Information and Explanation: An Inconsistent Triad and Solution

Mark Povich

Forthcoming in *European Journal for Philosophy of Science*

Abstract: An important strand in philosophy of science takes scientific explanation to consist in the conveyance of some kind of information (e.g., Lewis 1986; Railton 1981). Here I argue that this idea is also implicit in some core arguments of mechanists, some of whom (e.g., Craver 2014) are proponents of an ontic conception of explanation that might be thought inconsistent with it (Piccinini and Craver 2011; Zednik 2015). However, informational accounts seem to conflict with some lay and scientific commonsense judgments and a central goal of the theory of explanation, because information is relative to the background knowledge of agents (Dretske 1981). Sometimes we make lay judgments about whether a model is an explanation *simpliciter*, not just an explanation relative to some particular agent. And as philosophers of explanation, we would like a philosophical account to tell us when a model is an explanation *simpliciter*, not just when a model is an explanation relative to some particular agent. Thus, even if one's account of explanation is not concerned with explanation qua communicative or speech act, the account's reliance on the concept of information generates a *prima facie* conflict between the claims that 1) explanation is the conveyance of information, 2) information is relative to the background knowledge of an agent, and 3) some models are explanations not relative to the background knowledge of any particular agent. I sketch a solution to this puzzle by distinguishing informationally what I call "explanation *simpliciter*" from what I call "explanation-to," relativizing the latter to an individual's background knowledge and the former to what I call "total scientific background knowledge".

1. Introduction

Many venerable philosophical accounts of explanation have identified explaining with conveying information of some kind (e.g., Jackson and Pettit 1990; Lewis 1986; Railton 1981; Skow

2014)[1]. I will call these informational accounts of explanation. Usually informational accounts are causal (Ibid.) in the sense that explaining is said to be providing information about the explanandum phenomenon's causal history, but they need not be. There have been innumerable critiques of such causal accounts (e.g., Batterman and Rice 2014; Sober 1983; for responses see Povich 2018a and Skow 2014), but none, as far as I know, targets the apparent counterintuitive consequences of their reliance on information.

Informational accounts seem to conflict with some commonsense lay and scientific judgments and a central goal of the theory of explanation, because information is relative[2] to the background knowledge of agents (Dretske 1981). For example, it seems that sometimes a model is an explanation *simpliciter*, not just an explanation relative to some particular agent. We often speak of some thing or some text as being or providing an explanation without specifying nor intending implicitly to specify any particular audience; indeed, we often explicitly intend not to specify any particular audience. For example, take a claim like, "The Hodgkin-Huxley model explains the action potential." Here, and in similar examples, specification of any particular audience is often counter to the claim we intend to make. We precisely do not mean, "The Hodgkin-Huxley model explains the action potential to Lopez" or "The Hodgkin-Huxley model explains the action potential to this group of undergraduates". Obviously "explains" here should not be interpreted in the communicative sense. The Hodgkin-Huxley model cannot communicate with or to anyone. "Explains" as it occurs here is usually intended to mean something like "provides information about features explanatorily relevant to". The Hodgkin-Huxley

[1] I am not here concerned to show that *all* accounts of explanation rely on this assumption. Certainly, all accounts of explanation rely on representation, except an extreme ontic conception according to which explanations are the causal-mechanical structures represented, not the representations themselves. Proponents of the ontic conception such as Craver have backed away from this claim (Craver 2014; see below). If an informational account of representation is right, then my arguments may arguably apply to all accounts of explanation.

