologies, see "Using Optogenetics and Designer Receptors Exclusively Activated by Designer Drugs (DREADDs)," Fowler, C. et al. (2014).

populations begin and end. The brain is packed with neurons and neuroglia—images of neuropil reveals that the brain is not a tidy place—so strict guidelines are necessary. Otherwise, any finding contrary to the theory can be written off as the result of an inaccurate test. More importantly, applying techniques that activate, deactivate, or remove specific neurons or neural populations (e.g. DREADDs or optogenetics—see note 11) to a specific hierarchical level would be a tremendously useful method for testing the way system-level predictions (e.g. percepts, beliefs) are affected by the absence of a particular generative model.

Moreover and more specifically, much work has already been done in the mesolimbic "reward" system to reveal the development of anticipatory signals in the mechanisms of conditioning and addiction.[11] Undoubtably this can be incorporated into the UT framework, but apparently so can any brain functions. With dopaminergic systems in mind, UT should make testable predictions. Though Clark (2013) very briefly mentions dopamine neurons, he does not outline what to expect of the role of dopamine neurons and the various dopamineric pathways in predictive processing. Friston (2010) seems to suggest that dopamine neurons are a type of prediction neuron specialized for reward prediction, but it is not clear if, in a similar fashion, he would predict that all prediction neurons are specialized in some way. Such a prediction would be falsifiable, and may lead to important understanding of other specialized prediction neurons in the brain.

Rather than continue listing what needs to be clearly predicted by a scientific UT of brain functions, suffice to say that efforts to expand the original PC framework to include ever more complex features of cognition—language[12], psychopathology[13], social psychology[14], etc.—as exciting as the activity may be, should be curtailed and refocused on guiding scientific experimentation.

7. Possible Rebuttals

The above arguments—that the proposed UT of brain function allows for an explanation of sub-optimal psychology as optimal, thus lacks proper constraints, hence is not yet falsifiable, therefore is not scientific, so is not scientifically fruitful—can be rebutted in various ways. This final section acknowledges five possible rebuttals.

First, against the criticism of Bayesian computational models being under-constrained or manipulated pos hoc, one might argue that *any* model can be criticize for these reasons. The very definition of a computational model might be that is it fake evidence for the sake of argument, or incomplete evidence for the sake of discussion. Therefore, the issue of post hoc

11 See "A Neural Substrate of Prediction and Reward," Schultz, W. (1997).
12 E.g. "Words and the World: Predictive Coding and the Language-Perception-Cognition Interface," Lupyan, G. and Clark, A. (2015).
13 E.g. "Unifying treatments for depression: an application of the Free Energy Principle," Chekroud, A. (2015).
14 E.g. "The Role of Prediction in Social Neuroscience," Brown, E. and Brune, M. (2012).

selection of priors and likelihoods is moot—the point of modeling is to sculpt the best possible facsimile of the real thing. My reply is that though this might be the correct way to think about models, it only solidifies their status as non-scientific. If a model is maintained by being continually adjusted to fit the evidence, then it is not falsifiable.

Second, one might argue that the UT is not actually an attempt to explain all brain function, only brain function at a particular level of abstraction. Therefore is it unfair to say that the UT is lofty or precariously grand in scope. My response is that the person making it should read Friston's work; he intends to connect the principles of thermodynamics and statistical mechanics to behavioral psychology by utilizing information theory and neurophysiology at the micro-, meso-, and macro- scale. There is more than one level of abstraction involved here —all levels are purportedly being explained at once.

Third, in response to the intuition pump[15] that explained learned phobia in terms of hierarchical Bayesian PC, one might argue that internally-generated models are not actually permitted under the UT framework. As such, the claim that the framework allows one to explain sub-optimal psychology in terms of optimality is wrong. My response is that if such an explanation is prohibited by the framework, it has not been clearly stated, which is a problem with the UT.

Fourth, one might argue that the UT is indeed falsifiable through testing. Furthermore, it has been tested and continues to be tested. My response is that though testing of what may be deemed part of the UT framework has occurred, it has not been guided by specific and risky predictions that could falsify the whole framework. Rather, only particular implementations of the framework have been tested, and not with enough neural specificity to falsify the vague predictions for the underlying neural mechanisms of the framework.

Fifth, one might argue that vague unifying models are important to science—that this is the impetus for great discovery and profound understanding. My reply is that though exciting theories like UT are indeed an impetus for scientific action, what is far more important is scientific theory.

Conclusion

If sub-optimal psychology like learned phobia can easily be explained in terms of optimal Bayesian neuro-computation, can the proposed UT of brain function explain absolutely *any* state of a human mind? If so, is this a weakness of the framework rather than a strength? Is UT a non-scientific theory? This paper argues *yes* to each of these questions. Therefore, the UT should demonstrate exactly how it can be falsified. This will allow research endeavors to strive for direct neural evidence that either supports the theory's predicted implementation, or

15 Daniel Dennett's weird but useful term.

demonstrates its untenability. Furthermore, the complexity of the brain may very well perpetually thwart attempts at unified scientific explanations, thus we should be prepared for —perhaps even expectant of—the need for a less tidy collection of explanations, i.e. explanatory pluralism.

