An Interventionist Approach to Psychological Explanation

Michael Rescorla

Abstract: *Interventionism* is a theory of causal explanation developed by Woodward and Hitchcock. I defend an interventionist perspective on the causal explanations offered within scientific psychology. The basic idea is that psychology causally explains mental and behavioral outcomes by specifying how those outcomes would have been different had an intervention altered various factors, including relevant psychological states. I elaborate this viewpoint with examples drawn from cognitive science practice, especially *Bayesian perceptual psychology*. I favorably compare my interventionist approach with well-known *nomological* and *mechanistic* theories of psychological explanation.

§1. Explaining psychological phenomena

What is it to explain a mental or behavioral phenomenon? For example, suppose we want to explain why an object looks to have a certain shape. What features should we expect from an explanation of the perceptual shape-estimate? More generally, what features should we expect from explanations of perception, belief, action, language acquisition, memory, problem solving, decision-making, and other core psychological activities? What makes one psychological explanation better than another? To what extent does cognitive science already supply good psychological explanations of mental and behavioral phenomena?

How one answers these questions depends heavily upon one's background views regarding scientific explanation more generally. Some authors espouse a *nomological* approach (Fodor, 1981; 1987; 1994): psychological explanation deploys *psychological laws* that subsume the explanandum. Others espouse a *mechanistic* approach (Bechtel, 2008): psychological explanation identifies underlying *mechanisms* that produce the explanandum. I will advance an account grounded in *interventionism*, a theory of causal explanation developed by Woodward (2003) and Woodward and Hitchcock (2003a; 2003b). The basic idea is that psychology causally explains an explanandum by specifying how the explanandum would have been different had certain causal influences (such as sensory or psychological states) been suitably manipulated. I advance my account as a theory of *causal* explanation within psychology, leaving open that psychology may also offer *non-causal* explanations. Notwithstanding this restricted focus upon causal explanation, I will argue that an interventionist conception improves markedly upon the nomological and mechanistic conceptions.

Interventionism is already well-established as an appealing theory that honors many aspects of scientific practice (Woodward, 2003). Several authors (Campbell, 2007), (Woodward, 2008a, 2008b), including myself (Rescorla, 2014), have previously applied it to mental activity. However, prior applications within philosophy of mind tend to focus mainly on *causation* rather than *explanation*. Relatedly, prior applications do not engage in much detail with cognitive science practice. As a result, they do not convey very well how explanation works in cognitive science, nor do they adequately highlight the explanatory benefits afforded by psychological inquiry. I hope that my discussion will foster more robust appreciation of those benefits.

§2 critically reviews the nomological and mechanistic conceptions of psychological explanation. §3 presents key features of interventionism. §§4-6 develop an interventionist conception of causal explanation within psychology, taking *Bayesian perceptual psychology* as an illustrative case study. §§7-8 favorably compares my interventionist conception with the nomological and mechanistic conceptions.

§2. Laws and mechanisms

According to the *deductive-nomological (DN) model* of scientific explanation, scientists explain a phenomenon by showing how to deduce it from *laws of nature* (Hempel, 1965). More precisely, DN explanations instantiate the following schema:

(1)

$$\begin{array}{l} L_1, \ldots, L_n \\ \underline{C_1, \ldots, L_k} \\ E \end{array}$$

E is the explanandum, $L_1, \ldots L_n$ are laws, and $C_1, \ldots C_k$ specify particular circumstances. For example, E might be the acceleration of a physical body, $L_1, \ldots L_n$ might be physical laws, and $C_1, \ldots C_k$ might be particular masses and locations in the physical system. By deducing E from $L_1, \ldots L_n$ and $C_1, \ldots C_k$, we show that *E was to be expected* given certain deep uniformities in nature and given certain particular circumstances. We thereby illuminate why E occurred. Thus, scientific explanation subsumes the explanandum under a *covering law*. Of course, any complete version of the DN model must say what counts as a "law."[1]

Fodor (1981; 1987; 1994) applies the DN model to psychological explanation. In his words: "psychological explanation typically involves law subsumption" (1994, p. 3). On Fodor's

[1] Hempel supplements the DN model with the *deductive-statistical (DS) model* and the *inductive-statistical (IS) model*. The DS model is really just a special case of the DN model, in which the laws take a statistical form. For the IS model, one does not *deduce* the explanandum from the explanantia. Rather, one shows that the explanandum was *likely* to occur (at least to some degree) given the explanantia. Technically speaking, the "nomological conception of scientific explanation" includes the IS model along with the DN model. However, addressing the IS model would complicate my exposition without affecting the main thrust of my argument.

picture, cognitive science should delineate laws that describe interactions among sensory inputs, psychological states, and behavioral outputs. We explain some mental or behavioral outcome by citing appropriate psychological laws, combined with details of the psychological system. Many other philosophers explicitly endorse or implicitly presuppose the nomological conception of psychological explanation (Antony, 1995), (Aydede, 2000), (Horgan and Tienson, 1990), (Pietroski and Rey, 1995), (Schneider, 2005). Much of the surrounding literature has centered upon matters such as: which properties psychological laws must have to support good explanations; whether we can find laws with these properties; the extent to which such laws are already implicit in folk psychological practice; how the requisite laws compare to laws found in other scientific disciplines; and so on.

Critics have launched various objections to the DN model construed both as a general theory of scientific explanation and more specifically as a theory of *psychological* explanation. Some highlights:

- Numerous powerful counterexamples establish that subsumption under a covering law does not suffice for explanation (Salmon, 1989), (Woodward, 2003, pp. 154-155). For example, we can deduce a flagpole's height from the laws of optics, the sun's position, and the length of the shadow cast by the flagpole --- yet this does not count as a genuine explanation of the flagpole's height. When we show that some explanandum *was to be expected*, we do not necessarily explain it.
- Philosophers widely agree that scientific practice features few truly exceptionless generalizations (Fodor, 1987; 1991a), (Pietroski and Rey, 1995), (Woodward, 2003). Accordingly, proponents of nomological explanation usually propose that we qualify laws with *ceteris paribus* clauses. Critics respond that *ceteris paribus* laws are too

empty, untestable, vague, or otherwise problematic to figure in satisfying scientific explanations (Earman, Roberts, and Smith, 2002).

- Many philosophers argue that psychological practice offers few if any explanatory generalizations that count as laws (Bechtel and Wright, 2009), (Dennett, 1993), (Gauker, 2005), (Schiffer, 1991), even if we allow laws to include *ceteris paribus* clauses. Obviously, the force of this worry depends on how we demarcate the laws.[2]

All three worries have been extensively discussed in the literature over the past few decades.

Responding to the DN model's perceived failures, philosophers propose various non-nomological theories of scientific explanation. Recently, a *mechanistic* approach has gained popularity (Bechtel, 2008), (Craver, 2006). The rough idea is that explanation decomposes a complex system into parts, describes how the parts are organized, specifies operations of the parts, and exhibits how the explanandum results from joint operation of the parts. Thus, explanation unveils causal mechanisms that help produce the explanandum. To unveil these mechanisms, we need not subsume the explanandum under anything resembling a law.

From a mechanistic viewpoint, psychological explanation should isolate components of the psychological system and describe how joint operations of those components produce some mental or behavioral outcome. As Bechtel and Wright put it, "[t]he major tasks in developing mechanistic explanations in psychology are to identify the parts of a mechanism, determine their operations, discern their organization, and finally, represent how these things constitute the system's relationship to the target explanandum" (2009, p. 120). Proponents usually emphasize *neural* parts (Bechtel and Wright, 2009), (Piccinini and Craver, 2011). On this approach, psychological explanation should isolate regions of the brain and specify how activity in those

[2] Bechtel and Wright (2009) note that the phrase "law" is seldom used in scientific psychology. This does not strike me as an important datum for philosophical theorizing, since a psychological generalization might satisfy the traditional philosophical criteria for lawhood even though psychologists do not call it a law.

regions produces the explanandum. In contrast, Stinson (2016) allows that the "parts" specified by mechanistic psychological explanation may be abstract cognitive items (e.g. memory registers) that do not map straightforwardly onto neural regions.

A recurring worry facing the mechanistic conception of scientific explanation is that many successful scientific explanations seem non-mechanistic. Consider *the ideal gas law*:

(2) $PV = nRT,$

where P is pressure, V is volume, n is moles of the gas, R is the ideal gas constant, and T is temperature. (2) seems to support good explanations. For example, we can use (2) to explain why a gas exerts the pressure that it does. When you learn (2) along with V, n, and T, you have learned something illuminating about P. Yet (2) does not isolate anything resembling a mechanism. (2) does not even decompose a gas into component parts. From a mechanistic perspective, (2) is not explanatory. A genuinely mechanistic explanation must instead deploy statistical mechanics, which describes the gas as a collection of tiny interacting particles. I acknowledge that statistical mechanics can augment the explanatory power of (2). It does not follow (and is not true) that (2) itself is unexplanatory. Surely there is some good sense in which (2) helps us explain pressure even when unaccompanied by statistical mechanical details. While the mechanistic conception may isolate *one important class* of scientific explanations, it does not tell the whole story.

This worry arises with particular force for cognitive science, where it is widely accepted that good explanation can abstract away from underlying neural and computational mechanisms. When cognitive scientists study perception, action, decision-making, concept learning, navigation, and numerous other core phenomena, they often operate at an abstract psychological level that looks quite non-mechanistic (Weiskopf, 2011). I will provide examples in §§4-8.

The foregoing considerations may not constitute definitive arguments against the nomological and mechanistic conceptions of psychological explanation. Still, it seems well worth exploring alternative options. In what follows, I will advance an alternative *interventionist* conception.[3]

§3. Interventionism

Interventionism is a theory of *causal* explanation, i.e. explanation that illuminates causal influences upon the explanandum. There may be non-causal scientific explanations (Lange, 2016), such as dimensional explanations or mathematical explanations. Interventionists acknowledge that non-causal explanation is potentially illuminating, but they only seek to elucidate causal explanation. According to interventionists, a causal explanation specifies how the explanandum would have been different had certain explanantia been suitably manipulated. Causal explanation answers *what if things had been different questions* (or *w-questions*) about the explanandum.

Interventionists codify this intuitive idea by talking about *variables* and *interventions*. The values of a variable are possible states of the system under consideration. A variable must have at least two values. Roughly, an intervention on a variable X is an idealized experimental manipulation of X's value. Slightly more carefully, an intervention on variable X with respect to

[3] The literature offers several additional theories of scientific explanation, such as the *unificationist* conception (Kitcher, 1989) and the *kairetic* conception (Strevens, 2008). There is not enough space to discuss all existing theories in a single paper, so I have focused upon the two rival theories that seem to have been most influential within philosophy of cognitive science: the nomological and mechanistic conceptions. The unificationist conception faces serious problems (e.g. Woodward, 2003, pp. 358-373), and in any event it never found wide application among philosophers concerned with psychological explanation. Strevens (2008, pp. 464-468) briefly addresses how the kairetic conception applies to psychological explanation. He focuses exclusively on high-level propositional attitudes. He holds that psychological properties of propositional attitudes are *noncausally explanatorily relevant* to mental and behavioral outcomes. In contrast, I think that many good causal explanations found within cognitive science cite causally relevant psychological properties of mental states. My discussion is devoted to developing that viewpoint. More detailed discussion of the unificationist and kairetic conceptions must await another occasion.

variable Y is a change in X's value that changes Y's value *if at all* only through the change in X and not through an independent causal route. Specifically, an intervention must not change any confounding variables. Woodward (2003, pp. 94-114) offers a detailed theory of interventions. For present purposes, I forego further elucidation.