[2] An obvious way out of the puzzle is to deny the relativity of information. At least one account of information denies this relativity, but that account is controversial (Cohen and Meskin 2006; for objections, see Demir 2008 and Scarantino 2008; for a reply, see Meskin and Cohen 2008). For the purposes of this paper, I will assume that information is relative (see also fn. 16).

model provides information about features explanatorily relevant to the action potential. Provides such information to whom? In this and similar cases, we often intend no audience. One desideratum of an account of explanation should be to capture this "objectivity" of explanation.[3] We would like a philosophical account of explanation to tell us *when* a model such as the Hodgkin-Huxley model is an explanation *simpliciter*, not just when it is an explanation relative to some particular agent. However, relying on information seems only to give us the latter. Thus, there is a *prima facie* conflict between the claims that 1) explanation is the conveyance of information, 2) information is relative to the background knowledge of an agent, and 3) some models are explanations not relative to the background knowledge of any particular agent.[4] I sketch a solution to this puzzle by distinguishing scientific explanation *simpliciter* from what I call "explanation-to" and relativizing information in scientific explanations *simpliciter* to what I call "total scientific background knowledge" (TSBK).

To be clear, non-agent-relativity is the relevant sense in which some explanations are objective – the sense that generates the problem with which I am concerned. The Hodgkin-Huxley model is an objective explanation in the sense that it provides information – to no one in particular, that is, relative to no particular agent's background knowledge – about features explanatorily relevant to the action potential. I do not intend any stronger sense of objectivity than this; in particular, by "some explanations are objective," I certainly do not mean what some proponents of the ontic conception of explanation have meant, that some explanations are not representations, but the things represented.

In Section 2, I sketch a typical informational account of explanation, so that we have a clear example of such an account and provide textual evidence that such accounts have actually been held by prominent philosophers.

---

[3] Of course, pragmatists such as van Fraassen may disagree. I will not try here to convince them otherwise.

[4] To make this logically explicit, let "Ex" = "x is an explanation, "Cx" = "x conveys information," and "Rx" = "x is relative to the background knowledge of an agent". Then the triad can be transcribed as: $\forall x(Ex \rightarrow Cx)$, $\forall x(Cx \rightarrow Rx)$, and $\exists x(Ex \& \sim Rx)$. These are logically inconsistent. There may be more perspicuous transcriptions in higher-order logic, but this adequately gets the inconsistency across.

In Section 3, I show how an informational account of mechanistic explanation is implicit in some prominent mechanists' arguments. Specifically, an informational account is implied by Piccinini and Craver's (2011) argument that functional analyses are mechanism sketches and at least implies the crucial premise in Zednik's (2011; 2015) argument that dynamical systems models can be mechanistic explanations. I do not wish to saddle all mechanists with an informational account of explanation, but one is commonly implicit (or explicit) in some of their responses to putative counterexamples.

In Section 4, I explain why information is relative to background knowledge and illustrate this with Dretske's (1981, 78) example of the shell game. I can then spell out more clearly why informational accounts of explanation have the counterintuitive consequences they do.

In Section 5, I argue that informational accounts can avoid the counterintuitive consequences by distinguishing scientific explanation *simpliciter* from explanation-to (so called because it is an explanation *to* a particular agent). Information in scientific explanation *simpliciter* is relativized to what I call "total scientific background knowledge" (TSBK), the current total store of propositions that are known by the scientific community[5]. Information in explanation-to is relativized to the background knowledge of the particular agent to whom the explanation is given.

2. An informational account of explanation

In this section, I sketch a typical informational account of explanation (inspired by Lewis 1986), so that we have something concrete with which to work.

> **The Informational Account of Explanation (IAE)**: A model is an explanation (or is explanatory) when and only when it provides information about the explanandum phenomenon's causal history.

