Aporia vol. 23 no. 1—2013

Complementing Churchland's Novel Strategy

Nicholas Havrilla

Paul Churchland recently offered a novel strategy for closing the explanatory gap for consciousness. Churchland's strategy is to use novel predictions from color science to confirm an explanation of an aspect of human experience, color qualia. He does so by deriving predictions about novel colors from a hypothetical identity between neuronal coding vectors and subjective color experiences. The proposed strategy is novel because Churchland proclaims not to be using just explanatory criteria from relevant cognitive neuroscience but empirically predictive resources to demonstrate that the explanatory gap can indeed be closed. However, Churchland's use of novel predictions is ambiguous. His concept of novel prediction is intended to establish an empirically confirmed explanation to close the gap but his argument turns out to rest on a concept of explanation based on extra-empirical criteria.

In this paper, I will identify and articulate Churchland's ambiguity and then offer a genuine empirical solution to complement Churchland's intended strategy. Using Ian Hacking's famous distinction between representation and intervention, I will show that Churchland's argument attempts to provide a 'representational' explanation where an 'interventionist' explanation would be more appropriate for answering the explanatory gap. I will then import James Woodward's recent account of intervention-based explanatory strength to the discussion. Through Woodward's framework, I

Nich Havrilla is double majoring in mathematics and philosophy at Portland State University. His interests include learning theory, the foundations of statistics, and cognitive science. He will be entering Carnegie Mellon University's Logic, Computation and Methodology program this fall.

will reinterpret Churchland's novel predictions as interventions and show how they satisfy empirical criteria for explanatory strength. Before I get to my analysis of Churchland I will first give a brief clarification of some central concepts utilized in his argument: novel predictions and hypothetical identities.

I. Novel Predictions

A central concept of Churchland's argument is novel prediction. While novelty has historically been an empirical criterion of a theory, its meaning and use has recently shifted to unempirical and unificationist criteria. In particular, David Harker has shown how the value for novelty in theory confirmation reduces into nothing but extraempirical or unificatory criteria, strength, and simplicity.

'Novelty' has various meanings, even if all traditionally indicate empirical matters. Perhaps a common core is that some empirical results have special status for those theories that entail them. Furthermore, newness or unexpectedness is always associated with the concept. In the current landscape, three concepts of novelty are readily identifiable. One is temporal novelty, which gives probative or epistemic significance to confirmed predictions about previously unobserved phenomena (Sober & Hitchcock 4). Most find the temporal notion unsatisfactory for historical reasons. While there are cases in history that seem to corroborate this account, such as Fresnel's derivation of the bright spot, there are many cases where the result was known for some time yet the confirmation of the prediction was just as unexpected and 'novel', such as Einstein's derivation of the precession of Mercury's perihelion. Other concepts have been identified that highlight more clearly the basic features of the privileged role of some confirmed predictions. According to the heuristic conceptualization of novelty, an empirical result is novel if it was not used or involved in the construction of a theory. The epistemic point is to undermine ad hoc methodologies or habits of fitting theories to data (Harker 433). Another alternative concept of novelty not based in a temporal index is theoretical novelty (Sober & Hitchcock 4). Here, a result is novel if a theory entails it and if it is unexplainable (i.e. not derivable) or improbable relative to other theories. The associated epistemology underscores a theory's ability to account for anomalies of rival theories.

David Harker has recently analyzed various positions on the value of novel predictions. Harker argues that the value of novel predictions to a theory dissolves into more basic unificatory criteria. Specific positions on true novel predictions rest on two intuitions about confirmation: an increase in explanatory strength without loss of simplicity and a reason to prefer the successful theory to its rivals (448). Consider, for example, the heuristic account of novelty, which is endorsed on the basis of the disvalue of ad hoc hypotheses, theories built to fit particular phenomena (446). This disvalue boils down to a preference for simpler theories. The issue is not that the evidence wasn't used but that no additional assumptions were required to entail or explain the novel evidence (447). Thus, the value *really* comes from an increase in explanatory strength of a theory without loss of theoretical simplicity. Harker argues that all intuitions about novelty ought to be interpreted confirmationally or epistemologically in this sense. In a similar vein, theoretical novelty boils down to our preference for theories that indicate progress and explains what rival theories cannot (450). This objection to the epistemic value of novelty is particularly relevant for my own analysis of Churchland's strategy in the fourth section of this paper.