According to interventionists, we causally explain Y's value by revealing how interventions on some variable X would alter Y. An intervention on flagpole height with respect to shadow length (say, by extending the flagpole) yields determinate changes in shadow length. An intervention on shadow length with respect to flagpole height (say, by distorting the sun's rays with a prism) does not yield determinate changes in flagpole height. Intuitively, manipulating the flagpole's height is a way of manipulating its shadow, but manipulating its shadow is not a way of its manipulating its height. That is why one can explain shadow length in terms of flagpole height but not vice-versa.

Interventionists replace (1) with a similar but improved schema (Woodward and Hitchcock, 2003b). According to interventionism, science aims for explanations of the form

(3)
$$\begin{array}{l} f(X_1, X_2, ..., X_n) = Y \\ \underline{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n} \\ Y = y = f(x_1, x_2, ..., x_n) \end{array}$$

Here X_1, X_2, ..., X_n are *explanantia variables* and Y is the *explanandum variable*. The first line of (3) is an *explanatory generalization* that describes how Y's value depends upon the values of X_1, X_2, ..., X_n. A causal explanation satisfies two conditions. First, X_1, X_2, …, X_n, Y actually have the respective values x_1, x_2, …, x_n, y. Second, the generalization $f(X_1, X_2, ..., X_n) = Y$ specifies with at least approximate accuracy a non-trivial change in Y that results from interventions on X_1, X_2, ..., X_n. To clarify the second condition, say that a *test intervention* is an intervention that alters at least one of the variables X_1, X_2, ..., X_n, where the generalization $f(X_1, X_2, ..., X_n) = Y$ predicts that the altered values yield a *different* value for Y than its actual value. A *test counterfactual* specifies which value of Y is predicted, assuming that the relevant test intervention were to occur. In other words, a test counterfactual has the form:

> If an intervention had fixed $X_1 = x_1*$, $X_2 = x_2*$, …, $X_n = x_n*$, then it would have been the case that $Y = y*$,

where $x_i \neq x_i*$ for at least one i and where $y \neq y*$. The second condition requires that our explanation generate at least one approximately true test counterfactual.[4]

Interventionism sets a fairly low threshold for explanation. As long as we secure at least one approximately true test counterfactual, we have illuminated the explanandum by clarifying how it depends upon certain explanantia. We have answered at least one relevant w-question. Obviously, it would be desirable to answer more than a single w-question regarding the explanandum. Interventionists hold that a causal explanation is *better* or *deeper* to the extent that it generates a larger class of approximately true test counterfactuals.

As noted in §2, scientific practice features very few exceptionless generalizations. Interventionists respond by emphasizing *background conditions*. To illustrate, consider the formula for period of a pendulum:

(4) $$T = 2\pi\sqrt{\frac{L}{g}}\ ,$$

where T is period, L is length of the pendulum, and g is local acceleration of gravity. (4) only prevails when suitable background conditions are in place: amplitude is not too large; no obstacle impedes the pendulum; wind resistance is not too great; and so on. *Assuming background*

[4] Woodward (2003, pp. 209-220) extends interventionism to encompass singular causal claims, such as "The short circuit caused the fire." Woodward holds that singular causal claims answer certain w-questions and hence are minimally explanatory. We need not evaluate this aspect of Woodward's position.

conditions where such exceptional factors do not arise, we can use (4) to say how period would have been different if length had been different. More generally, an explanatory generalization *G* will typically prevail only against appropriate background conditions, and we will generally not be able to articulate those background conditions fully and explicitly. Nevertheless, we can use *G* to answer numerous w-questions. We can do so whenever we are confident that appropriate background conditions obtain (as scientific practice shows that we often are).

Note here an important contrast between interventionism and the traditional literature on nomological explanation. Proponents of the DN model try to salvage the truth of explanatory generalizations by introducing *ceteris paribus* clauses. In contrast, interventionists concede that explanatory generalizations used in science are often false. For example, (4) is false, because it fails when appropriate background conditions do not obtain. Interventionists hold that false explanatory generalizations can be explanatory. A false generalization is explanatory when appropriate background conditions obtain, so that the generalization yields approximately true counterfactuals. Although this analysis may initially sound counterintuitive, it seems to accord rather well with actual scientific practice.

§3.1 Laws and counterfactuals

Some readers may suggest that interventionism is a variant of the DN model. On this analysis, interventionists do not so much reject the DN model as offer a novel account of "law." We can rescue the DN model simply by defining "law" along the following lines: *explanatory generalization that generates at least one approximately true test counterfactual*.

The issue here is partly terminological. Clearly, one may stipulatively define "law" however one pleases. As Woodward emphasizes (2003, pp. 285-288), though, proponents of the

DN model have *not* traditionally defined “law” by invoking test counterfactuals. They have instead chosen other defining properties. As a result, the traditional literature on laws obfuscates the boundary between explanatory and unexplanatory generalizations.

The defining property most commonly cited in the traditional literature is that laws “support counterfactuals.” This feature is supposed to demarcate “lawlike” generalizations from mere “accidental” generalizations. The generalization

Water boils at 100° Celsius.

is lawlike, because it “supports” the counterfactual

If we were to heat this sample of water to 100° Celsius, then it would boil.

The generalization

All the coins in my pocket are quarters.

is accidental, because it does not “support” the counterfactual

If we were to add this coin to my pocket, then it would be a quarter.

Proponents of the DN model assign great weight to the distinction between generalizations that “support counterfactuals” and generalizations that do not. Yet the distinction does not in fact seem like the crucial one we should be studying, because many unexplanatory generalizations “support counterfactuals.” For example, Salmon’s (1971) famous generalization

(5) All males who take birth control pills regularly fail to get pregnant.

cannot be used to explain why a man fails to become pregnant after taking birth control pills, even though it supports a counterfactual along the following lines:

If John were to take the birth control pill regularly, then he would fail to get pregnant.

From an interventionist viewpoint, the traditional literature errs by emphasizing counterfactuals *in general* rather than the sub-category of *test counterfactuals*. (5) supports counterfactuals, but it does not support approximately true *test* counterfactuals. Intervening on whether John takes the birth control pill would have no impact on whether he becomes pregnant. That is why (5) is unexplanatory even while it supports certain counterfactuals.

Readers are free to define the phrase "law" however they like. Speaking for myself, I find it most helpful to eschew this phrase along with its traditional connotations. Doing so encourages us to emphasize more important questions, such as: *When does an explanatory generalization help explain some phenomenon?* Interventionism sheds considerable light upon this question --- far more so than traditional versions of the DN model.

§3.2 Causation and mechanism

Interventionism about causal explanation comes packaged with an appealing theory of *causal relevance*. The rough idea: X is causally relevant to Y just in case Y's value would change if suitable interventions altered X's value. For example, flagpole height is causally relevant to shadow length, because intervening on flagpole height yields a determinate change in shadow length. Woodward (2003, pp. 25-93) develops these ideas at length. We need not delve into the details. The key point is that, if we accept an interventionist theory of causal relevance, then an interventionist explanation (3) adduces at least one variable that is causally relevant to the explanandum variable. If $X = (X_1, X_2, ..., X_n)$ is a vector of the explanantia variables, then (3) is explanatory only if some test intervention on X would yield a change in Y's value. So (3) is explanatory only if X is causally relevant to Y. Individual explanantia variables X_i may be causally relevant to Y as well.[5]

[5] Strictly speaking, one can embrace interventionism about causal explanation without embracing interventionism about causal relevance (Saatsi and Pexton, 2013). However, much of the motivation for interventionism about causal explanation lies in the nexus with interventionism about causal relevance.

An intervention on *X* with respect to *Y* cannot alter any variables that are causally relevant to *Y* and that lie on a causal route independent of the causal route (if any) from *X* to *Y*. To state this constraint, we must cite the relation of causal relevance between variables. Since *intervention* is elucidated in terms of *causal relevance* and *causal relevance* is elucidated in terms of *intervention*, interventionists do not purport to have supplied a non-circular analysis for either notion. Instead, interventionists want to illuminate how the two notions relate to one another and to the scientific practice of supplying causal explanations.

By placing causation at center stage, interventionism accords with the mechanistic conception. Interventionism differs from the mechanistic conception by denying that explanation must decompose a system into *components* whose joint operations produce the explanandum. Take the ideal gas law. (2) yields numerous test counterfactuals, e.g. counterfactuals describing how pressure *P* exerted by the gas would change if an intervention altered volume *V* or temperature *T*. So interventionists regard (2) as explanatory. They hold that it underwrites *non-mechanistic causal* explanations that illuminate how causally relevant variables (*V* and *T*) impact *P*. Thus, while interventionists and mechanists both emphasize causal antecedents that produced the explanandum, interventionists hold that we can illuminate those causal antecedents without limning anything like an underlying mechanism. Although interventionists deny that good explanation must expose underlying mechanisms, they agree that exposing those mechanisms often *improves* explanation (Woodward, forthcoming). Supplementing a non-mechanistic explanation with mechanistic details will often generate additional test counterfactuals. For example, statistical mechanical explanation of pressure improves upon (2). It generates additional test counterfactuals that describe how *P* would change if an intervention altered the velocities of individual particles. It also describes how *P* would change in various conditions where (2) breaks down. Thus, interventionists hold that non-mechanistic causal explanation is possible but that mechanistic details often improve the explanation.

Often, not always. From an interventionist perspective, adding further mechanistic details to an explanation need not improve it and may even make it worse (Woodward, 2008b).

To illustrate, suppose that a bridge will collapse precisely when the weight on it exceeds 5000 kg. If I want to explain why the bridge collapsed, then adducing the specific weight 8356 kg seems like no advance over saying that the weight exceeded 5000 kg. Indeed, the more specific explanation seems worse to the extent that it misleadingly intimates explanatory import for the particular weight 8356 kg. From an interventionist perspective, these intuitive verdicts are readily explicable. I do not generate additional approximately true test counterfactuals by citing the particular weight 8356 kg, so I gain no explanatory benefit by citing it. Moreover, depending on the context, citing the particular weight 8356 kg may misleadingly suggest that any lower weight would not have caused the bridge to collapse. Moral: adding further details about the explanans does not always improve an explanation.[6]

[6] If we regiment explanation using variables, then there are at least four putative explanations of the bridge collapse to consider. The first cites a binary variable *T* whose two values reflect whether the weight on the bridge exceeds 5000 kg. The second cites a binary variable *U* one value of which is 8356 kg and the other value of which corresponds to all other possible weights. The third cites a binary variable *V* whose two values reflect whether the weight was less than 8356 kg. The fourth cites a continuum valued variable *W* whose values are all possible weights. *T* supports a good explanation of why the bridge collapsed. *U* does not: there is no determinate answer as to whether the bridge would collapse if an intervention altered the weight from 8356 kg, because the answer depends on whether the altered weight exceeded 5000 kg. Similarly for *V*. *W* can figure in good explanations of why the bridge collapsed, since each possible value of *W* has a determinate implication for whether the bridge collapses. The intuitive statement “The bridge collapsed because the weight on it was 8356 kg” is misleading to the extent that it suggests a regimented explanation using *U* or *V*, acceptable to the extent that it suggests a suitable regimented explanation using *W*. The question remains: how do explanations that cite *T* compare to explanations that cite *W*? Woodward (2008b) suggests that causal explanations are better when they are “proportional,” meaning roughly that they describe the explanans in just enough detail to explain the explanandum. From this perspective, an explanation that cites *T* is superior to an explanation that cites *W*. Franklin-Hall (2016) argues against proportionality. I remain neutral regarding proportionality. In particular, I remain neutral as to whether *T* is a better explanans variable than *W*. What matters for my purposes is that *W* does not seem like a better explanans variable than *T*. We gain no explanatory benefit by citing the fine-grained *W* rather than the binary *T*. (Thanks to an anonymous referee for suggesting that I discuss the bridge example.)

