Synopsis

Causality, Mechanisms, and Psychology
Saturday, 24 February 2007, Center for Philosophy of Science, University of Pittsburgh

Morning Session
Chair: Edouard Machery

9:00-9:30 Continental Breakfast

9:30-10:00 Position Statement: Phillip Wolff, Psychology, Emory University

10:00-10:15 Response: Carl Craver, Philosophy, Neuroscience, & Psychology Program
Washington University in St. Louis

10:15-10:45 Coffee

10:45-12:45 Discussion

12:45 - 2:00 Break for lunch

Afternoon Session
Chair: Kenneth F. Schaffner

2:00-2:30 Position Statement: J. D. Trout, Philosophy, Loyola University Chicago

2:30-2:45 Response: Carl Craver, Philosophy, Neuroscience, & Psychology Program
Washington University in St. Louis

2:45-3:15 Coffee

3:15-5:15 Discussion

Organizing Committee

Peter Machamer (chair) and John Norton, Department of History and Philosophy of Science and Center for Philosophy of Science, University of Pittsburgh

Sponsors

Center for Philosophy of Science, University of Pittsburgh
Department of History and Philosophy of Science, University of Pittsburgh

Report on Workshop I: Causality, Mechanisms, and Psychology

Holly Andersen
Graduate Student, Dept. of History and Philosophy of Science, Univ. of Pittsburgh

Phillip Wolff opened the conference by motivating the need for and then outlining his framework for the representation of causation. It follows common linguistic usage of causal verbs, which are sorted into three main categories of similarity: cause, enable, and prevent. These three kinds of causal verbs differ according to features of the affector and patient. These differences are whether or not the result takes place (it does for cause and enable, it does not for prevent); whether the patient has a tendency towards that cause (no for cause, yes for enable and prevent), and whether there is a concordance between the tendencies of the affector and patient (yes in the case of enable, no for cause and prevent).

The results of Wolff's empirical data were formalized into a model of representation where the above causal factors are treated as vectors which form patterns of forces in space. Adding vectors in the appropriate way yields predictions of how subjects will label specific situations: as instances of causing, enabling, or preventing. The idea is that the vectors in people's own representations of causation to some extent replicate or mimic the spatial patterns of force vectors out in the world, where we perceive not just kinematic but also dynamic relationships. One benefit of this kind of representation is that it provides a more accurate way of accounting for static causal relationships.

Wolff generalized his account to not just physical forces, however, but also to social ones, treating intentions and desires as force vectors located in the same space as physical forces. For cases of people wanting to cross the road to see a friend, representing that intention as a vector pointed in the direction of the friend was plausible. In the discussion period, a number of people raised issues with this. Are intentions and desires generally (not just specifically spatial ones like 'wanting to be over there') located in space or represented by us using spatial information? Is this vector space of dynamic representation really a spatial one, or could it be more consistently thought of as some kind of logical space representing logical relations rather than purely spatial ones? Carl Craver raised this concern in his response presentation: when we say that the poison disoriented the hamster, there doesn't seem to be anything straightforwardly spatial about our understanding of the situation, and no particular reason to think that our representation of the causal structure of the situation to any extent replicates the actual causal structure in the world.

One participant was concerned that Wolff managed to achieve this unity between physical and social causation by assumption, and that linguistic commonality masked further differences between these kinds of causation. In response to a number of questions, Wolff emphasized that his is not a metaphysical theory about what causation really is, but rather codifies pre-existing linguistic usage. This also makes it not quite a normative theory of how we ought to use causal language, although it does provide grounds to label some uses normal or not normal.

Wolff's Dynamic Model of causal representation also includes transitive dynamics, or how to compound causes and predict which causal label (of cause, prevent, or enable) will be used to describe the overall situation when two or more causal sentences are compounded. Carl Craver described an alternative breakdown of causal relationships in his presentation, developed by his colleague Northcott: it included a difference between letting and abetting. Although it

was not fully developed, Wolff suggested that the further differences Craver pointed to could be understood in terms of compounded causes: abetting is enabling, and letting could be preventing a prevention. This kind of solution could prima facie be levied to deal with cases like 'intending to call one's mother more often,' where a seemingly unspatial intention could be parsed in terms of spatially directed vectors like one in the future towards the phone and towards oneself dialing the number. This part bears further development, both for whether or not it captures the full range of intentional examples, and to what extent the model formalizes the way we really think and reason about these intentional cases.

In his response, Craver expressed enthusiasm for Wolff's work, but still managed to raise some substantive issues with it. One of these was the problem of relevance, where we don't always know what the relevant forces are, such as hexing a pile of salt which causes it to dissolve. In the case of the transitive dynamics, the problem takes this form: powder caused the flame to turn blue; the blue flame caused the house to burn. We don't want to have the blue-turning to transitively cause the house to burn. Hume's problem also still survives in Wolff's account: even though he speaks of us perceiving dynamics, we still don't see them, but merely kinematics. What Wolff seems to be getting at is akin to what Hume calls 'habits'.