II. Hypothetical Identities

The other concept central to Churchland's argument is 'hypothetical identity'. Recently, William Bechtel and Robert McCauley have elaborated extensively on its epistemic value and relationship to the philosophy of mind, particularly psycho-neural identities.[1] Hypothetical identity, as an epistemic tool of cognitive science, is a relatively recent idea in the philosophy of science. As opposed to traditional conceptualizations of mind-brain identities found in philosophy of mind (e.g. U.T. Place or J.C.C. Smart's formulations), hypothetical identities are more so a method for discovery than a reductive theory. Instead of identity being strictly the end of research, it can also be the beginning through hypothesizing an identity (Bechtel 236). It is a tool used between disciplines as a heuristic, particularly between neuroscience and psychology. Its purpose is to generate research that leads to the development of more refined and accurate hypotheses (McCauley & Bechtel 737).

Criteria for evaluating the quality of hypothetical identities are loosely available. According to McCauley and Bechtel, hypothetical identities are falsifiable and their justification is the same as for any other scientific hypothesis (751). They are vindicated by the predictive and explanatory progress they initiate (754). However, the more hypotheses the identity informs and the more successful those hypotheses prove, the more likely the hypothetical identity will become a scientifically reliable reduction (McCauley 8). In sum,

[1] See McCauley & Bechtel (2001); Bechtel; and McCauley.

the assumed identities ought to be as rigorously assessed as any other serious scientific theory, which includes their heuristic value.

A motivation for comparing hypothetical identities in science to classical mind-brain identity theory is noticing that no longer is "philosophical cleverness" or "metaphysical comfort" the basis from which the identity theory gets support and plausibility (McCauley 5). For McCauley and Bechtel, psycho-neural identities should be hypotheses supported by empirical evidence and metaphysical considerations of identity claims should be marginalized. Empirical and explanatory adequacy should support hypothetical identities. This final nuance of hypothetical identities is significant for Churchland's argument, as he uses a hypothetical identity to derive novel predictions, a methodological criterion of science, to resolve the explanatory gap, traditionally a problem in the philosophy of mind.

III. Churchland's Novel Strategy

Churchland uses novel predictions, derived from the hypothetical identity of color experience and neuronal coding vectors, to settle the explanatory gap, what amounts to an ontological debate. He attempts to show that subjective qualia experience is reducible to physical processes. The novel predictions give special confirmation to the hypothesized identity relation and demonstrate the predictive and explanatory power that a physical theory can have of subjective qualia (528). In this section, I will present his argument and underscore its significance: importing the value of novel prediction, which is traditionally a criterion deciding between scientific theories, into the philosophy of mind debate. This will lead into the next section, where I will identify an ambiguity in Churchland's execution of his strategy.

The physical theory holding the identity with subjective experience is the color-opponency theory. The specification of color-opponency is the Hurvich-Jameson net (H-J net), which is an attempt to explain the structure of the color experience. It does so in terms of coding vectors: at the input of the visual system are three types of cone cells responsible for three regions of the visible spectrum and at the output are three kinds of color coding cells that code for any visual stimulus (along the axes blue/yellow, green/red and black/white) (529). The H-J net accounts for Munsell's classic spindle theory of colors as a sub-region of the now cubical-shaped activation space of the H-J net (538). Through the assumed identity relation, the color qualia space is the opponency theory activation space, which, in virtue of its cubical shape, implies that the opponency theory can account for representations of color not possible by the spindle theory.

From the H-J net and its account of possible color experiences are derived novel predictions of colors that lie outside the spindle. The predicted facts are what he calls "chimerical colors". Chimerical colors are paradoxical by definition to be predicated of a real physical object, but, nonetheless, appear to be represented in experience (545). They are a class of after-images outside of Munsell's spindle produced by fatiguing certain opponent-cells. What is generated in the opponency activation space are colors with distinct hues that appear to be darker or as dark as black. Colors with hue that are as dark as black are paradoxical by definition because it would seem that there is nothing as dark as black. Yet, when certain opponent cells are fatigued in a certain way, humans will experience such paradoxical colors represented as after-images (546). The color-opponency theory explains and predicts such chimerical colors along with other color experiences that lie outside of the classic spindle (such as "self-luminous" (547) and "hyperbolic" colors (553)).