In particular, adding further *mechanistic* details does not always improve an explanation. For example, economists explain the inflation rate using models that describe how factors such as the money supply, interest rates, and inflationary expectations influence price levels (Mankiw, 1997, pp. 146-182). They also model how factors such as currency levels and reserve-deposit ratio requirements influence the money supply (Mankiw, 1997, pp. 475-493). These models yield numerous approximately true test counterfactuals concerning the inflation rate. To render the models more mechanistic, we might add further mechanistic details about how the money supply changes. We might describe gear configurations of currency printing presses, or silicon chip states in the computers through which the central bank communicates reserve-deposit ratio requirements, and so on. However, economic explanation invariably neglects gear configurations and silicon chip states. From an interventionist viewpoint, the neglect is readily explicable. In adducing specific printing press gear configurations that occurred when the money supply changed, we do not thereby generate new approximately true test counterfactuals involving economic variables. Depending on the context, we may even misleadingly suggest that any different gear configurations would have caused a different economic outcome. Details about gear configurations or silicon chip states do not by themselves improve explanation of inflation, the money supply, or any other economic variable.

Some mechanists might acknowledge that details about printing press gear configurations do not improve economic explanation. They might say that only certain privileged mechanistic details improve explanation. They must then answer the question: which mechanistic details improve explanation? Interventionism offers a principled, systematic answer: mechanistic details improve explanation when they help us answer additional w-questions about the explanandum.

§3.3 Incorporating interventionist notions into the mechanistic account?

Craver (2006, 2014) advances a mechanistic account that incorporates interventionist elements. He distinguishes two kinds of scientific description:

- *Phenomenal models*, such as Snell's law, which summarize a body of data without providing any explanation for the data.
- *Mechanistic explanations*, which limn the mechanisms that produce an explanandum.

"In many cases," Craver writes, "the distinction between explanatory and non-explanatory models is that the latter, and not the former, describe mechanisms" (2006, p. 367). He grants that scientific explanation sometimes provides only a *mechanism sketch* that "characterizes some parts, activities, and features of the mechanism's organization" while leaving numerous mechanistic details unspecified (2006, p. 360). He acknowledges that the word "explanation" is used in various ways within scientific practice and that not all usages align with the mechanistic conception (2006, p. 367). Still, he insists that mechanistic explanation has a privileged status, because it constitutes a normative ideal to which scientific practice aspires (2014, pp. 37-41). Scientific models are explanatory only to the extent that they reach this normative ideal.

Craver develops his mechanistic viewpoint using interventionist ideas. He says that mechanistic models are explanatory because they answer a wide range of w-questions (2006, p. 358, p. 374). He acknowledges that phenomenal models "typically allow one to answer some w-questions" (2006, p. 358), but he says that they do not answer a suitably wide range of w-questions to count as explanatory. Thus, while interventionists hold that an explanation need only yield a *single* approximately true test counterfactual, Craver holds that an explanation must yield a suitably wide range of approximately true test counterfactuals. A natural question for Craver is why non-mechanistic "phenomenal" descriptions are non-explanatory even when they yield approximately true test counterfactuals. If mechanist models are explanatory by virtue of generating a wide range of approximately true test counterfactuals, then shouldn't a phenomenal model count as at least *a little* explanatory when it generates *some* approximately true test counterfactuals? Craver does not answer this question in a clear, consistent way.[7] He offers no principled basis for requiring explanations to meet such a high standard.

Another problematic aspect of Craver's position concerns the explanatory value contributed by additional mechanistic details. Craver appears to hold that adding mechanistic details always helps us answer additional w-questions and thereby improves explanation (2014, pp. 40-41). I critiqued that thesis in §3.2. In response to my critique, Craver might retrench to a weaker thesis: adding mechanistic details improves an explanation *just in case* the new details help us answer additional relevant w-questions. I endorse this weaker thesis. I also note that the weaker thesis is not very congenial to Craver's mechanistic viewpoint. According to the weaker thesis, the mechanistic details that improve explanation are simply those details that yield new approximately true test counterfactuals. The explanatory power of a scientific model is determined by the test counterfactuals that it generates, not the amount of mechanistic detail that it incorporates. Better then to jettison mechanistic locutions and emphasize what really matters: the w-questions that a scientific theory answers.

§4. Interventionist explanation within scientific psychology

The rest of the paper develops an interventionist conception of causal explanation within psychology. Psychology may also provide *non-causal* explanations, but I focus solely on causal psychological explanations, i.e. psychological explanations that illuminate causal influences

[7] Craver writes in one passage that "phenomenal models are at best shallow explanations" (2006, p. 374), which suggests that they may be explanations after all (even if not particularly satisfying ones). However, this passage clashes with Craver's repeated, emphatic insistence that phenomenal models are unexplanatory.

upon the explanandum.[8] I will pursue the following idea: *scientific psychology causally explains mental and behavioral outcomes by specifying how those outcomes would have been different had an intervention altered various factors, including relevant psychological states*.

Cognitive science invokes a wide range of psychological states, events, and processes. Here is an extremely partial list:

- *Ordinary propositional attitudes*, e.g. belief, desire, and intention.
- *Perception*, e.g. a perceptual estimate of the shape, color, size, or location of some perceived object.
- *Mental imagery*, e.g. rotating a mental image to compare it with another mental image.
- *Navigation,* e.g. updating a cognitive map through dead reckoning.
- *Computational states* posited by Turing-style models, e.g. storing a mental representation in a memory register.
- *Associations*, e.g. an associative bond of a certain strength between a conditioned stimulus and an unconditioned stimulus.
- *Activation weights* among nodes in a connectionist network.

Each item in this list has proved controversial at various points in intellectual history. I offer the list simply to evoke the varied posits found in scientific psychology.

[8] Cummins (2000) emphasizes a mode of psychological explanation that he calls *functional analysis*, which explains a psychological capacity (e.g. the capacity to perceive depth, or to speak English) by decomposing it into less sophisticated capacities. Functional analysis is arguably an example of *non-causal psychological explanation*: it explains a psychological capacity not by specifying causal influences upon the capacity but rather by revealing how the capacity decomposes into more basic capacities. While I agree with Cummins that functional analysis plays a role in cognitive science practice, I do not think that it exhausts psychological explanation. In many cases, the explananda of interest to scientific psychology are not *capacities* but *particular states or events*. For example, we might want to explain why someone perceived an object as having a certain depth, or why she understood a particular English utterance as expressing a particular propositional content. Cummins does not say what it takes to explain such outcomes. He focuses exclusively upon functional analysis of psychological capacities.

We may formalize these posits using variables. For example, we may introduce variables corresponding respectively to

- Possible intentions (e.g. intentions to reach one's arm to various possible locations).
- Possible aspects of perception (e.g. possible perceptual estimates of an object's shape).
- Possible orientations of a mental image.
- Possible cognitive maps of the environment.
- Possible contents of some memory register.
- Possible strengths of the associative bond between two stimuli.
- Possible activation weight between two nodes in a connectionist network.

These are *psychological variables*, in the sense that their values reflect possible psychological states or events.

Cognitive science also cites *non-psychological* aspects of the individual or the environment. Here is an extremely partial list:

- *Distal stimuli* (e.g. the size or color of a perceived object).
- *Proximal stimulations* of the individual's sensory organs (e.g. a pattern of retinal stimulation).
- *Bodily motions* (e.g. the trajectory of one's arm).
- *Motor commands* (i.e. electromagnetic impulses transmitted from the brain to the musculature).
- *Efference copy* (i.e. a copy of a motor command, transmitted back to the brain).

In each case, one can introduce a *non-psychological variable* whose values reflect possible non-psychological states or events (e.g. a variable that reflects possible arm trajectories).

I will not try to clarify the boundary between psychological and non-psychological variables. Some cases may be borderline, indeterminate, or controversial. In practice, it is usually fairly clear whether a variable is psychological or non-psychological.[9]

Psychological variables figure as both explananda and explanantia in scientific psychology. One might want to explain why an individual acquires some belief; or why the perceptual system estimates that an object has a certain shape; or why an associative bond of a certain strength is formed. Or one might cite an individual's perceptual states to explain her beliefs; or her beliefs and desires to explain her intentions; or her cognitive map to explain why she forms a plan to move in a certain direction. Psychological explanation may also employ non-psychological variables as explananda (e.g. one can cite intentions to explain bodily motions) or explanantia (e.g. one can cite proximal sensory stimulation to explain perceptual states).

On an interventionist conception, scientific psychology causally explains by delineating explanatory generalizations that describe the interaction among psychological and non-psychological variables. The generalizations yield true (or approximately true) test counterfactuals specifying how some psychological or non-psychological explanandum variable would change were interventions to alter certain psychological or non-psychological variables. The generalizations conform to schema (3), where X_1, X_2, ..., X_n and Y may be either psychological or non-psychological variables. At least one variable must be psychological in order for the explanation to count as psychological. Thus, the boundary between psychological and non-psychological explanation is only as clear as the boundary between psychological and non-psychological variables. I am not trying to clarify those boundaries. I am trying to clarify

[9] *Neural variables* describe possible neural states. Are any neural variables also psychological variables? That depends on whether any neural states are psychological states --- more carefully, on whether any neural state-types are psychological state-types. This is a controversial question. Luckily, I do not need to take a stand here. Nothing in my treatment turns upon whether any neural variables are psychological variables.

what it takes for psychological generalizations to yield good causal explanations, assuming an antecedent demarcation of the psychological.[10]

The previous paragraph delineates an interventionist template for psychological explanation. My first main thesis is *normative*: causal psychological explanation should conform to the interventionist template. This thesis follows from interventionism more generally. My second main thesis is *descriptive*: current cognitive science already offers numerous explanations that conform well to the interventionist template. I now defend the second thesis.

Experimental psychology isolates various "effects," such as the *McGurk effect* (MacDonald and McGurk, 1978), the *Garcia effect* (Garcia and Koelling, 1966), the *ventriloquism effect* (Alais and Burr, 2004), the *spacing effect* (Madigan, 1969), the *Stroop effect* (Stroop, 1935), the *bystander effect* (Darley and Latané, 1968), and many others. We can summarize these effects through generalizations couched with varying degrees of precision and rigor. Great mathematical precision is sometimes possible, especially within perceptual psychology. Consider *Weber's law* (Palmer, 1999, pp. 671-672):

(6) $$\frac{\Delta I}{I} = k \,,$$

where I is the magnitude of some distal stimulus dimension (e.g. the length of a line), ΔI is the *just noticeable difference* (JND) for the stimulus, and k is a constant called the *Weber fraction*. (6) is really a generalization schema, since different stimuli have different Weber fractions. Rearranging terms in (6), we obtain

[10] *Folk psychology* offers numerous singular psychological explanations of mental and behavioral outcomes (e.g. "John went to the restaurant because he wanted to meet Sam there.") How interventionists should assess these singular explanations depends, in part, on the general issues about singular causal explanation raised in note 4. These matters deserve their own dedicated paper. But it seems clear that anything resembling *scientific* psychological explanation requires generalizations rather than mere singular causal statements. Folk psychology also offers various platitudes, such as the belief-desire law. Whether those platitudes are (or can be converted into) generalizations that conform to the interventionist template (3) is a tricky question that I will not address here.

(7) $\Delta I = kI$,

which displays the JND as a function of the stimulus magnitude. (7) generates test counterfactuals that describe how the JND would change if an intervention were to change the stimulus magnitude. From an interventionist viewpoint, (7) explains JNDs. A similar diagnosis applies to many other generalizations that summarize psychological effects, although few such generalizations match the precision, accuracy, and scope of Weber's law.