To close, this paper was motivated by fascination and admiration for the work being done by Clark, Friston, and the other philosophers and neuroscientists who have eagerly constructed a grandly unifying theory of brain function. It was also motivated by the suspicion that the current unified theory might be too good to be true. If this attempt at a UT has an Achilles heel, we need to locate it so that a stronger UT can replace it. We cannot yet locate a point of weakness in the UT because the UT is not yet in "the scientific game"—or to extend the Achilles metaphor, the UT has not yet fully entered the battle.

Sources

Alink, A., C. M. Schwiedrzik, A. Kohler, W. Singer, and L. Muckli. (2010) "Stimulus Predictability Reduces Responses in Primary Visual Cortex." *Journal of Neuroscience* 30.8: 2960-966.

Anderson, Michael L., and Tony Chemero. (2013) "The Problem with Brain GUTs: Conflation of Different Senses of "prediction" Threatens Metaphysical Disaster." *Behavioral and Brain Sciences Behav Brain Sci* 36.03: 204-05.

Bear, Mark F., Barry W. Corners, and Michael A. Paradiso. (2016) *Neuroscience: Exploring the Brain*. Philadelphia: Wolters Kluwer.

Berkes, P., G. Orban, M. Lengyel, and J. Fiser. (2011) "Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment." *Science* 331.6013: 83-87

Bowers, J. S., & Davis, C. J. (2012) "Bayesian just-so stories in psychology and neuroscience." Psychological Bulletin, 138(3), 389–414.

Brown, Elliot C., and Martin Brüne. (2012) "The Role of Prediction in Social Neuroscience." *Frontiers in Human Neuroscience Front. Hum. Neurosci.*6

Chekroud, Adam M. (2015) "Unifying Treatments for Depression: An Application of the Free Energy Principle." *Frontiers in Psychology Front. Psychol.* 6

Clark, Andy. (2013) "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." Behavioral and Brain Sciences Behav Brain Sci 36.03: 181-204.

Clark, Andy. (2014) *Mindware: An Introduction to the Philosophy of Cognitive Science*. New York: Oxford UP, 2014.

Dayan, Peter, Geoffrey E. Hinton, Radford M. Neal, and Richard S. Zemel. (1995) "The Helmholtz Machine." *Neural Computation* 7.5: 889-904.

Egner, Tobias, and Christopher Summerfield. (2013) "Grounding Predictive Coding Models in Empirical Neuroscience Research." *Behavioral and Brain Sciences Behav Brain Sci* 36.03: 210-11.

Fowler, Christie, Lee, Brian, Kenny, Paul. (2014) "Using Optogenetics and Designer Receptors Exclusively Activated by Designer Drugs (DREADDs)" *Report on Progress*. www.dana.org

Friston, Karl. (2003) "Learning and Inference in the Brain." *Neural Networks* 16.9: 1325-352.

Friston, Karl. (2010) "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience* 11 (2), 127–138.

Friston, Karl. (2012) "Prediction, Perception and Agency."*International Journal of Psychophysiology*83.2: 248-52.

Gerrans, Philip. (2013) "Unraveling the Mind."*Behavioral and Brain Sciences Behav Brain Sci* 36.03: 214-15.

Hohwy, Jakob. (2013) *The Predictive Mind*. Oxford: University.

Kahneman, D., Slovic, P. and Tversky, A. (1982) *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Lupyan, G., and A. Clark. (2015) "Words and the World: Predictive Coding and the Language-Perception-Cognition Interface." *Current Directions in Psychological Science* 24.4: 279-84.

Marcus, G. F., and E. Davis. (2013) "How Robust Are Probabilistic Models of Higher-Level Cognition?" Psychological Science 24.12: 2351-360.

Popper, K. (1934). *The Logic of Scientific Discovery*.

Rao, R., & Ballard, D. (1999) "Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects." Nature Neuroscience, 2, 79–87.

Schultz, W. (1997) "A Neural Substrate of Prediction and Reward." *Science*275.5306: 1593-

599.

Spratling, Michael W. (2013) "Distinguishing Theory from Implementation in Predictive Coding Accounts of Brain Function." *Behavioral and Brain Sciences Behav Brain Sci* 36.03: 231-32.

Summerfield, C., and Egner, T. (2009) “Expectation (and attention) in visual cognition.” Trends Cogn. Sci. (Regul. Ed.) 13, 403–409.

Weiss, Yair, Eero P. Simoncelli, and Edward H. Adelson. (2002) "Motion Illusions as Optimal Percepts." Nature Neuroscience Nat Neurosci 5.6: 598-604.

Xu, F., & Tenenbaum, J. B. (2007) “Word learning as Bayesian inference.” Psychological Review, 114(2).