The predictions are derived from a series of generalizations that can be expressed as the following functions:

$$A_{g/r} = 50 + (L\text{-}M)/2$$

$$A_{b/y} = 50 + [(L+M)/4] - (S/2)$$

$$A_{w/b} = 50 + [(L+M+S)/6] - (B/2)$$

The three generalizations code four types of input (S, M, L, B) into three kinds of outputs ($A_{g/r}$ = green/red axis, $A_{b/y}$ = blue/yellow axis, $A_{w/b}$ = white/black axis) (530). The three outputs constitute a three dimensional vector (e.g., <50, 50, 50>) that determine color experience. The significance of the functions, if they are accurate, is that they present the range of possible activation points in experience (533). Furthermore, corresponding fatigue/potentiation vectors emerge for any "extremal activation triplet" or, to put it simply, when the normal activation vector (i.e., <50, 50, 50>) is shifted by a fatiguing or potentiating stimulus (e.g., <50, 0, 50>) (541). The functions above guarantee that any color experience will lie in the classic spindle. However, the additions of the f/p shifts create vectors that land outside the spindle (544). By shifting the activation vector, "impossible" or "chimerical" color experiences can now be generated.

According to Churchland, these novel predictions of colors outside of the spindle lend special confirmation to the H-J net as an explanation of color experience. With the identity relation initially assumed, they test the plausibility of a reduction of subjective color experience to the opponency theory (556).

The significance of his argument is in importing the value of novel prediction, which is traditionally a criterion deciding between scientific theories, into the philosophy of mind debate. This strategy is a realization of the sentiments expressed by Bechtel and McCauley when they assert that philosophical theories shouldn't be safe from our best scientific theories. The classic mind-brain identity theory is argued for on the basis of scientific explanation that satisfies methodological criteria for theory appraisal. In fact, Bechtel and McCauley assert that mind-brain identity theory construed more appropriately with hypothetical identities, what they call Heuristic Identity Theory or HIT, resolves classic philosophical challenges, such as the "correlation objection" (754). Hypothetical identities, like any scientific theory, are inferred in virtue of their empirical and explanatory success. Churchland uses this feature and highlights novelty as a criterion for explanatory success. The results of the H-J net effectively fill the explanatory gap. Churchland says: "These predictions provide no less than an empirical test of the identity theory itself, in one of its many possible (physically specific) guises" (528). However, his execution of this novel strategy is suspect because it is ambiguous as to whether the H-J net is a potential or actual explanation. As I will show, this ambiguity hinges on the unempirical nature of the novel predictions he utilizes.

IV. Churchland's Ambiguity

As briefly indicated in the first section of this paper, recent philosophy of science has deemed novel predictions a thoroughly unempirical criterion. Churchland's use is consistent with this trend. This raises an ambiguity as to whether the explanation the predictions facilitate actually fills the explanatory gap or is simply a potential explanation that satisfies unempirical criteria for theoretical adequacy.

There are at least two interpretations of Churchland's argument. Maximally, the explanatory gap could be filled by the explanation facilitated by the novel predictions. Minimally, the explanation could be only a conjecture among other competing explanations. The difference between the two possibilities is that in the former Churchland would have aspired to solve the explanatory gap by providing an explanation, while in the latter he would have answered only the weaker problem of whether it is even possible to conjecture an explanation to fill the gap, regardless of its plausibility. Resolving this ambiguity has consequences for Churchland's argument. As I will show, this distinction can usefully be mapped onto Ian Hacking's famous distinction between representations and interventions. But first, I need to clarify Churchland's sense of novel prediction.