Cummins (2000) offers an opposing analysis. He denies that one can explain a psychological explanandum through a generalization that summarizes some psychological effect (2000, p. 119):

> No laws are explanatory in the sense required by DN. Laws simply tell us what happens; they do not tell us why or how... In psychology, such laws there are are almost always conceived of, and even called, effects: the Garcia effect (Garcia, 1966), the spacing effect (Madigan, 1969), the McGurk effect (MacDonald and McGurk, 1978), and many, many more. But no one thinks that the McGurk effect explains the data it subsumes. No one not in the grip of the DN model would suppose that one could explain why someone hears a consonant like the speaking mouth appears to make by appeal to the McGurk effect. That just *is* the McGurk effect.

Bechtel (2008) and Craver (2006) concur.

In contrast, I maintain that summaries of psychological effects are sometimes explanatory. A well-chosen summary does not just tell us what happens. It tells us what *would* have happened had an intervention altered certain factors. It thereby illuminates the explanandum. To illustrate:

- We do not explain *the McGurk effect itself* by citing the McGurk effect. But we can explain *particular perceptual events* by invoking the McGurk effect. For example, suppose we want to explain why someone heard /da/ even though /ba/ was uttered. A proper appeal to the McGurk effect illuminates this perceptual event by pinpointing a key causally relevant factor: visual appearance of lip movements associated with /ga/. Our explanation is illuminating because it supports helpful test counterfactuals (e.g. the perceiver would *not* have heard /da/ had the lips *not* looked like they were articulating /ga/).
- Weber's law helps us explain why a stimulus has a certain JND. When we use Weber's law to derive the JND, we pinpoint a key causal influence upon the JND: the stimulus magnitude. We also specify how the JND would have changed had an intervention altered the stimulus magnitude. We thereby provide valuable information that illuminates why a certain JND arises.

A generalization that merely summarizes some psychological effect can be explanatory.

Still, I think that Cummins is right to find such generalizations unsatisfying. They are clearly rather shallow and superficial. We would like deeper psychological explanations. From an interventionist perspective, there is a straightforward reason why mere summaries of psychological effects seem unsatisfying: they answer a fairly limited range of w-questions. For example, a summary of the McGurk effect does not specify how speech perception would have changed under various further manipulations: interventions that insert a temporal lag between lip movements and sound (as in a badly dubbed movie); interventions that alter the perceiver's attention; interventions that blur visual input; and so on. Analogous points apply to other generalizations that merely summarize psychological effects. At best, these summaries generate relatively few test counterfactuals. Even when the summaries are explanatory, they are *minimally* explanatory. A satisfying scientific psychology should move beyond minimally explanatory generalizations, articulating generalizations that answer a wider range of w-questions.

To what extent does current scientific psychology already do so? In some areas (e.g. social psychology), the explanatory generalizations on offer arguably do not go much deeper than careful summaries of psychological effects. In other areas, though, psychology has isolated far more informative generalizations. §5 illustrates by discussing *Bayesian models of perception*.

§5. Explaining perception

The perceptual system estimates distal conditions (e.g. shapes, sizes, colors, and locations of perceived objects) based upon proximal sensory stimulations (e.g. retinal stimulations). As a simple example, consider *shape from shading*. Suppose that you perceive an object whose shading is compatible with two conflicting hypotheses: light comes from overhead and the perceived object is convex; or light comes from below and the perceived object is concave. Then you will perceive the object as convex rather than concave (Palmer, 1999, pp. 244-245). Thus, a certain pattern of retinal illumination is reliably mapped by your perceptual system into a percept that estimates convexity (rather than concavity). *Perceptual psychology* studies the mapping from proximal sensory stimulations to perceptual estimates. The science offers numerous extremely precise generalizations that describe how aspects of the percept (e.g. perceived shape, size, color, or location) depend on proximal sensory input.

Generalizations along these lines play a central role within Burge's (2010) analysis of perceptual psychology. Burge emphasizes what he calls *formation laws*: "laws that determine transformation of sensory registrations --- sensory states that correlate highly with a type of stimulation --- into perceptual representational states with representational content" (2010, p. 345. Burge (2010, p. 383) writes that perceptual psychology "explains by citing laws or law-like patterns of operation that lead from given registrations of proximal stimulation to perceptual states that specify particulars as having specific attributes." Formation laws dictate which percepts ensue from which sensory registrations. Perceptual psychology explains by subsuming percepts under appropriate formation laws.

I set aside the question whether "formation laws" should be called "laws." The key point I want to emphasize is that, from an interventionist perspective, many formation laws *do* seem explanatory. A well-chosen formation law exhibits how intervening on proximal sensory input would alter the percept --- e.g. how intervening on shading would alter perceived shape. Suitable formation laws illuminate the percept by identifying causally relevant antecedents (salient aspects of the proximal stimulus) and by exhibiting how the percept would have been different had we suitably manipulated those antecedents.

Nevertheless, we would like to move beyond formation laws that merely dictate which proximal sensory inputs are mapped into which perceptual states. How exactly does the perceptual system transform proximal sensory inputs into perceptual states? To answer this question, we must "look inside the black box," clarifying the psychological processes that mediate between sensory inputs and percepts.

§5.1 Bayesian perceptual psychology

Helmholtz (1867) proposed that the perceptual system executes an *unconscious inference* from sensory input to a "best" hypothesis regarding distal conditions. In the 1990s, perceptual psychologists began using *Bayesian decision theory* to develop Helmholtz's proposal.

Bayesian decision theory hinges on the notion of *subjective probability*. *Bayes's Rule* describes how one should update subjective probabilities in light of new evidence:

> ***Bayes's Rule***: When one receives evidence e, one should update $p(h)$ by replacing it with $p(h \mid e)$.

$p(h)$ is the probability of h --- usually called the *prior probability* --- and $p(h \mid e)$ is the *probability* of h given e --- usually called the *posterior probability*. An invaluable aide to computing the posterior is Bayes's theorem:

> ***Bayes's Theorem***: $p(h \mid e) \propto p(e \mid h) p(h)$,

meaning that the left-hand side is proportional to the right-hand side. Bayes's theorem expresses the posterior in terms of the prior probability and the *prior likelihood* $p(e \mid h)$. Another key element of Bayesian decision theory is *expected cost minimization* (equivalently, *expected utility maximization*). This rule instructs one to choose an action that minimizes expected cost, where "expected cost" is determined by one's probabilities and a cost function.

On a Bayesian approach, perception executes an unconscious *statistical* inference (Feldman, 2015), (Knill and Richard, 1996), (Rescorla, 2015). The perceptual system allocates probabilities over hypotheses drawn from a hypothesis space, where each hypothesis h reflects some aspect of the distal environment (e.g. the shape of a perceived object). The perceptual system encodes prior probabilities $p(h)$ and prior likelihoods $p(e \mid h)$, where each e corresponds to possible sensory input (e.g. possible retinal stimulation; possible proprioceptive input). After receiving input e, the perceptual system reallocates probabilities across the hypothesis space in rough accord with Bayes's Rule, yielding a posterior $p(h \mid e)$.

To illustrate, consider shape from shading. Let *s* reflect possible shapes, θ reflect possible lighting directions, and *e* reflect possible patterns of retinal illumination. Stone (2011) offers a Bayesian model with the following elements:

A prior probability $p(s)$ over possible distal shapes.

A prior probability $p(\theta)$ over possible lighting directions. The prior assigns higher probability to overhead lighting directions.

A prior likelihood $p(e \mid s, \theta)$, which codifies the likelihood that distal shape *s* and lighting direction θ cause retinal illumination *e*.

Upon receiving retinal illumination *e*, the perceptual system reallocates probabilities in rough accord with Bayes's Rule, yielding a posterior $p(s \mid e)$.

Any Bayesian perceptual model must describe how the perceptual system transits from the posterior $p(h \mid e)$ to a specific perceptual estimate $\hat{h}$. For example, a Bayesian model of shape perception must describe how the perceptual system transits from a posterior over possible shapes to a specific shape-estimate that goes into the percept. Bayesian models vary in how they handle the transition from the posterior to the estimate $\hat{h}$. The most common modeling strategy posits unconscious expected cost minimization. The "action" is selection of estimate $\hat{h}$. The cost function reflects various possible factors: the penalty for an incorrect answer; which distal properties are more "important"; and so on. If one chooses a suitable cost function, then expected cost minimization reduces to a simpler decision rule (e.g. selecting the median of the posterior).

Perceptual psychologists have employed the Bayesian paradigm to illuminate numerous perceptual phenomena, including perceptual estimation of size, shape, orientation, weight, color, speed, location, depth, and so on.

§5.2 Priors as explanantia

Bayesian perceptual modeling exhibits how key features of the percept depend upon the priors. We may schematize many Bayesian models using an equation of the form:

(8) $\hat{h} = \Phi(\textit{prior probability}, \textit{prior likelihood}, e)$.

(8) depicts perceptual estimate $\hat{h}$ (e.g. estimated shape, size, color, or location) as a function of three variables: the prior probability; the prior likelihood: and proximal sensory input *e*. (8) displays how $\hat{h}$ would have changed had an intervention altered one of those variables. For example, a suitable Bayesian model of shape perception can answer the questions *How would the perceptual shape-estimate have been different had an intervention altered the prior p(θ) over lighting directions?* and *How would the final shape-estimate have been different had an intervention altered the prior p(s) over shapes?* Bayesian perceptual models isolate crucial explanantia (perceptual priors) that causally influence the explanandum (the percept), and they describe how the explanandum would have been different had an intervention altered the explanantia. They thereby answer numerous w-questions.

If we hold the priors fixed, then we can convert an equation of the form (8) into an equation of the form:

(9) $\hat{h} = \Gamma(e)$.

(9) describes a mapping from proximal sensory inputs *e* to perceptual estimates $\hat{h}$. Priors do not figure as explicit variables in (9). Instead, they are built into our choice of Γ. Different priors induce a different Γ. (8) explicitly depicts how the percept would have changed had the priors changed, while (9) does not. So (8) generates a much wider range of test counterfactuals than (9). (8), unlike (9), exhibits how the percept systematically depends upon the priors. From an interventionist perspective, (8) is far more explanatory than (9). This interventionist verdict

accords with actual scientific practice, which does not rest content with (9) but instead strives to articulate the deeper generalization (8).

In summary, Bayesian perceptual models specify how perceptual outcomes would have been different had an intervention altered the subject's prior probabilities or prior likelihoods. By answering so many w-questions, Bayesian models go far beyond the minimally explanatory generalizations surveyed in §4.

§5.3 Intervening on priors

Interventionism requires that there be a well-defined notion of "intervention" for each explanans variable. Intervening on sensory input *e* is relatively straightforward. For example, one can intervene on retinal input simply by altering the light that hits the retina. But how does one intervene on the priors?

It is well-established that experience can alter the priors. When environmental statistics change, the priors change. However, an "intervention" must target a specific explanans variable without altering any other variable that independently influences the explanandum. For example, an intervention on the prior probability must not alter the prior likelihood, because the prior likelihood exerts independent influence on the percept. In a notable experiment along these lines, Adams, Graf, and Ernst (2004) manipulated the lighting direction prior $p(\theta)$. They used performance in a shape-estimation task to infer the peak of each subject's prior. They then exposed subjects to deviant stimuli indicating that the light source was shifted by as much as 30° from that peak. The result was an altered percept: the same stimulus caused a different shape-estimate before the experimental manipulation than it did afterwards. Performance also changed in a separate lightness-estimation task. Why did both shape-estimates and lightness-estimates change, even though deviant stimuli occurred only in the shape-estimation task? The best explanation, Adams, Graf, and Ernst (2004) argue convincingly, is that exposure to the deviant stimuli in the shape-estimation task caused a shift in $p(\theta)$, which affected performance in both tasks. In other words, the experimental manipulation changed the lighting direction prior. The experimental manipulation does not seem to have altered $p(S)$ or $p(e \mid s, \theta)$. So the experimental manipulation was plausibly an intervention (or close to an intervention) on $p(\theta)$ in the technical sense demanded by interventionists.[11]

Let us consider a more detailed example: *velocity estimation*. The perceptual system estimates velocities of distal objects based upon local measurements of retinal image motion. The inference from local motion measurements to distal velocities is a non-trivial problem, for several reasons. First, local motion measurement is noisy, especially in low contrast scenes. Second, a given local motion measurement is often compatible with multiple distal velocities. This is called *the aperture problem*: motion of an edge across an aperture is ambiguous, since many possible velocities are compatible with what one can measure through the aperture. See Figure 1. The perceptual system somehow transits from the overall pattern of local motion measurements to an estimate of distal velocity.