There was some concern during the discussion period as to how representative his sample of undergraduates from privileged socioeconomic backgrounds are of general linguistic usage. While in certain regards, this is definitely not a representative sample, it is unclear whether or not they would have markedly distinct usages of causal verbs. Wolff said the usage carried over to foreign language groups as well. On the other hand, it could be quite interesting if one were to do additional work with other groups and uncover systematic differences in causal verb usage.

J.D. Trout discussed his work on how the sense of understanding one sometimes gets when encountering an explanation is not a particularly good indicator of whether or not the explanation is a good one. The sense of understanding refers to the 'aha!' moment, or as Peirce put it, feeling the key turn in the lock. We tend to think that we have got a good explanation when it is accompanied by this feeling, but as Trout argues, there are several psychological factors which account for this feeling and which are poor means of picking out legitimately good explanations.

One of these psychological factors involved in the sense of understanding is the overconfidence bias, where people tend to be overly confident in their own judgments, placing ten thousand to one odds on answers where they are only correct 85% or so of the time. This overconfidence leads to our labeling poor explanations as good ones, because we feel the sense of understanding which is really associated with overconfidence.

The other is hindsight bias, where people overestimate their ability to have predicted something which occurred: they 'knew it was going to happen', even though they couldn't have successfully predicted before the event occurred. Evidently Trout has tricked people by invoking their hindsight bias, where in a description of his work, he tells them the opposite of what they found and people claim it to be obvious. It was a mistaken sense of understanding that led them to believe they understood his work well enough to predict the results, and yet get it so wrong.

Trout explained that actuarial models, including statistical prediction rules (SPR's), needn't get at the underlying causal structure of patterns of variable correlations to nevertheless outperform experts in making predictions in cognitive tasks. This is true of parole boards, where SPR's more accurately gauge recidivism rates than the boards do, and in APA hiring practices (ouch). For tasks requiring perceptual discrimination instead of primarily cognitive reasoning,

humans do quite well, even though they can't generally explain what it is they are using for perceptual discrimination. A way to get humans to perform better on judgment tasks is to represent the difficulty of the task in the stimulus itself, such as having to judge the texture of a building from a grainy photo, where it is apparent from the perceptual stimulus itself that this is tough. Other means of calibrating judgments via metacognitive control were discussed, such as making subjects explicitly consider alternate or opposite hypotheses.

Craver's response included a defense of Salmon's ontic explanation view, where the linguistic entities we get a sense of understanding from are not the explanations themselves, and that we should not conflate understanding with misunderstanding – sometimes we think we understand and are just wrong. If explanation is thought of as bringing representations to bear on the world, said Craver, then we are focused on the relationship between representation and world, which is the wrong place to focus. Instead, we should focus on explanatory structures and relations in the world, such as causal structures (and multi-level mechanisms). Craver offered a slightly different account of explanation, based on factors like the number of prototypes under which a phenomenon can be subsumed and the degree of fit of those prototypes. He had a substantial list of different kinds of explanations one could give, each of which could then be associated with a different kind of understanding.

The discussion period raised several issues for Trout's claims, including the fact that humans themselves need to be involved in choosing which variables are kept track of, for which SPR's can then codify statistical calculations. As Wolff pointed out, this is also an impractical way to make most decisions. While school admission and parole decisions could utilize SPR's, we simply lack the models for many other decisions and must rely on our own expertise. There is also the further point that for decisions which are made infrequently or based on a small number of cases, instead of massive numbers, humans seem to have the advantage.

John Norton added the confirmation bias to the psychological features which give rise to mistaken senses of understanding, and Ken Schaffner added another kind of explanation to the mix: the sort found in professional journals like Cell, where a team of authors is somehow trying to persuade or convince others of their position.

It was somewhat unclear what the upshot was of the difference in performance of experts and SPR's, if Trout was suggesting that as a matter of policy we should start replacing APA hiring committees and parole boards with SPR's. The problem of gaming was raised: if the SPR's were not getting at genuinely causal relationships between the variables, and people knew what factors were considered in the model, they could play to those factors. Trout responded that the actuarial models can evolve through time to take this into account, such that a previously important indicator of future behavior is eventually no longer an indicator. It seems like, in the case of recidivism and parole, the lag time between gaming the model and the model compensating could be years, where the model will always be somewhat behind and trying to catch up to the ongoing changes in behavior of the subjects.

Trout also drew some conclusions for standard analytic epistemology, which evidently resists the idea that models can outperform them in making predictions.