Of the standard accounts available in the literature, Churchland's use of novel prediction fits most comfortably with the heuristic account. It is the H-J net's ability to account for unexpected elements of experience outside of the spindle that grants it novelty and leads Churchland to the identity inference. It seems that the "theoretical" account may fit his description of the chimerical color's novelty. In describing novel phenomena he associates such phrases as "paradoxical", "impossible" or the more telling, "by prior semantic lights, flatly self-contradictory" (528, 545). Because the theoretical account relates novelty to such terms, Churchland's use seems compatible. His argument claims that chimerical colors are improbable and unexplainable relative to ordinary semantics. However, in every case in the history of science where theoretical novelty is demonstrated, it is against some specified theory or a collection of theories in background knowledge. Churchland would appear to claim the results of the H-J net novel with respect to ordinary language and experience. The obvious problem is that ordinary ideas on colors may be irrelevant. If such ideas were relevant, it could be that many scientific theories are novel simply because they cannot be reconciled so easily in ordinary language. For the H-J net's result to be novel in the theoretical sense, there must be some rival theory or knowledge specified that would give meaning to novelty, to ascribe a low probability to the results. A look at the classic cases of novelty (e.g., Einstein's GTR, Fresnel's bright spot) entailed results that were genuinely unexpected on the basis of scientific knowledge. How this could work in Churchland's argument is having the opposing position, dualism, associate a low probability to the experience of the chimerical colors.

The heuristic account is more plausible. Churchland claims it was not the motive of the original H-J net's proposal that chimerical colors be explained. They were unanticipated as results because the explanatory target was the ordinary experience of colors represented by the spindle in terms of the visual system. Churchland describes the results as "excess empirical content" that would support a reduction in addition to the H-J net's accommodative success of the spindle (554). Here Churchland is expressing the unexpectedness of the results in terms of motive and intention, which is a version of the heuristic account of novelty (Harker 434).

Novel predictions, and the heuristic account in particular, have recently been shown to be an unempirical criterion of theory appraisal. As mentioned in the first section of this paper, Harker has argued convincingly that the heuristic account actually dissolves into more fundamental theoretical virtues. Churchland's argument uses novelty to give support to the H-J net and identity theory. Out of standard accounts only the heuristic account works. But if Harker's analysis is correct the heuristic account is a specifically unempirical route for confirmation, satisfying extra-empirical criteria for explanation. I believe importing Hacking's famous distinction between representation and intervention into the discussion illuminates why the ambiguity identified at the beginning of this section and the unempirical status of the heuristic account creates an ambiguity for his argument; and furthermore, since it is his intention to fill the gap, it represents a flaw.

According to Hacking, representations are theories, and, as such, hypothetical in nature (273). Here, novel predictions are a deciding factor between competing scientific representations. Hacking's interventions, in turn, concern manipulating and involving experimental setups (272). Their causal nature allows interventions to persist, even through higher-level theory change (Chalmers 161). From Hacking's perspective on epistemology, representations compete to explain the same phenomena but are subject to change; interventions manipulate phenomena, create other phenomena, and, once demonstrated, become permanent causal facts (274). Once established, all theories will have to somehow accommodate them.

Returning to Churchland's ambiguity, if his aim to present a theory, or representation, as a possible explanation then he has achieved his goal in virtue of the use of novel prediction, in the heuristic sense. However, he can use either an empirical criterion for explanation or an extra-empirical one. The difference is that the former makes use of the persistent empirical store of Hackingesque interventions while the latter is based on conjectural criteria. If Churchland's intention is to supply an explanation to fill the explanatory gap, the argument is lacking because Churchland hasn't exploited the empirical dividends of the chimerical colors. Essentially all Churchland has achieved is to show that the H-J net satisfies extra-empirical criteria based on strength and simplicity. But, as Hacking points out, the representational status of the explanations implies that they are subject to change and, worse, may be wrong. To put it simply, they are hypothetical. What would achieve the maximal goal of filling the gap is a sense of novel prediction rooted in empirical criteria, such as Hacking's interventions, exploiting the novel results' permanent place as interventions. The difference between the purely representational status of the H-J net and to what extent it is embodied in interventions would show the chimerical results of the color experiment to be novel and explanatory in virtue of their empirical nature, rather than a lofty representational criterion.