INSERT FIGURE 1 ABOUT HERE

[11] Experimental manipulation of the lighting prior alters additional mental states, especially ancillary beliefs. For example, the subject comes to believe that she participated in a psychology experiment. If Bayesian perceptual models are on the right track, then a change in ancillary belief influences the percept (if at all) only by altering the priors, the prior likelihoods, or the cost function. In the experimental manipulation performed by Adams, Graf, and Ernst (2004), $p(\theta)$ changes but $p(S)$ and $p(e \mid s, \theta)$ do not. The cost function also remains fixed. Thus, any changes in ancillary belief influence the percept (if at all) only by altering $p(\theta)$. For this reason, the experimental manipulation still counts as an intervention on $p(\theta)$ with respect to the percept even though it alters various ancillary beliefs.

Weiss, Simoncelli, and Adelson (2002) offer a Bayesian velocity estimation model that illuminates how the human visual system overcomes the aperture problem. The model applies in circumstances where one's eyes are fixated while observing a pattern moving in a two-dimensional fronto-parallel plane. Under these circumstances, the velocity *v* of the image cast by the pattern on the retina serves as a decent proxy for the pattern's own distal velocity. The model takes as input an *image intensity function I*(*x*, *y*, *t*), which gives the light intensity at retinal location (*x*, *y*) at time *t*. To estimate *v*, the model deploys the following elements:

- *Prior probability p*(*v*) *over velocities*. The model employs a prior that favors slow speeds, reflecting the environmental regularity that objects tend to move slowly. More specifically, the prior is a Gaussian centered at speed 0 with variance σ_p^2. A generalized model offered by Sotiropoulos, Seitz, and Seriès (2011) allows Gaussian priors with non-zero mean $\mu = (\mu_x, \mu_y)$, where μ_x and μ_y are the *x* and *y* components of velocity μ.
- *Prior likelihood p*(*I* | *v*). The model assumes that image intensity changes only because of the moving pattern's translational motion, not because points on the pattern change in apparent brightness. The model also assumes that local motion measurements are corrupted by Gaussian noise with variance σ^2.

The model combines the prior and prior likelihood into:

- *Posterior probability p*(*v* | *I*), reflecting the probability of velocity *v* given the overall pattern of retinal image intensity *I*.

Given input *I*, the model selects the velocity-estimate $\hat{v}$ that maximizes *p*(*v* | *I*). The generalized model offered in (Sotiropoulos, Seitz, and Seriès, 2011) expresses $\hat{v}$ as

(10) $$\hat{v} = -\begin{bmatrix} \sum I_x^{\ 2} + \frac{\sigma^2}{\sigma_p^{\ 2}} & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^{\ 2} + \frac{\sigma^2}{\sigma_p^{\ 2}} \end{bmatrix}^{-1} \begin{bmatrix} \sum I_x I_t - \frac{\sigma^2}{\sigma_p^{\ 2}} \mu_x \\ \sum I_y I_t - \frac{\sigma^2}{\sigma_p^{\ 2}} \mu_y \end{bmatrix},$$

where I_x, I_y, and I_t are the derivatives of *I* with respect to *x*, *y*, and *t*, and where sums are taken over pixels in the stimulus image. For our purposes, the details of (10) are less important than its schematic form:

(11) $$\hat{v} = \Phi(\mu, \frac{\sigma^2}{\sigma_p^{\ 2}}, I_x, I_y, I_t)\ .$$

(11) displays estimated velocity $\hat{v}$ as a function of several variables: the mean μ of the velocity prior; the ratio $\frac{\sigma^2}{\sigma_p^{\ 2}}$; and the spatial and temporal derivatives of image intensity *I*. Using the Bayesian model, one can explain a variety of motion illusions that are otherwise quite difficult to accommodate within a single unified framework. In the words of Born and Bradley (2005, p. 179): "[t]he model can explain a remarkable range of psychophysical observations." For example, low-contrast stimuli typically appear to move slower than high-contrast stimuli. The Bayesian model explains this as follows: lower contrast stimuli induce noisier local motion measurements, which lead to a relatively "spread out" likelihood function, which allows the slow motion prior *p*(*v)* to dominate computation of the posterior, which yields a lower speed estimate.

(10) answers numerous w-questions of the form: *How would the velocity estimate have been different had an intervention altered the velocity prior?* (10) specifies how $\hat{v}$ would have changed had an intervention altered the prior's mean μ (corresponding to a change in anticipated velocity) or variance (corresponding to increased or decreased uncertainty regarding anticipated velocity). It thereby depicts how the explanandum (estimated velocity) depends upon the

explanans (the velocity prior). To test (10), Sotiropoulos, Seitz, and Seriès (2011) experimentally manipulated the velocity prior. They exposed subjects to two-dimensional moving stimuli. When subjects repeatedly encountered fast-moving stimuli, the velocity prior shifted so as to favor speeds faster than zero. As a result, velocity estimates increased in accord with (10). The visual system interpreted the same stimulus as moving faster after the experimental manipulation. Quite plausibly, this experimental manipulation counts as an intervention on the velocity prior. So (10) yields approximately true test counterfactuals describing the relation between μ and $\hat{v}$.

I submit that we find the motion estimation model explanatorily powerful precisely because it yields so many approximately true test counterfactuals. The model explains velocity estimation by revealing how the velocity percept would change if an intervention altered a key antecedent mental state: the velocity prior.

To derive (10) from Bayes's Rule, we make numerous assumptions about the stimulus: that the perceived pattern moves with a single translational velocity; that points on the pattern do not change in apparent brightness as they move; that the image intensity function is smooth enough to allow a Taylor series approximation; that measurement noise is independent across spatial location; etc. The derivation also makes several restrictive assumptions based more upon mathematical convenience than psychological realism. For example, it assumes that the velocity prior is a Gaussian, even though evidence suggests that the true velocity prior is "heavier tailed" than a Gaussian (Stocker and Simoncelli, 2006). So (10) only prevails for certain stimuli, and it only prevails under restrictive assumptions (some false) about the perceptual system. A similar diagnosis applies to all other Bayesian models found in perceptual psychology.

From an interventionist perspective, these facts are not troubling. As long as a generalization generates approximately true test counterfactuals, it supports satisfying scientific explanations. That (10) meets this standard has been experimentally confirmed. Even though we derived (10) using restrictive assumptions (some false) about the perceptual system, and even though (10) fails against certain background conditions, it still achieves the goal that causal explanation of perception should achieve: it illuminates how the percept depends upon key causally relevant factors.

I have adduced actual experimental interventions on perceptual prior probabilities. The literature depicts additional experimental manipulations that plausibly intervene upon prior probabilities (Ernst, 2007), (Flanagan, Bittner, Johansson, 2008), (Knill, 2007) or prior likelihoods (Sato and Kording, 2014), (Seydell, Knill, Trommershäuser, 2010). No doubt further such experimental interventions will emerge as the field develops. Even when researchers have not directly confirmed test counterfactuals generated by some Bayesian perceptual model, the model's other predictive successes often provide strong evidence that those counterfactuals are at least approximately true. Overall, then, explanatory practice within Bayesian perceptual psychology conforms nicely to the interventionist template. Bayesian perceptual psychologists try to isolate (and often successfully isolate) explanatory generalizations that generate approximately true test counterfactuals.

§5.4 Non-deterministic Bayesian perceptual modeling

So far, I have emphasized Bayesian models that can be schematized using (8). These models are *deterministic*. However, many Bayesian perceptual models involve indeterministic elements that flout schema (8).

One source of indeterminacy is *noise*. Human neural activity, including perceptual processing, is very noisy. For that reason, the same stimulus can cause different percepts on different occasions. To model trial-by-trial variation in human perceptual response, Bayesian modelers frequently add indeterministic perturbations to the model. In this spirit, Stone (2011) offers a *deterministic* Bayesian model of shape perception based on shading cues, and he then corrupts the model with indeterministic Gaussian noise. The resulting non-deterministic model yields a formula for $O(c)$, the probability that one perceives shape c. Here $O(c)$ is an *objective probability* rather than a *subjective probability*.

Another possible source of indeterminacy is *multistable perception*. Certain ambiguous stimuli, such as the Necker cube, cause the perceptual system to "flip" between rival percepts. Bayesian perceptual psychologists sometimes model multistable perception in indeterministic terms. They replace expected cost minimization with an indeterministic strategy for transiting from the posterior $p(h \mid e)$ to the estimate $\hat{h}$. The usual idea is that the perceptual system *samples* an estimate from the posterior, where the frequency distribution of sampled estimates approximately matches the posterior. Moreno-Bote, Knill, and Pouget (2011) offer a model of this kind. The model concerns an ambiguous stimulus composed of two gratings that drift in opposite directions. Perceived depth ordering of the gratings spontaneously reverses. See Figure 2. By varying the gratings' wavelengths and speeds, we vary the frequency $freq(o)$ with which depth ordering o appears. The Bayesian model expresses $freq(o)$ as a function of three factors: a prior $p(o)$ over depth orderings; a prior likelihood $p(\Delta\lambda \mid o)$ that relates the depth ordering o to the difference in wavelength $\Delta\lambda$ between the gratings; and a prior likelihood $p(\Delta v \mid o)$ that relates the depth ordering to the difference in speed Δv between the gratings.

INSERT FIGURE 2 ABOUT HERE

How should interventionists handle non-deterministic Bayesian perceptual modeling? The interventionist schema (3) features a *deterministic* generalization that describes how interventions on certain explanantia variables would alter *Y*'s value. Taking (3) as our guide, non-deterministic models look unexplanatory. For example, Stone's (2011) model of shape perception does not depict how changes to the priors would *determinately* change the shape-estimate. Should interventionists conclude that Stone's model does not explain the perceptual state estimate?

To avoid that conclusion, interventionists might generalize schema (3) to allow non-deterministic explanations. They might allow that a non-deterministic generalization involving explanandum variable *Y* can explain *Y*'s value. Woodward and Hitchcock (2003b) mention this option but do not endorse it.

I favor a different line. I concede that non-deterministic Bayesian models do not explain perceptual estimates, but I insist that they explain *objective probabilities* or *frequencies* of perceptual estimates. Take Stone's (2011) model of shape perception. Because it is non-deterministic, the model yields no generalization of the form (3). I grant that the model does not explain the perceptual shape-estimate. Nevertheless, the model yields a determinate equation of the form

(12) $O(c) = \Phi(\mathit{prior}, \mathit{prior\ likelihood}, e)$.

It thereby depicts how interventions on priors would alter the objective probability of a given shape-estimate. Even if the model does not explain the shape-estimate itself, it explains the objective probability of a given shape-estimate. Similarly, the model offered by Moreno-Bote, Knill, and Pouget (2011) yields a determinate equation of the form:

$\mathit{freq}(o) = \Phi(\mathit{prior\ probability}, \mathit{prior\ likelihood}, e)$;

or, more specifically,

(13) $freq(o) = \Phi(p(o), p(\Delta\lambda \mid o), p(\Delta v \mid o), \Delta\lambda, \Delta v)$.