Report on Workshop II: Causality, Mechanisms, and Psychology

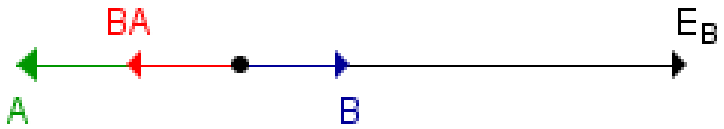
Johannes Persson
Visiting Fellow, Center for Philosophy of Science, University of Pittsburgh
Department of Philosophy, Lund University

Position statement I:
Force dynamics in causal meaning and reasoning
Phillip Wolff

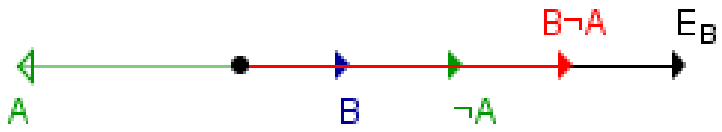
Phillip Wolff began his position statement by a series of noteworthy observations. For instance, (a) we distinguish between cause and enable. "A cold wind *caused* him to close the window" but "A crank *enabled* him to close the window." Moreover, (b) we use causal talk also in static situations: "Pressure will cause the water to remain liquid at slightly below 0°C." These observations challenge traditional theories of causation. (a) is not obvious from a dependency perspective. (b) is not to be expected on a transference view. Wolff's preferred view, the force dynamics theory, predicts these phenomena. It builds on the link between cause and force, and on the categories of affector, patient, and endstate: The wind (affector) caused the boat (patient) to heel (endstate). It is important to note that endstate is not defined by patient or affector.

The force dynamics theory predicts that different causal concepts will be employed depending on the patient's tendency for the endstate (Y/N), concordance between affector and tendency of the patient (Y/N), and approached endstate (Y/N): Cause when N-N-Y; enable Y-Y-Y; prevent Y-N-N; and despite Y-N-Y. The theory received such support from a series of experiments where subjects were exposed to animations with transportation vehicles, heading in a certain direction, which were suddenly exposed to strong winds (simulated by fans). At the workshop, Wolff showed the boat & cone set up. Wolff also presented work on the extension of the force dynamics theory in the direction of social causation. In a similar experimental setup subjects were invited to categorize animations with a woman (patient) and her partner (affector) on different sides of the street of a two way junction, and with a police (affector, acting as a traffic light) in the middle. Again, the predictions of the force dynamics theory were in accordance with how the subjects responded.

Finally, Wolff presented an extension of the force dynamics theory, the transitive dynamics model. According to this, people can construct causal structures by linking together two or more force dynamics patterns. For instance, one can form an opinion on the causal relation between *vegetation and landslides* by linking force dynamics models of *vegetation prevents erosion* and *erosion causes landslides*. According to Wolff, this is done by adding the resultant force vector from the first to the model of the second. To make the model more complete for purposes of causal reasoning, Wolff stipulated how to that negations of the affector and result consist in reversal of the vectors. To exemplify,



Here patient B has a tendency for the endstate E_B, but neither the affector A nor the result BA (i.e, endstate is not approached) is collinear with this tendency. It is a Y-N-N pattern, so *A prevents B*. Now, reverse the affector A (not-A causes B):



This is a Y-Y-Y, so *Not-A enables B*.

Carl Craver
Comments on Wolff

In his response Carl Craver challenged the argument for the force dynamics theory along the three dimensions of metaphysics, epistemology, and psychology.

The metaphysical challenge consisted in the claim that causation does not always involve a spatial endpoint (and forces making a difference with regard to spatial position of the patient). Examples seem to abound: The poison caused the hamster to become disoriented; her embrace caused him ululate.

The epistemological challenge consisted in whether we really perceive these dynamics of forces, and to what extent Wolff's model is a model of causal perception. Craver argued that Wolff rather gave an account of the psychological habit (to speak with Hume) by which we infer hidden causes.

Craver's primary worry about the psychological aspect was that one seems to learn about forces gradually by learning which physical/spatial relationships make a difference. This suggests that counterfactuals are more fundamental than forces. And by introducing the notion of the affector's difference making instead of the patient's tendency for the endstate, it seems that we can more easily account for finer differences between various meanings of "enable", as for instance the help/let distinction: A *lets* B pass the street: Does A (affector) make a difference with regard to the result? Yes. Does A's difference making and B's tendency concord? Yes. Is the result achieved? Yes. (Note the difference in A *helps* B to pass the street: N-Y-Y).

Three themes from the discussion:
1. Can the model be applied to the non-spatial?
Several questions concerned the extended applicability of the model. Heat differences, color changes, and a number of non-spatial intentions (for red jumpers, phone calls, etc) were reported. A special worry on this matter concerned multiple causes of different character. The force dynamics theory seems to presuppose a common metric.

Granted that the model should be applicable to non-spatial cases, in what sense is "physicalistic" force dynamics a fundamental feature of the theory? That Wolff intended an extension by analogy is clear, but the reason why was not evident. A number of Wolff's assumptions about the mental representation of causation, given a "physicalist" theory, might survive a less "physicalistic" setting. Among these assumptions are that the theory copies or reproduces causation in the world; causation is ultimately based on local interactions; local

interactions are deterministic; and that noncontiguous causal links involve chains (implies need for mechanism).