I believe an empirical criterion of explanatory strength can be constructed on Hackingesque grounds or at least by supplementing it with James Woodward's work on interventionist explanation. Furthermore, in this context, explanatory strength will depend on an "interventionist" prediction concept that is distinguished from the traditional theoretical concept through its independence from theory, a genuine intervention.

Such a concept of explanation, if satisfied by Churchland's argument, would make good on Churchand's novel strategy to answer the explanatory gap.

V. Interventionist Explanatory Criteria

Churchland's use of novelty is thoroughly unempirical. An empirical use must employ Hacking's idea of scientific intervention. An experiment that generates the experience of chimerical colors could be similar to past novel interventions, such as Fresnel's bright spot. Here, knowledge articulated in some generalization is used to generate new and unanticipated phenomena. If Churchland's argument were to show the H-J net to be necessary for the derivation of chimerical colors, then the theory would be more than just representational. In order for the net to be more than mere representation, it would minimally have to satisfy a criterion of explanation rooted in interventionist concepts. The key to breaking out of these unempirical tendencies is connecting the novel predictions of chimerical colors, as interventions, to the explanatory strength of the theory that entails them. In James Woodward's theory of causal explanation, novelty and its connection to explanatory strength can be articulated in a manner that makes novel predictions distinct from extraempirical methods of science.

In *Making Things Happen*, James Woodward gives precise meaning to intervention through causation to develop an empirical concept of explanation. His concept is "invariance" or "invariant relationships" (325). Woodward's invariance is a concept of scientific generality and an empiricist notion of explanation. It intentionally avoids all unificationist associations such as scope, simplicity, and strength. According to Woodward, explanations are such only if they appeal to generalizations that are invariant. An invariant generalization is explanatory because it holds under interventions on the value of its variables as predicted by the generalization. Such predictions are "testing interventions" because they test and establish the invariance of the generalization. The invariant generalizations that result from such interventions can be used to answer a range of questions about the conditions under which their explananda would have been different—"what-if-things-had-been-different questions" or "w-questions" (191).

An intervention is an experimental manipulation and allows for understanding causation when it has the right structure (28). Woodward expresses the core idea through a "switch." An intervention will cause *x*, and act as a switch for all other variables that cause *x*. Thus, *x* ceases to depend on anything but the intervention; any causal path from the intervention to the dependent variable *y* must go through *x* (98). In other words, for something to be a cause, it must indicate what it is to manipulate a variable. Causes are variable changes: changes in the value of one variable produce changes in the value of another (45). Moreover, for Woodward, causation is a counterfactual notion. Thus, the meaning of a causal claim is exhausted by the different (including hypothetical) experiments associated with it.

For epistemological purposes, Woodward's ideas on the quantity and quality of interventions are central. According to Woodward, it is the number of interventions and *important* interventions over which a generalization holds that determines the degree of invariance. In Woodward's theory of scientific explanation, explanatory strength reduces to the degree of invariance (257). The model of explanatory strength is not direct or explicit. For example, the degree can be determined by comparisons between generalizations. If one is a subset of another the latter can then answer a larger set of w-questions, and hence be more explanatory (260). But this is not all. *Judgments* on the importance of some interventions also determine degree of invariance (264). This allows incorporating a more traditional model of method in Woodward's theory. Judgments concern the kinds of interventions over which a generalization is invariant and that in turn can include *new* interventions. For Woodward, important interventions are always determined as such relative to the generalization's domain or discipline expectations (262).

VI. Reinterpreting the Novelty of Chimerical Colors

If we accept Woodward's criteria for explanation, an increase in range of invariance is an increase in its explanatory strength. Confirmation is not an intended target of Woodward's account of explanation. Yet, novelty can be smoothly related to it through his ideas on testing interventions. Novel predictions are the epistemic tools by which new and important *kinds* of interventions are referenced and incorporated into the domain of a scientific generalization. These confirm a generalization by facilitating its explanatory depth. Testing interventions establish invariance and important testing interventions establish a high degree of invariance. Moreover, like classic novel predictions, novel interventions can be seen as hypothetico-deductive in nature: predictions derived from the generalization that test and establish the generalization as explanatory. This brings us back to Churchland.