Equation (13) expresses the frequency *freq*(*o*) with which ordering *o* is perceived as a function of the priors and sensory input. (13) generates approximately true test counterfactuals describing how *freq*(*o*) would vary if an intervention were to alter the priors. It thereby explains *freq*(*o*), even if it does not explain the perceived ordering *o* itself.

I conclude that interventionism, although geared towards deterministic explanation, can handily accommodate indeterministic Bayesian perceptual modeling.

§6. Bayesian cognitive science

Inspired by the success of Bayesian perceptual psychology, cognitive scientists have offered Bayesian models for numerous additional mental processes. Researchers have applied Bayesian modeling to many areas: *motor control* (Wolpert, 2007); *language acquisition* (Chater and Manning, 2006); *high-level cognition* (Chater and Oaksford, 2008), (Griffiths, Kemp, and Tenenbaum, 2008), including *social cognition* (Baker and Tenenbaum, 2014), *intuitive physics* (Sanborn, Masinghka, and Griffiths, 2013), and *causal reasoning* (Gopnik, Glymour, Sobel, Schulz, and Kushnir, 2004); *human and nonhuman navigation* (Madl et al., 2014), (Petzschner and Glasauer, 2011); *mental disorders*, such as *schizophrenia* (Fletcher and Frith, 2009) and *autism* (Pellicano and Burr, 2012); and so on. The resulting theories are often far more rigorous and better-confirmed than alternative theories. A Bayesian model of a mental process generates detailed counterfactuals describing how the process would have proceeded differently had we intervened on the priors. One reason why Bayesian models often seem so much more fruitful than rival theories is that they generate so many detailed test counterfactuals.

Whether the test counterfactuals are *true* depends, naturally, upon the specific model. We have seen that test counterfactuals generated by Bayesian perceptual psychology are often at least approximately true. For a non-perceptual example, consider *Bayesian sensorimotor psychology*. This research program studies the mental processes that transform *intentions* into *motor commands*. For instance, if I form an intention to pick up a ball, then my motor system must transform my intention into motor commands that help execute the intention. Bayesian models postulate that the motor system transforms intentions into motor commands through unconscious inference and decision-making (Bays and Wolpert, 2007), (Rescorla, 2016), (Wolpert, 2007). The motor system uses Bayesian inference to estimate environmental state (including the subject's own bodily state) based on sensory feedback. The resulting environmental state estimates guide the motor system as it selects motor commands that promote the individual's intentions. These motor commands are selected through expected cost minimization. Bayesian sensorimotor models generate well-confirmed counterfactuals that describe how motor outcomes would have changed had an intervention altered the motor system's priors or the individual's intentions.[12] Thus, Bayesian sensorimotor psychology explains motor outcomes by revealing how those outcomes depend upon antecedent mental states.

Bayesian perceptual psychology and Bayesian sensorimotor psychology are the best-developed areas of Bayesian cognitive science. Other areas do not achieve such massive explanatory success. Nevertheless, all areas of Bayesian cognitive science *try* to articulate generalizations that generate true test counterfactuals. In many cases, the generalizations are approximately true. Thus, Bayesian cognitive science goes substantially beyond the minimally explanatory generalizations discussed in §4, such as Weber's law. Future work should bolster my

[12] See (Campbell, 2007) for discussion of what it is to intervene on an intention.

assessment by analyzing additional Bayesian case studies. Future work should also evaluate the extent to which non-Bayesian cognitive science improves upon minimally explanatory generalizations.

§7. Comparison with the nomological conception

How does my interventionist conception of psychological explanation compare with the nomological and mechanistic conceptions? I discuss the nomological conception in this section and the mechanistic conception in §8.

Interventionism and the DN model both emphasize explanatory generalizations. The difference concerns *which* generalizations count as explanatory. DN theorists prioritize the distinction between lawlike and accidental generalizations. Interventionists do not invoke the notion of law. Instead, they prioritize test counterfactuals. Weber's law is minimally explanatory, because it supports a restricted range of test counterfactuals. Generalizations such as (10) and (13) are more explanatory, because they support a wider range of test counterfactuals.

Interventionism accords much better than the DN model with actual cognitive science practice. Cognitive scientists do not treat a generalization as explanatory simply because it is a law. Instead, they pursue generalizations that generate approximately true test counterfactuals --- the more such counterfactuals, the better.

To illustrate, consider a famous puzzle: *the moon illusion*. We can describe the moon illusion through the following law:

(14) The moon looks larger on the horizon than at its zenith, *ceteris paribus*.

Fodor frequently cites (14) as a paradigmatic psychological law (Fodor, 1991a), (Fodor, 1991b, p. 280), (Fodor and Lepore, 1992, pp. 151-152). And (14) does indeed seem as a close to a psychological law as one could hope to find. In particular, it supports counterfactuals along the following lines:

(15) If the moon had been at the horizon rather than its zenith, then it would have looked larger.

Nevertheless, (14) does not seem explanatory. Intuitively, (14) does not provide the slightest indication which features of the distal or proximal stimulus are responsible for the change in apparent lunar size. After all, lunar elevation *in itself* hardly seems like a plausible causal influence upon apparent lunar size. No vision scientist would take (14) seriously as even the start of an explanation for why the moon appears larger on some occasions than others. (14) merely states a phenomenon that we wish to explain.[13]

From an interventionist viewpoint, counterfactuals such as (15) are irrelevant to causal explanation. What matters are *test* counterfactuals. (14) does not support test counterfactuals that are even approximately true. As lunar elevation changes during the nighttime, a crucial additional variable typically changes: distance cues afforded by surrounding terrain. An intervention on lunar elevation should hold ground terrain distance cues fixed. Rock and Kaufman (1962) performed such an intervention. They used optical tricks to reverse ground terrain cues, so that the moon appeared in zenith position with horizon ground terrain surround and in the horizon position with an empty sky surround. This experimental intervention reversed the moon illusion: the horizon moon looked smaller than the zenith moon. Hence, (14) does not accurately specify how perceived lunar size would change if an *intervention* altered lunar elevation. Rather than isolate a causally relevant variable, (14) highlights an irrelevant concomitant (lunar elevation).

[13] Fodor seems to recognize that (14) looks unexplanatory, because he usually cites it as an explanandum rather than an explanans. I doubt that Fodor can consistently classify (14) as unexplanatory, since it counts as a law according to the traditional criteria of lawhood.

In this respect, (14) compares unfavorably with even a minimally explanatory generalization such as Weber's law. At least Weber's law isolates a crucial variable (stimulus magnitude) that causally influences the explanandum (JND). For a more pointed comparison, Rock and Kaufman (1962) show that a suitable generalization

(16) $S = f(G)$,

holds against normal background conditions, where G is a binary variable that reflects whether ground terrain surround is present and S is a binary variable that reflects whether perceived lunar size exceeds an appropriately chosen reference standard. (16), unlike (14), supports true test counterfactuals. (16), unlike (14), specifies how to manipulate the percept by intervening on a key explanans variable. So (16) is explanatory, if minimally so. My verdict accords with the widespread scientific consensus that Rock and Kaufman achieved considerable explanatory progress by isolating a key causal influence on apparent lunar size.

Obviously, we would like a deeper explanation for the moon illusion than (16). Unfortunately, the psychological processes that generate the moon illusion are highly controversial (Hershenson, 1989), (Kaufman and Kaufman, 2000). Thus, (16) is only the beginning of a fully satisfying explanation. But it *is* a beginning, whereas (14) is not. This contrast in explanatory status supports the interventionist conception over the nomological conception.

Over the past few decades, philosophers have extensively debated whether scientific psychology can supply explanatory generalizations that deserve the honorific "law" (Davidson, 1980), (Fodor, 1987), (Dennett, 1993), (Schiffer, 1991). One might therefore ask which, if any, of the explanatory generalizations (7), (10), (12), (13), and (16) count as laws. I can remain neutral on this question, because my positive account does not employ the notion *law*.

Depending on how one demarcates the laws, some of the generalizations may count as laws. Depending on how one demarcates the laws, some explanations offered within cognitive science may conform to the nomological conception. My primary complaint about the nomological conception is not that it fails to recognize certain psychological explanations as explanations. My primary complaint is that the nomological conception (as traditionally developed) does not elucidate *why* these explanations are explanatory. From an interventionist viewpoint, it does not much matter whether a psychological generalization counts as a law. What matters is how many (if any) approximately true test counterfactuals a psychological generalization generates.

In summary, interventionism provides what the traditional literature on laws does not: a principled account that illuminates why generalizations such as (7), (10), (12), (13), and (16) are explanatory while generalizations such as (14) are unexplanatory.

§8. Comparison with the mechanistic conception

Bechtel (2008) illustrates the mechanistic viewpoint with various case studies drawn from cognitive science. At least some of Bechtel's case studies yield approximately true test counterfactuals, so these count as genuine explanations from an interventionist viewpoint. *Some* psychological explanations therefore conform to the mechanistic conception. Nevertheless, the mechanistic conception does not fit well with numerous other psychological explanations, ranging from minimally explanatory generalizations such as Weber's law or (16) to the far deeper Bayesian explanations canvassed in §§5-6.

Consider (16). This is a paradigmatic "phenomenal" description of the sort that Craver and other mechanists disapprovingly compare with genuine explanations. (16) has a purely input-output character: it says that stimuli with certain properties reliably cause percepts with certain properties. It does not even begin to suggest computational or neural mechanisms that mediate between proximal sensory input and the percept. It tells us nothing about the mechanisms that generate perceived lunar size, except of course that those mechanisms are sensitive to presence or absence of ground terrain surround. From a mechanistic viewpoint, it is quite mysterious why (16) constitutes such a marked advance over (14) in our understanding of the moon illusion. From an interventionist viewpoint, the advance is apparent: (16), unlike (14), yields approximately true test counterfactuals.

Or consider the Bayesian perceptual models surveyed in §5. The models are highly non-mechanistic. They provide no clue regarding how the perceptual system encodes prior probabilities or prior likelihoods. Nor do they decompose the perceptual system into neural or computational components. Nor do they identify underlying neural processes or mental computations through which the perceptual system executes Bayesian inference. The non-mechanistic character of Bayesian modeling is widely acknowledged within cognitive science (Griffiths, Kemp, and Tenenbaum, 2008), (Jones and Love, 2011), (Knill and Richards, 1996) and philosophy (Colombo and Hartmann, 2017), (Herschbach and Bechtel, 2011). Some authors invoke it to *critique* the Bayesian paradigm (Jones and Love, 2011), (Herschbach and Bechtel, 2011). I think we should instead regard it as evidence against the mechanistic conception of psychological explanation. Many Bayesian models, including the motion estimation model, are explanatory.

When confronted with putatively non-mechanistic psychological explanations, mechanists sometimes reply that the putative counterexamples are *mechanism sketches* (Piccinini and Craver, 2011). They say that a theory counts as explanatory if it sketches a mechanism, with full mechanistic details to be disclosed at a future date. This reply seems rather forced when applied to minimally explanatory generalizations. For example, (16) does not meet Craver's (2006, p. 360) criteria for a mechanism sketch, because it provides no hint how the relevant perceptual mechanism decomposes into computational or neural parts. The appeal to mechanism sketches seems more compelling when applied to Bayesian models. A Bayesian model delineates abstract computational features of neural processing, so one might plausibly contend that it sketches a neural mechanism. Colombo and Hartmann (2017) argue along these lines. They say that Bayesian modeling *constrains* neural mechanisms, thereby providing a first step towards full-blown mechanistic explanation.