2. Does the model represent anything in the world?
An underlying assumption is that the force dynamics model copies causation in the world. Two reasons why this might not be true are (a) that it doesn't seem crucial to ground force dynamics in all the details of actual dynamics (direction and origin of salient forces are often enough for causal perception), and (b) physics no longer assumes that Newtonian forces are in the world. Basing the model on *naïve physics* obviously weakens the point that we are capturing causation in the world.

3. Is this just another descriptive model of causation?
Another point concerned the normative claims of the theory. Are people who deviate from the predicted perceptions wrong or merely non-conformists? Does the theory imply that there is one correct model of causation?

The discussion indicated that there might be at least a claim of unity on offer here. The force dynamics theory implies counterfactuals in a straightforward way (remove the force and see if the result vector changes). But it doesn't stand and fall with counterfactual models. It has no problems with over determination. Wolff also argued that the theory can account for type-level probabilistic causation.

Position statement II
The causes of genuine understanding
J. D. Trout

J. D. Trout's position statement pointed to the epistemic risks of employing certain kinds of explanatory arguments. Sometimes good explanations are portrayed as either constituted by or as giving rise to a sense of coherence, an "aha"-experience, or "feeling the key in the lock turn", i.e. to a sense of understanding. But bad explanations can induce a similar sense of understanding, and good explanations can fail to induce understanding.

Trout argued that it is risky to base one's judgments on senses of understanding because of two well-known psychological phenomena: the overconfidence bias and the hindsight bias. Explanatory beliefs often reflect both these biases. Moreover, there are no internal "certain signs" that distinguish between genuine and counterfeit understanding.

Our sense of understanding needs calibration against, for instance, the most plausible alternative hypothesis. That there is indeed a problem here that needs to be fixed can be seen in the fact that in many areas simple linear models outperform experts' judgments.

Two implications of Trout's statement were (a) that explanation should not be confused with understanding (genuinely explanatory beliefs are produced by more reliable mechanisms), and (b) that psychological findings can inform science (and philosophy) about reliable and less reliable epistemic strategies. Relying on understanding does not always lead to falsehood but one cannot rest content with it since it is not always a reliable cue.

Craver comment II

In his comment Carl Craver sided with Trout in distinguishing sharply between explanation and understanding. Explanation is not "seeing it work". Craver followed Wesley Salmon in understanding explanations as structures in the world.

Then Craver interpreted Trout's argument as involving the premises that the sense of understanding is just a sense of confidence, and that the psychological sense of understanding is the *product* of inherently unreliable psychological processes (hindsight bias and overconfidence). Both these premises are problematic. Does such a causal story give us necessary and sufficient conditions for understanding? No. It is hardly necessary since there are cases where we have understanding without prediction. And confidence is clearly not sufficient. We often have confidence in things we do not claim to understand.

As an alternative to Trout's account of understanding, Craver offered Churchland's PDP model of explanation (to understand is to activate a prototype).

Three themes from the discussion

1. Calibration and scientific practice

Several questions concerned the accuracy of Trout's description of scientific practice. It seems that the scientific community to some extent handles problems of bias. Scientists learn what it takes for a claim to be likely to hold, and what kinds of things are expected from those who participate. Different kinds of peer review might take care of calibration issues.

A related discussion concerned how scientific argumentation is typically built. A good paper, it was said, typically proceeds by running through a couple of alternatives by more or less conclusive arguments. The conclusion often consists in putting forward one of these as an alternative explanation. Hence no effect of overconfidence? Or is this a short version of testing against alternative hypotheses correcting hindsight and overconfidence bias?

2. How can psychological findings inform epistemology?

No epistemologist would suggest that we use unreliable rather than reliable procedures. It was suggested that psychology might still help by showing that some strategies are more unreliable than others. Imagining and introspection, for instance, are sometimes used in philosophy. Overconfidence and hindsight bias show that these are often unreliable (and maybe even in what contexts they are).

3. Overconfidence vs. underconfidence and confidence

Reporter's additional note: Being confident might be a mental state, but being *over*confident is often understood as a relation between confidence and the world. We are moreover prone to be overconfident in hard cases and underconfident in easy. But if so, is it not also the case that we understand better in easy cases than in hard (in which case overconfidence does not really explain understanding as much as stating that we are sometimes confident when reliable evidence is lacking).

Remarks on Phillip Wolff's Position Statement

John D. Norton
Center for Philosophy of Science and
Department of History and Philosophy of Science
University of Pittsburgh

Do We Pictures Causes as Vectors?