Chimerical colors satisfy this interventionist definition of novelty. They are new and unexpected results generated by an intervention and predicted by the H-J net's functional generalizations. The H-J net as presented here also appears to satisfy Woodward's definition of an invariant generalization—for instance, it is change-relating (it is, after all, three functions with four independent variables and three dependent variables). Initially, the H-J net was well received for subsuming Munsell's classic spindle for color experience. The H-J net subsumes that spindle and, through further testing interventions, explains a class of color experiences that go beyond the spindle—the very definition of an important testing intervention or novel intervention. Woodward's sense of explanatory strength can be applied to the case to articulate an empirical criterion for explanation quite independent from extraempirical criteria. As an empirical criterion, interventions measure the explanatory strength of a theory. In Churchland's case, the confirmation of chimerical colors establishes a higher degree of invariance for the H-J net's functional generalizations, which by Woodward's criterion fill the explanatory gap.

VII. Conclusion

Under close analysis, Churchland's novel strategy for closing the explanatory gap for colors rests on unempirical unificationist ideas of explanation. I have shown how an interventionist model of explanation fits Churchland's concerns and strategy better than his classic unificationist one. These advantages generalize beyond Churchland's concerns.

Avoiding unificationist criteria and other units of analysis closely associated with representations is helpful for the philosophy of all special sciences, which are known for lacking unificationist criteria. In fact, many philosophers deny that laws are applicable to the special sciences, including cognitive science. Yet plenty of science occurs in the special sciences, particularly investigations into causal relations. In fact, as early as the 1970s, cognitive scientists such as Newell expressed a worry over the abundance of empirical data and lack of unifying theories in cognitive science (1973). In response to this, many, including Newell, Ketelaar, and Ellis have made attempts to supply unifying theories to areas in cognitive science, attempting to justify their status as explanatory and scientific.[2] To do this they use novel prediction-based methods. Yet in order to solve the perceived demarcation problem these dissolve into unificationist criteria. If I am correct, novel predictions can be had through empirical methods without unificationism. The color case bears witness to this. Novelty is a crucial epistemic tool for the experimental arm of science.

[2] See, for example Newell (1990), and Ketelaar & Ellis.

Works Cited

Bechtel, William. "Decomposing the Mind-Brain: a Long-Term Pursuit." *Brain and Mind*. 3.2 (2002): 229–242.

Chalmers, Alan. "Experiment and the Growth of Experimental Knowledge." *In the scope of logic, methodology, and philosophy of science*. Ed. Gardenfors, Peter, Wolenski, Jan, Kijiana-Placek, Katarzynan. Netherlands: Kluwer Academic Publishers, 2002. 157–170.

Churchland, Paul. "Chimerical Colors: Some Phenomenological Predictions from Cognitive Neuroscience." *Philosophical Psychology*. 18.5 (2005): 527–560.

Hacking, Ian. *Representing and Intervening*. Cambridge: Cambridge UP, 1983.

Harker, David. "On the Predilections for Predictions." *British Journal for the Philosophy of Science* 59 (2008): 429–453.

Ketelaar, Timothy, and Bruce J. Ellis. "Are Evolutionary Explanations Unfalsifiable? Evolutionary Psychology and the Lakatosian Philosophy of Science." *Psychological Inquiry*. 11.1 (2000): 1–21.

McCauley, Robert. "About Face: Philosophical Naturalism, The Heuristic Identity Theory, and Recent Findings about Prosopagnosia." *New Perspectives on Type Identity*. Ed. S. Gozzano & C. Hill. Cambridge: Cambridge UP, 2012. 186–206.

McCauley, Robert, & William Bechtel. "Explanatory Pluralism and Heuristic Identity Theory." Theory & Psychology. 11.6 (2001): 736–760.

Newell, Allen, "You can't play 20 questions with nature and win: Projective comments on the papers in this symposium." *Visual Information Processing* Ed. W.G. Chase. New York: Academic Press, 1973. 283–308.

——. *Unified Theories of Cognition*. Cambridge, Mass: Harvard UP, 1990.

Sober, Elliot & Hitchcock, Christopher. "Prediction Versus Accommodation and the Risk of Overfitting." *British Journal for the Philosophy of Science* 55.1 (2004): 1–34.

Woodward, James. *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford UP, 2003.