In response, I deny that constraints upon neural mechanism suffice for psychological explanation. To illustrate, consider motion perception. Many computations that underlie primate motion perception occur in the middle temporal brain region (V5). We know a huge amount about V5, its internal structure, its contribution to motion perception, and its place in neural architecture (Born and Bradley, 2005), (Zeki, 2015). Notably, though, much of this knowledge does not illuminate why the perceptual system estimates a particular velocity. For example, we have detailed knowledge about the neural pathways into V5 from V1, V2, V3, and other brain areas. These are mechanistic details. They concern how certain regions of the brain are organized. These particular mechanistic details do not in themselves illuminate why the perceptual system estimates one velocity rather than another. No practicing scientist would say that we have explained perceived velocity, even in a superficial way, simply by specifying the neural pathways into V5. A mechanism sketch that merely specifies pathways into V5 does not explain why the perceptual system estimates one velocity rather than another.

Some mechanism sketches explain the explanandum. Others do not. So the mere fact that a Bayesian model sketches a mechanism does not entail that it explains the explanandum. We cannot elucidate the explanatory value of Bayesian models simply by noting that they supply mechanism sketches or that they constrain underlying mechanisms. We need a systematic account that demarcates the explanatory mechanism sketches from the unexplanatory mechanism sketches. Interventionism provides the needed demarcation: the explanatory mechanism sketches are those that generate suitable test counterfactuals. For example, a specification of the neural pathways into V5 may predict that certain kinds of brain damage would impair motion perception, but it does not pinpoint any intervention that would produce a specific velocity estimate. Therefore, it does not explain why the perceptual system estimates one velocity rather than another.[14] The Bayesian motion estimation model explains why the perceptual system estimates one velocity rather than another, because it describes how intervening on the priors would produce specific velocity estimates. If we want to evaluate whether a model explains an explanandum, the crucial question is which test counterfactuals the model generates, not whether the model sketches a mechanism.

I conclude that interventionism outperforms the mechanistic conception in its handling of the numerous non-mechanistic explanations offered within cognitive science.

Interventionism also provides useful guidance regarding how one might improve upon non-mechanistic psychological explanations. Everyone agrees that it would be desirable to clarify the neural mechanisms that underlie psychological activity. In particular, we would like to supplement existing Bayesian models by specifying how neural states encode priors and how neural activity approximately implements Bayesian inference. However, not all mechanistic details improve explanation. If we learn that a certain pattern of neural firing occurs in the causal

[14] Some well-confirmed test counterfactuals relate V5 to perceived velocity. Researchers have confirmed counterfactuals of the form: *If we microstimulate certain cells in V5, then certain changes in the velocity percept will occur* (Zeki, 2015). Accordingly, I count some mechanistic details about V5 as explanatory. My point in the main text is that some notable mechanistic details about V5 are *not* explanatory.

chain from proximal stimulation to percept, this mechanistic detail does not necessarily improve our understanding of why the percept occurs. The specific pattern of neural firing may be explanatorily important, or it may be more analogous to the specific weight 8356 kg on the collapsing bridge or to specific printing press gear configurations through which the money supply changes. We would like a principled account that clarifies which details about neural mechanisms are explanatory. Which ways of "filling in the mechanism sketch" with neural details improve psychological explanation? Interventionism provides a systematic, satisfying answer: neural details improve psychological explanation when they generate new approximately true test counterfactuals. For example, a theory of the neural substrate for Bayesian perceptual inference will constitute an explanatory advance if it clarifies how the percept would change if we intervened on various neural states, or if it isolates neural conditions under which an idealized Bayesian perceptual model breaks down. The scientific literature already offers some speculative theories that attempt this sort of advance (Pouget, Beck, Ma and Latham, 2013).

§9. Advantages of the interventionist conception

An interventionist approach to psychological explanation offers notable advantages over the nomological and mechanistic conceptions. Like the nomological conception, interventionism emphasizes the pivotal role that generalizations play within psychological explanation. Interventionism improves upon the nomological conception by clarifying what makes a psychological generalization explanatory. Like the mechanistic conception, interventionism places causation at center stage. Interventionism improves upon the mechanistic conception by elucidating how one can expose explanatorily important causal structure without limning anything like a mechanism that produces the explanandum. Hence, interventionism preserves virtues of the nomological and mechanistic conceptions while improving upon both.

Acknowledgments

I presented excerpts from this material at a conference on Bayesian Theories of Perception and Epistemology at Cornell University, July 2015; during a symposium at the Philosophy of Science Association Biennial Meeting in Atlanta, November 2016; and during a symposium at the Society for Philosophy and Psychology Annual Meeting, Baltimore, July 2017. I am grateful to all participants and audience members for helpful feedback, especially David Chalmers, David Danks, Steven Gross, Gualtiero Piccinini, Susanna Siegel, and Scott Sturgeon. Thanks also to Nicholas Shea and to three anonymous referees for this journal for comments that significantly improved the paper. My research was supported by a fellowship from the National Endowment for the Humanities. Any views, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect those of the National Endowment for the Humanities.

Works Cited

Adams, W., Graf, E., and Ernst, M. 2004. "Experience Can Change the "Light-From-Above' Prior." *Nature Neuroscience* 7: 1057-1058.

Alais, D., and Burr, D. 2004. "The Ventriloquism Effect Results from Near-Optimal Bimodal Integration." *Current Biology* 14: 257-262.

Antony, L. 1995. "Law and Order in Psychology." *Philosophical Perspectives* 9: 429-446.

Aydede, M. 2000. "Computation and Intentional Psychology." *Dialogue* 39: 365-379.

Baker, C., and Tenenbaum, J. 2014. "Modeling Human Plan Recognition Using Bayesian Theory of Mind." In *Plan, Activity, and Intent Recognition: Theory and Practice,* eds. G. Sukthankar, R. P. Goldman, C. Geib, D. Pynadath, D., and H. Bui. Waltham: Morgan Kaufmann.

Bays, P., and Wolpert, D. 2007. "Computational Principles of Sensorimotor Control that Minimize Uncertainty and Variability." *Journal of Physiology* 578: 387-396.

Bechtel, W. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Lawrence Erlbaum Associates.

Bechtel, W., and Wright, C. 2009. "What is Psychological Explanation?" In *Routledge Companion to the Philosophy of Psychology,* eds. J. Symons and P. Calvo. New York:

Routledge.

Born, R., and Bradley, D. 2005. "Structure and Function of Visual Area MT." *Annual Review of Neuroscience* 28: 157-189.

Burge, T. 2010. *Origins of Objectivity*. Oxford: Oxford University Press.

Campbell, J. 2007. "An Interventionist Approach to Causation in Psychology." In *Causal Learning: Psychology, Philosophy, and Computation*, eds. A. Gopnik and L. Schulz. Oxford: Oxford University Press.

Chater, N., and Manning, C. 2006. "Probabilistic Models of Language Processing and Acquisition." *Trends in Cognitive Science* 10: 335-344.

Chater, N., and Oaksford, M., eds. 2008. *The Probabilistic Mind*. Oxford: Oxford University Press.

Colombo, M., and Hartmann, S. 2017. "Bayesian Cognitive Science: Unification and Explanation." *The British Journal for the Philosophy of Science* 68: 451-484.

Craver, C. 2006. "When Mechanistic Models Explain." *Synthese* 153: 355-376.

---. 2014. "The Ontic Account of Scientific Explanation." In *Explanation in the Special Sciences: The Case of Biology and History*, eds. M. Kaiser, O. Scholz, D. Plenge, and A. Hütteman. Dordrecht: Springer.

Cummins, R. 2000. "'How Does It Work?' versus 'What Are the Laws?': Two Conceptions of Psychological Explanation." In *Explanation and Cognition*, eds. F. Keil and R. Wilson. Cambridge: MIT Press.

Darley, J. and Latané, B. 1968. "Bystander Intervention in Emergencies: Diffusion of Responsibility." *Journal of Personality and Social Psychology* 8: 377-383.

Davidson, D. 1980. *Essays on Actions and Events*. Oxford: Clarendon Press.

Dennett, D. 1993. "Back from the Drawing Board." In *Dennett and his Critics*, ed. B. Dahlbom. Malden: Blackwell.

Earman, J., Roberts, J., and Smith, S. 2002. "*Ceteris Paribus* Lost." *Erkenntnis* 57: 281-302.

Ernst, M. 2007. "Learning to Integrate Arbitrary Signals from Vision and Touch." *Journal of Vision* 7: 1-14.

Feldman, J. 2015. "Bayesian Models of Perceptual Organization." In *The Oxford Handbook of Perceptual Organization*, ed. J. Wagemans. Oxford: Oxford University Press.

Flanagan, J., Bittner, J., Johansson, R. 2008. "Experience Can Change Distinct Size-Weight Priors Engaged in Lifting Objects and Judging Their Weights." *Current Biology* 22: 1742-1747.

Fletcher, P., and Frith, C. 2009. "Perceiving is Believing: A Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia." *Nature Reviews Neuroscience* 10: 48-58.

Fodor, J. 1981. *Representations*. Cambridge: MIT Press.

---. 1983. *The Modularity of Mind*. Cambridge: MIT Press.

---. 1987. *Psychosemantics*. Cambridge: MIT Press.

---. 1991a. "Replies." In *Meaning in Mind*, eds. B. Loewer and G. Rey. Cambridge: Blackwell.

---. 1991b. "You Can Fool Some of the People All of the Time, Everything Else Being Equal: Hedged Laws and Psychological Explanation." *Mind* 100: 19-34.

---. 1994. *The Elm and the Expert*. Cambridge: MIT Press.

Fodor, J., and Lepore, E. 1992. *Holism: A Shopper's Guide*. Cambridge: Blackwell.

Franklin-Hall, L. 2016. "High-level Explanations and the Interventionist's 'Variables Problem.'" *The British Journal for the Philosophy of Science* 67: 553-577.

Garcia, J., and Koelling, R. 1966. "The Relation of Cue to Consequence in Avoidance

Learning." *Psychonomic Science* 4: 123-124.

Gauker, C. 2005. "The Belief-Desire Law." *Facta Philosophica* 7: 121-144.

Gopnik, A., Glymour, G., Sobel, D., Schulz, L., and Kushnir, T. 2004. "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets." *Psychological Review* 111: 3-32.

Griffiths, T., Kemp, C., and Tenenbaum, J. 2008. "Bayesian Models of Cognition." In *The Cambridge Handbook of Computational Cognitive Modeling,* ed. R. Sun. Cambridge: Cambridge University Press.

Helmholtz, H. von. 1867. *Handbuch der Physiologischen Optik*. Leipzig: Voss.

Hempel, C. 1965. *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science*. New York: Free Press.

Herschbach, M., and Bechtel, W. 2011. "Relating Bayes to Cognitive Mechanisms." *Behavioral and Brain Sciences* 34: 202-203.

Hershenson, M. 1989. *The Moon Illusion*. Hillsdale: Lawrence Erlbaum Associates.

Horgan, T., and Tienson, J. 1990. "Soft Laws." *Midwest Studies in Philosophy* 15: 256-279.

Jones, M., and Love, B. 2011. "Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contribution of Bayesian Models of Cognition." *Behavioral and Brain Sciences* 34: 169-188.

Kaufman, L., and Kaufman, J. 2000. "Explaining the Moon Illusion." *Proceedings of the National Academy of Sciences* 97: 500-505.

Kitcher, P. 1989. "Explanatory Unification and the Causal Structure of the World." In *Scientific Explanation*, eds. P. Kitcher and W. Salmon. Minneapolis: University of Minnesota Press.

Knill, D. 2007. "Learning Bayesian Priors for Depth Perception." *Journal of Vision* 7: 1-20.

Knill, D. and Richards, W., eds. 1996. *Perception as Bayesian Inference*. Cambridge: Cambridge University Press.