Phillip Wolff's reviewed his intriguing theory that people think of causation as vectors. It attracted some skeptical commentary both in question time and in informal discussion. There seemed to be two lines of thought. First, the connectedness of the world and the sorts of causal tasks people face seem much more complicated than a simple vector model and also different from the spatialization of causes that a vector model suggests. Second the idea that people think of causes as Newtonian force vectors seems too simple and, to be blunt, reminiscent of a naive sort of physics chauvinism.

My own feeling is that neither of these objections is telling. First, it may well be the case that the connectedness of the world differs from and even far outstrips what a simple vector model can capture. That is compatible with people holding a simple vector model. It is a commonplace of psychological research that people's mental pictures of many processes do not conform precisely to the realities. Second, we should not condemn an idea because we have a prejudice that any model that looks too much like physics must be oversimplified and thus wrong. The thesis that people picture causes as vectors can be formulated independently of Newtonian physics and is open to experimental testing. That is what should decide.

While neither of these objections are telling, they do raise some doubt and draw attention to the question of whether Phil's experimental results do indeed establish the presence of the vector structure in our pictures.

What Does it Take to Show We See Causes as Vectors?

What does it take to demonstrate experimentally that people picture causes as vectors? The task is simpler than it may first seem. As an analogy, imagine that we want to establish that some space's geometry is Euclidean. There is no need to check that every one of Euclid's many theorems hold. All that is needed is to establish that Euclid's five postulates hold. Is it the case that we can: (1) always connect two points with a straight line; (2) that we can always extend a straight line interval; and so on? If we have five "yes'es" for the five postulates, the job is done.

The task is of comparable difficulty for a vector space. We merely need to identify the defining properties of such a space and then show that they hold. Of the senses of vector space defined in the literature, the one that comes closest to what Phil described is an inner product vector space. Dropping some technical details, causes would form an inner product space if they have the following properties:

> (1) Any cause C can be expressed as the linear sum of component causes
> $C = a_1 C_1 + a_2 C_2 + ... + a_n C_n$

Two important ideas enter here:
(1a) It is possible to "add" causes to produce a new cause.
(1b) It is possible to multiply causes by a scalar (e.g. a real number).

That multiplication is what happens when we write "$a_1 C_1$": we multiply the cause C_1 by the real a_1. So if we have a cause C_1, it makes sense to talk of $2C_1$, $3C_1$, etc.; and, in some suitable setting, they will have twice and thrice the causal power of C_1.

(2) Angles between vectors: it makes sense to say that vectors are orthogonal, parallel and anything in between. This is realized by a "dot product" on the space. (The dot product of vectors V_1 and V_2 is zero if they are orthogonal.)

Two important consequences follow:
(2a) If the vectors $C_1, C_2, \ldots, C_n$ are pairwise orthogonal, then the vector space is n dimensional, since one must in general specify the n numbers $a_1, a_2, \ldots a_n$ to fix an arbitrary cause C.
(2b) The dot product of a vector with itself gives us its magnitude, the norm of the vector.

Doing Without Spatialization

The experimental exercise is to show that these properties are realized in our mental pictures of causes. The cases Phil showed us strongly stacked the deck in favor of a positive result since causes and effects were spatialized. That is obvious in the case of boats being blown by the wind from fans, since the causal power of the wind is manifested in displacements in space of the boats. That same spatialization is present, however, when a girl's inclination for or against a boyfriend is operationalized by whether she crosses the street to meet him or takes a path orthogonal to him. How will things work out when the causes and effects are not spatialized?

Perhaps the most important property to be established is orthogonality, for that distinguishes a simple one dimensional space from higher dimensioned spaces. In the case of wind causes, that is easy to do. People seem to recognize that a north easterly wind blowing across a boat's northerly course can be decomposed into two orthogonal parts, one aligned with the motion that retards its northerly motion; and one orthogonal to it that has no effect on the north-south motion.

Should we not expect this experimental template to be usable in cases in which there is no obvious spatial operationalization? What of medicines, to use Carl Craver's example (if we replace poison with medication)? It seems quite credible that people might model medicines as forming a vector space of causes. Orthogonality is realized in the sense that a medicine may be compounded of several different component medicines that act independently of oneanother. One cures ailment A_1; the other cures ailment A_2. There may even be an angle between two medicines $M_1$and M_2 in the sense that the first acts solely on ailment A_1 and the second largely on ailment A_2 but with a smaller effect on A_1. The causal powers of individual medicines are quite credibly linear in the dose, conforming to the scaling of (1b). Perhaps also people are

willing to assent that the total amount of medicine taken may have a joint effect, say, on one's kidneys, since that organ may be tasked with eliminating the medicine. So the summation of medicines in (1) would be realized.

Of course I am just guessing that experimental investigations will yield these defining characteristics of a vector space in people's pictures. Perhaps in some cases the vector model will be recoverable. Perhaps in others it will not be. The latter might not be an unhappy outcome. As long as the failure is systematic, it will then be possible to replace one defining property by another and thereby to have shown that we picture these causes as conforming to a slightly different structure.

Remarks on Wolff and Trout

Carl Craver
Philosophy, Neuroscience, & Psychology Program
Washington University in St. Louis

Wolff

Wolff offers an original and powerful model for the psychology of human causal judgment. Wolff's papers address themes in metaphysics, epistemology, and psychology. Psychology is the central theme, though, and it is the most promising aspect of Wolff's project.

Metaphysics

I do not think that Wolff's model works well as a metaphysics of causation. A metaphysics of forces faces its own difficulties, of course, but my worry is more specific. Very few of the causal interactions that we routinely perceive and judge are accurately represented as simple forces. Suppose that a toxin causes a laboratory rodent (e.g., a hamster) to become disoriented. The forces required to apply Wolff's schema (the tendency of the hamster not to be disoriented, the inhibitory force of the drug on that tendency, and so on) misrepresent the causal factors at work in the situation (such as the physiological processes involved in locomotion, proprioception, spatial navigation, and the action of the toxin on receptors in the central nervous system). This is no problem at all for a psychological model, of course, but it is a serious problem if one wants to claim, as Wolff sometimes does, that we directly perceive the forces at work in the world. Indeed, if his model is correct, it seems that we regularly misperceive the causes at work as forces when they in fact involve much more specific kinds of activities (such as binding to receptors, opening channels, changing the behavior of the cochlea). "Force" would seem to be a filler term, or schema term, to be filled in with more specific activities and processes.

Epistemology

I also do not think that Wolff ameliorates Hume's skeptical worries about causation. Although he claims at times that people "see" the underlying dynamics of causal systems, he usually puts "see" in scare-quotes. What he means to say, I think, is that we reliably infer the dynamics of the system on the basis of kinematics. However, he recognizes that this is an inference, and that it is a fallible one. Something like Hume's argument can be reformulated in Wolff's language: All meaningful concepts must be grounded in perception. We do not directly perceive dynamics. We only directly perceive kinematics. Kinematics under-determine dynamics. So it is not possible to ground dynamic concepts in observation, and our ideas of dynamics are not meaningful. I do not see that Wolff's research gives us a reason to find this argument suspect. Wolff's psychological model is anti-Humean, in that it constructs our causal judgments and perceptions out of forces rather than regularities, but that model is an alternative view of the psychological mechanisms underlying the "habit" inferring the causal structure of the world.

Psychology

Wolff's psychological hypothesis stands as a clearly articulated alternative to covariational and counterfactual analyses of causation and to those analyses that emphasize the importance of kinematics in our perception of causation. I find the model most plausible and the experiments

subtle and well-crated. Rather than focus on the many strengths of this model, however, let me suggest just a few constructive criticisms.

The Problem of Relevance. It seems to me that Wolff's model presupposes one central aspect of human causal judgment: the assessment of which causal factors are relevant to a given effect. Kyburg's example of hexed salt is a classic illustration: Touch a pile of table salt with a wand, and then put the salt in water. The salt dissolves. Few people will judge that the wand (or the spell behind it) caused the salt to dissolve. The point is that there must be some way to distinguish the relevant forces acting from the irrelevant forces acting. Consider cases where multiple forces are clearly present: Her embrace caused him to blush. What is the relevant force? The force of the arms on the body? The sudden rise in air pressure as the bodies approach one another? Her desire to display affection? In perceiving and making judgments about the causal structure of the world, we explicitly and implicitly assess the relevance of different causal factors. Wolff's model seems to presuppose that we already know how to distinguish relevant from irrelevant forces. This is precisely the kind of problem that probabilistic and counterfactual models of causation are well equipped to handle.

A special case of this problem pertains to Wolff's treatment of CAUSE as a transitive relation. On his psychological model, if people judge that A causes B, and that B causes C, those people should judge that A causes C. There are examples in the philosophical literature, however, that can be used to show that people sometimes fail to make this inference. Suppose that a powder is thrown into the fire causing the flame to be blue, and the flame causes the house to burn to the ground. We are not tempted to think that the powder caused the house to burn. Yet there is a cause-cause configuration in this case. I see this as related to the problem of relevance discussed above. The blueness of the flame is irrelevant to the house fire. Here is an interesting place where the psychology of causal judgment and the philosophical literature on the transitivity of causation might make contact with one another.

Enable and Let. Wolff's notion of "enable" covers both cases of helping (in which the agent aids the patient in the direction of the end-point though the patient would reach the endpoint in the absence of the agent) and letting (in which the agent is necessary for reaching the endpoint). (Wolff draws this distinction in "Representing Causation"). In the papers distributed before the workshop, enable and let are both included under the heading "enable" as cases in which the tendency of the patient is toward an endpoint, the agent and patient concord, and the endpoint is approached. (A parenthetical question: in what sense is the tendency of the patient toward the endpoint if the action of the agent is a necessary condition for the patient to reach the endpoint? Phrased as a question about the psychological mechanism: by what process or inference does one assess the tendencies of the patients and the objects? For example, is the tendency of a person to die or to live? There is a sense in which both are endpoints. What psychological factors influence the choice among them?)