Lange, M. 2016. *Because Without Cause*. Oxford: Oxford University Press.

MacDonald, J., and McGurk, H. 1978. "Visual Influences on Speech Perception Processes." *Perception and Psychophysics* 24: 253-257.

Madigan, S. 1969 "Intraserial Repetition and Coding Processes in Free Recall." *Journal of Verbal Learning and Verbal Behavior* 8: 828-835.

Madl, T., Franklin, S., Chen, K., Montaldi, D., and Trappl, R. 2014. "Bayesian Integration of Information in Hippocampal Place Cells." *PloS One* 9: e89762.

Mankiw, G. 1997. *Macroeconomics,* 3rd ed. New York: Worth Publishers.

Moreno-Bote, R., Knill, D., and Pouget, A. 2011. "Bayesian Sampling in Visual Perception." *Proceedings of National Academy of Sciences* 108: 12491-6.

Palmer, S. 1999. *Vision Science*. Cambridge: MIT Press.

Pellicano, E., and Burr, D. 2012. "When the World Becomes Too Real." *Trends in Cognitive Science* 16: 504-510.

Petzschner, F., and Glasauer, S. 2011. "Iterative Bayesian Estimation as an Explanation for Range and Regression Effects: A Study on Human Path Integration." *Journal of Neuroscience* 31: 17220-17229.

Pietroski, P., and Rey, G. 1995. "When Other Things Aren't Equal: Saving Ceteris Paribus Laws from Vacuity." *The British Journal for the Philosophy of Science* 46: 81-110.

Piccinini, G., and Craver, C. 2011. "Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches." *Synthese* 183: 283-311.

Pouget, A., Beck, J., Ma., W. J., and Latham, P. 2013. "Probabilistic Brains: Knowns and

Unknowns." *Nature Neuroscience* 16: 1170–1178.

Putnam, H. 1975. *Mind, Language, and Reality: Philosophical Papers, vol. 2*. Cambridge: Cambridge University Press.

Rescorla, M. 2014. 2014. "The Causal Relevance of Content to Computation." *Philosophy and Phenomenological Research* 88: 173-208.

---. 2015. "Bayesian Perceptual Psychology." In *The Oxford Handbook of the Philosophy of Perception,* ed. M. Matthen. Oxford: Oxford University Press.

---. 2016. "Bayesian Sensorimotor Psychology." *Mind and Language* 31: 3-36.

Rock, I., and Kaufman, L. 1962. "The Moon Illusion, II." *Science* 136: 1023-1031.

Saatsi, J., and Pexton, M. 2013. "Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Noncausal Explanations." Philosophy of Science 80: 613-624.

Salmon, W. 1971. "Statistical Explanation." In *Statistical Explanation and Statistical Relevance,* ed. W. Salmon. Pittsburgh: University of Pittsburgh Press.

---. 1989. "Four Decades of Scientific Explanation." In *Scientific Explanations: Minnesota Studies in Philosophy of Science XIII,* eds. P. Kitcher and W. Salmon. Minneapolis: University of Minnesota Press.

Sanborn, A., Masinghka, J., and Griffiths, T. 2013. "Reconciling Intuitive Physics and Newtonian Mechanics for Colliding Objects." *Psychological Review* 120: 411-437.

Sato, Y., and Kording, K. 2014. "How Much to Trust the Senses: Likelihood Learning." *Journal of Vision* 14: 1-13.

Schiffer, S. 1991. "Ceteris Paribus Laws." *Mind* 100: 1-17.

Schneider, S. 2005. "Direct Reference, Psychological Explanation, and Frege Cases." *Mind and Language* 20: 423-447.

Seydell, A., Knill, D., and Trommershäuser, J. 2010. "Adapting Internal Statistical Models for Interpreting Visual Cues to Depth." *Journal of Vision* 10: 1-27.

Sotiropoulos, G., Seitz, A., Seriès, P. 2011. "Changing Expectations about Speed Alters Perceived Motion Direction." *Current Biology* 21: R883-R884.

Stinson, C. 2016. "Mechanisms in Psychology: Ripping Natures at its Seams." *Synthese* 193: 1585-1614.

Stocker, A., and Simoncelli, E. 2006. "Noise Characteristics and Prior Expectations in Human Visual Speed Perception." *Nature Neuroscience* 4: 578-585.

Stone, J. 2011. "Footprints Sticking Out of the Sand, Part 2: Children's Bayesian Priors for Shape and Lighting Direction." *Perception* 40: 175-190.

Stroop, J. 1935. "Studies of Interference in Serial Verbal Reactions." *Journal of Experimental Psychology* 18: 643-662.

Strevens, M. 2008. *Depth*. Cambridge: Harvard University Press.

Weiskopf, D. 2011. "Models and Mechanisms in Psychological Explanation." *Synthese* 181: 313-338.

Weiss, Y., Simoncelli, E., and Adelson, E. 2002. "Motion Illusions as Optimal Percepts." *Nature Neuroscience* 5: 598-604.

Wolpert, D. 2007. "Probabilistic Models in Human Sensorimotor Control." *Human Movement Science* 26: 511-524.

Woodward, J. 2003. *Making Things Happen*. Oxford: Oxford University Press.

---. 2008a. "Mental Causation and Neural Mechanisms." In *Being Reduced,* eds. J. Hohwy and J. Kallestrup. Oxford: Oxford University Press.

---. 2008b. "Cause and Explanation in Psychiatry: An Interventionist Perspective." In

Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology, eds. K. Kendler and J. Parnas. Baltimore: Johns Hopkins Press.

---. Forthcoming. "Explanation in Neurobiology: An Interventionist Perspective." In *Integrating Psychology and Neuroscience: Prospects and Problems*, ed. D. Kaplan. Oxford: Oxford University Press.

Woodward, J., and Hitchcock, C. 2003a. "Explanatory Generalizations, Part I: A Counterfactual Account." *Nous* 37: 1-24.

---. 2003b. "Explanatory Generalizations, Part II: Plumbing Explanatory Depth." *Nous* 37: 181-199.

Wright, C., and Bechtel, W. 2007. "Mechanisms and Psychological Explanation." In *Philosophy of Psychology and Cognitive Science*, ed. P. Thagard. Amsterdam: Elsevier.

Zeki, S. 2015. "Area V5 --- A Microcosm of the Visual Brain." *Frontiers in Integrative Neuroscience* 9: Article 21.

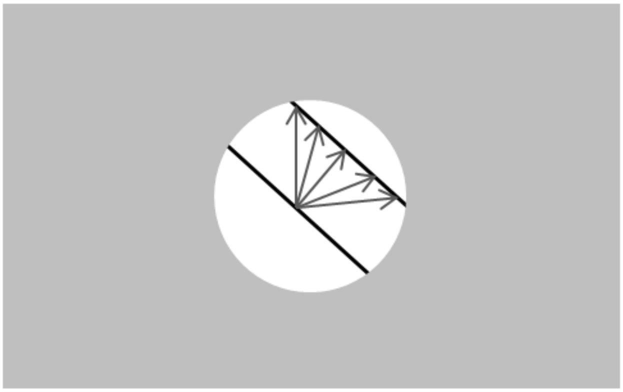

Figure 1. Motion of a straight edge, as observed through a small aperture. The observation is consistent with infinitely many velocity vectors, as illustrated by the five grey vectors.

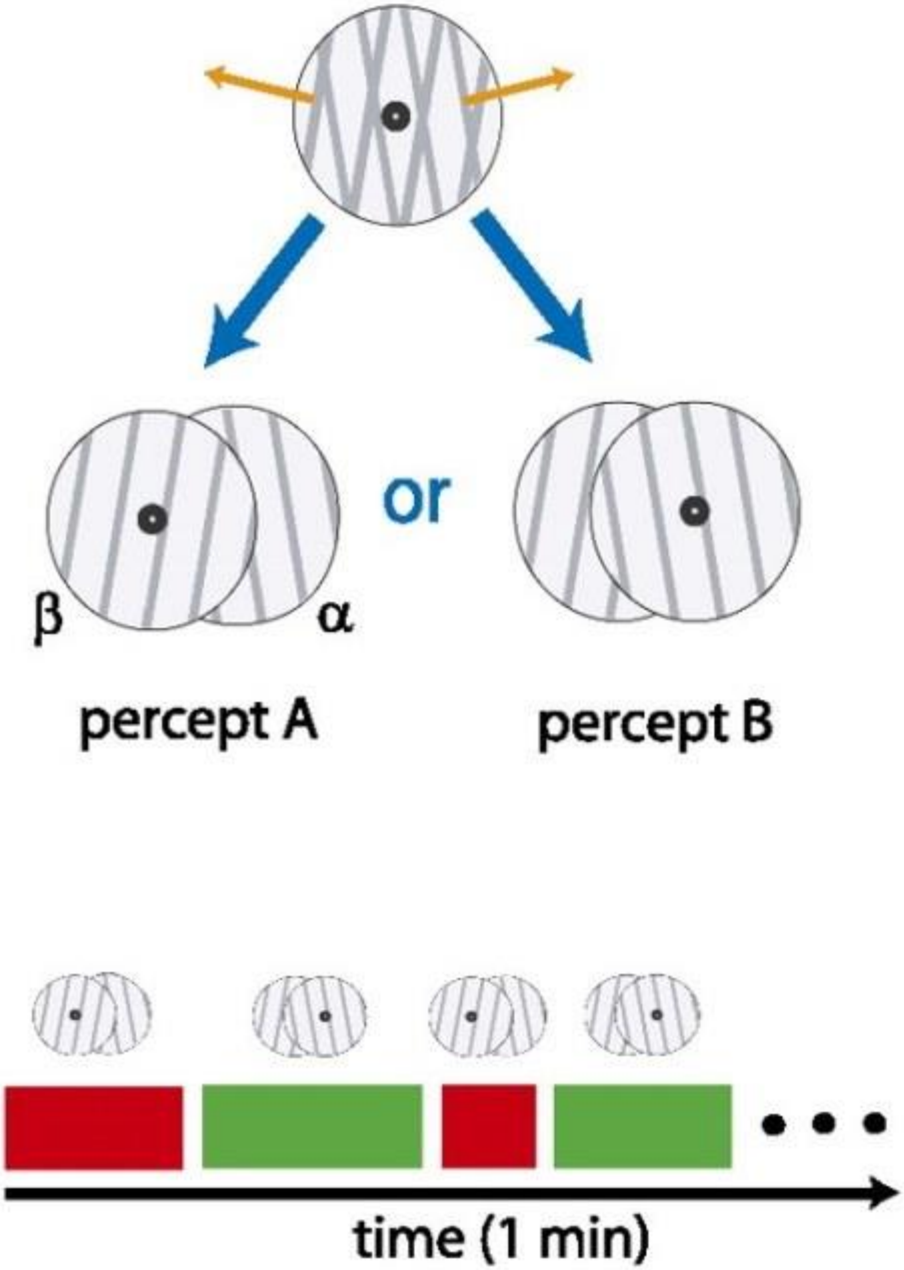

Figure 2. The ambiguous stimulus is shown at the top. Two percepts A and B are possible, corresponding to two depth orderings for superimposed gratings moving in opposite directions. The percept flips between A and B. *freq*(*o*) is the frequency with which ordering *o* is perceived. In this figure, both gratings have the same wavelength, i.e. the two gratings have the same spacing between lines. *freq*(*o*) changes when there is a disparity in wavelength Δλ, i.e. when lines are spaced differently in one grating than in the other. *freq*(*o*) also changes in response to changes in the relative speed Δ*v* with which the gratings move. Rpt. from (Moreno-Bote, Knill, and Pouget, 2011) with permission from Proceedings of the National Academy of Sciences.