But let's return to enable and let. One way to capture the distinction between enable and let would be to reformulate Wolff's taxonomy as recommended by Robert Northcott, a philosopher of science at the University of Missouri, St. Louis. The basic idea is to replace the "tendency of the patient" in Wolff's Schema with the heading "difference making?" "Difference-making"

means “Does the agent change the outcome from what it would have been in the agent’s absence?” The result of this adjustment is the below eight-place grid.

Northcott’s Taxonomy

	Make a Difference (E or ~E)	Concord?	End?
Cause	Y	N	Y
Enable_{NC} (Let)	Y	Y	Y
Help	N	Y	Y
Despite	N	N	Y
Too much of a good thing	Y	Y	N
Help to avoid	N	Y	N
Prevent	Y	N	N
Nail in Coffin	N	N	N

What is interesting about Northcott’s revision is not merely that it fills in the eight possible force conditions that are recognized but not labeled in Wolff’s work. The interesting point is that if we conjoin Wolff’s model with an appeal to difference-making we get a model that seems to be more predictively adequate than Wolff’s force model alone (because it accommodates the help/let distinction). In his comments, Wolff suggested that prevent might be treated as a two-factor model (described in the comments here). The question then becomes whether there is an experiment for distinguishing Wolff’s model and Northcott’s difference-making model. (I think that some work could be done to embellish Northcott’s model a bit. For example, it would be good to explicitly recognize the distinction between hinder and prevent, which is symmetrical to the help/let distinction).

Acquisition of force concepts. One worry not mentioned above about the psychological aspect of Wolfe’s project was that children seem to learn about forces gradually by learning which physical/spatial relationships make a difference. Consider Renee Baillergeon’s experiments on children’s acquisition of the concept of “support” (or at least their acquisition of dispositions to look longer at anomalous kinds of support phenomena). Without going into all of the details, the development of the concept of support (assuming that these experiments can be so interpreted) begins with children recognizing that contact is required for support, but not recognizing that the contact must be from below, or that the amount of contact makes a difference to whether or not one box can support another. Children seem to first learn (I assume that “learn” is at least close to the right verb here) the idea of support through contact, then support from below, and then support by amount of contact from below. Baillergeon shows that different kinds of causal relations (such as collision) develop independently and seem to exhibit relatively fixed developmental patterns. For present purposes, what matters is that this pattern of learning about

forces seems to reflect children's learning of the relevance of different properties and relations to the forces at work in any given causal situation. That is, children seem to learn about forces, and they seem to do so by assessing which properties are relevant. If this is correct, then Wolff's force model cannot be psychologically fundamental (at least developmentally fundamental) as the ability to learn to see force dynamics seems to presuppose some more fundamental capacity to assess relevance relations, and this is the primary domain of covariational (Humean) and counterfactual views of the psychology of human causal judgment.

Trout

My comments focused exclusively on Trout's views on explanation. I embrace Trout's central claim that explanation must be kept distinct from the sense of understanding. The sense of understanding is an unreliable guide to the quality of one's explanation. It is, after all, possible to misunderstand things, that is, to have the sense of understanding without knowing the explanation. Furthermore, there might be explanations that nobody can understand. Some phenomena are so complex that they cannot be apprehended, let alone understood, by limited cognitive agents such as us. Despite the fact that these complex systems overwhelm our cognitive abilities, it seems wrong to say that they lack explanations. They have explanations; we just can't understand them. In such circumstances (as one finds in the sciences that study gene regulation and interaction and in the study of information processing mechanisms in the brain), one relies on visual diagrams, computer simulations, and other forms of intellectual scaffolding to represent mechanisms for which we cannot evoke the sense of understanding. For similar reasons, it will not do to gloss explanation as "being able to see how it works". As Mary Hegarty's work amply demonstrates, our ability to visualize mechanisms is sorely limited beyond relatively simple machines with just a few component parts represented in two dimensions. The nature of scientific explanation and the psychological sense of understanding need have very little to do with one another.

Despite this central agreement, Trout and I disagree on many things. For starters, we disagree about the history of the philosophical literature on explanation. Where Trout views Salmon, Hempel, and Kitcher as embracing the idea that "understanding" is required of explanation, I read all of them as rejecting that view. It is precisely because these thinkers recognized the limitations of the sense of understanding that they found it necessary to develop a philosophical analysis of this kind of scientific achievement. Theirs is a normative project offered as a corrective to reliance on an intuitive sense of "fit", "understanding" and "making sense". Salmon (following Coffa) advocated an ontic view of explanation:

> The linguistic entities that are called 'explanations' are statements reporting the actual explanation. Explanations, in this view, are fully objective and, where explanations of nonhuman facts are concerned, they exist whether or not anyone ever discovers or describes them. Explanations are not epistemically relativized, nor (outside of the realm of human psychology) do they have psychological components, nor do they have pragmatic dimensions. (1989, 133)

It is true that Salmon later wrote more about the importance of scientific understanding, and he claimed that understanding could be achieved either through knowledge of mechanisms or

through unification under argument schemata. Nothing in this, however, involves abandoning the ontic conception. Indeed, if one reads Hempel generously, it is possible that even he might have endorsed an ontic view of sorts (as Salmon noted). However, even if one is unwilling to be that charitable (it is a stretch, I confess) it is clear that Hempel did not intend to characterize the sense of understanding or to reduce the sense of understanding to a kind of inferential "fit". As he wrote in his *Introduction to the Philosophy of Science*:

> …man has long and persistently been concerned to achieve some understanding of the enormously diverse, often perplexing, and sometimes threatening occurrences in the world around him… Some of these explanatory ideas are based on anthropomorphic conceptions of the forces of nature, others invoke hidden powers or agents, still others refer to God's inscrutable plans or to fate.
>
> Accounts of this kind undeniably may give the questioner a sense of having attained some understanding; they may resolve his perplexity and in this sense 'answer' his question. But however satisfactory these answers may be psychologically, they are not adequate for the purposes of science, which, after all, is concerned to develop a conception of the world that has a clear, logical bearing on our experience and is capable of objective test. (PNS, 47-8).

Similar points could be made for the unificationist Kitcher, and even more persuasively for the recently emerged causal-mechanical Kitcher.

I agree with Trout that there are serious problems for the covering-law and unificationist models, but I disagree as to where the diagnosis lies. I think that it is wrong to seek the norms of explanation in features of a representation or in the relationship between a representation and the thing that it represents. The norms of explanation fall out of the commitment on the part of the investigator to correctly describe the causal structure of the world (thanks to Jim Greeno for helping me to see that this is the right way to put this). The norms of explanation can be made explicit, then, by asking what the causal structure of the world is and how we can most reliably discover it. Thinking this way thrusts our attention away from a focus on internal cognitive structures and arguments and out onto the world that those structures and arguments represent when they correctly describe the relevant portion of the causal structure of the world. That, I think, is a more perspicuous way saying how "internalist" or "inferentialist" models of explanation miss their target.

Trout argues that the psychological sense of understanding is the *product* of inherently unreliable psychological processes (hindsight bias and overconfidence). I argued that these conditions are not necessary for the psychological sense of understanding, as one can achieve the sense of understanding without hindsight bias and without overconfidence. One has the sense that one understands the events of 9/11 (if only dimly) without the sense that one could have predicted them. And one can be quite confident in certain kinds of facts (such as that Prince changed his name to a symbol) without having any sense of understanding why that fact obtained. Furthermore, one might even be confident that something happened, and confident in one's ability to have predicted it, without having the sense of understanding. One might, for example, know only that one can predict storms by consulting barometer readings but have no

understanding of why this predictive technique works so well. So I do not think that this will suffice as a psychological account of the sense of understanding.

One should note also that Trout's psychological hypothesis is radical in that it denies the common-sense view, stated commonsensically, that the sense of understanding is a feeling or judgment that accompanies success in fitting some phenomenon into one's conceptual framework. One might make this common sense framework more precise. One might, for example, adopt Churchland's PDP model of explanation. There are other possible representational frameworks that one might adopt (perhaps Wolff's could work here?), but Churchland's is familiar and lends itself especially nicely to the task of making this common-sense notion explicit. According to Churchland's model, to understand is to activate a prototype. According to the model I have in mind, the sense of understanding is a judgment or affective response that is roughly proportional to the number of prototypes that can be brought to bare on the phenomenon to be explained, to the fit between prototype to explanandum phenomenon, and to the centrality of the applied prototypes in one's intellectual economy (something like the unifying potential of the prototype). This is an alternative starting point for thinking about how one might build a psychological model of the sense of understanding. We could focus attention on how on fit is measured, on whether the sense is affective, propositional, or whatever, on how prototypes are weighted in this process and how relevant prototypes are distinguished from irrelevant prototypes. This is all speculative, of course, but if one takes Churchland's model as a starting point, one could think very concretely about how such a "sense of understanding" mechanism might be implemented in a connectionist system. This "representational fit" model is not committed to the idea that the sense of understanding results from a mistake hardwired into our cognitive systems (as Trout suggests). It is not hardwired, because one could imagine the features of such a mechanism changing with learning, and it is not a mistake (a misleading cognitive spandrel) but rather a central component of information processing in the mind/brain. Indeed, if one is inclined to tell how-possibly stories, one might speculate that the sense of understanding (understood as an affective response to representational fit) has the function guiding us to build robust models— models that can withstand scrutiny from multiple independent perspectives.