To be clear, I do not intend to defend **IAE** as an account of explanation or causal explanation. **IAE** is

[5] Scientific anti-realists can replace "knowledge" with their preferred substitute. Ditto for "veridical information" in Section 2.

merely an instance of the kind of informational account with which this paper is concerned. **IAE** a causal instance of an informational account since the kind of information that is deemed explanatory is information about causes. Other instances are certainly available, all of which are afflicted by the problem with which my paper is concerned. The problem afflicts any account of a form such as, "A model is an explanation (or is explanatory) when and only when it provides information about X," a typical instance of which results when "X" is "the explanandum phenomenon's causal history". Another instance of which would be, "A model is an explanation (or is explanatory) when and only when it provides information about the explanandum phenomenon's mechanism". But causal and mechanistic accounts of explanation are but two instances. An instance that is neither causal nor mechanistic is, "A model is an explanation (or is explanatory) when and only when it provides information about the explanandum phenomenon's grounds" (e.g., Skow 2016; grounds are ontologically distinguished from causes and mechanisms). One might even give a DN-inspired instance such as, "A model is an explanation (or is explanatory) when and only when it provides information about the laws governing the explanandum phenomenon". Here, I use the causal version as a typical instance; I return to the mechanistic version in the next section. All arguments that follow apply, *mutatis mutandis*, to all informational accounts of explanation, regardless of what is substituted for "X".[6]

To provide information, in the intended sense, about an explanandum phenomenon's causal history (or mechanism or grounds or whatever) is to reduce uncertainty about its causal history or to reduce the space of possible causal histories responsible for it (Dretske 1981; Skow 2014)[7]. A causal

[6] Accounts of explanation that rely, for example, on understanding may be immune from the problem that motivates this paper insofar as they make no use of the concept of information. Of course, one could give an informational account of explanation that does rely on the concept of understanding, such as, "A model is an explanation (or is explanatory) when and only when it provides information that results in understanding". I thank an anonymous reviewer for pointing this out and for pressing me to clarify this section.

[7] This is a stipulation about the intended sense of "information" with which I am concerned. This is the sense of "information" at issue in the quotes that follow. If one does not accept this account of information, the puzzle with which I

history is merely the explanandum phenomenon's entire causal chain or network (Lewis 1986). For my purposes, I need not commit to any particular account of causation. To provide information about an explanandum phenomenon's causal history, it is not necessary to reduce the space of possible causal histories to one, which would be to provide "complete" information. Also, to provide complete information in this sense is not necessarily the same as providing a complete *explanation*. Due to the (often pragmatic) norms governing completeness of explanation, a complete explanation could be provided without complete information. I have made **IAE** non-comparative so that I will not have to address norms of completeness (Craver and Kaplan 2020). I take explanatory information to be veridical.[8]

I include "model" in **IAE** because when we make judgments about explanations, these explanations are usually embodied in models. Similarly, we would like a philosophical account to tell us when something is an explanation, and that something is usually a model.[9] By "model" or "representation," which I use synonymously, I mean a structure, broadly construed (e.g., a concrete replica, a mathematical equation, a diagram, or a linguistic description), that is interpreted to represent a target system (Weisberg 2013). Although I am ignoring Weisberg's (2013) distinction between models and model descriptions, this should not affect the points that follow.

It might be thought that information is too weak a basis for explanation ([redacted] personal communication; Salmon 1984). Part of the motivation for causal-mechanical accounts was to avoid the verdicts on which the covering law model floundered. One kind of case was where merely informational relations were counted as explanatory. For example, the barometer provides information

---

am concerned and all the arguments to follow can be cast simply in terms of possibility reduction. There is a prima facie conflict between the claims that 1) explanation is possibility reduction, 2) possibility reduction is agent-relative, and 3) some explanations are not agent-relative. I thank an anonymous reviewer for pressing me here.

[8] I do not take much to hang on this (cf. Lewis 1986, 226) and I will not address idealization here (see Craver 2014; Weisberg 2013). This should not affect my argument or solution, for information is still relative to background knowledge, whether or not it is veridical.

[9] Even proponents of the ontic conception recognize both that the term "explanation" is used this way and that their account needs to address these questions (Craver 2014; Craver and Kaplan 2020).

about the storm, but does not explain the storm (Salmon 1989)[10]. However, **IAE** does not entail that the barometer explains[11] the storm, because, although a barometer reading reduces uncertainty about the occurrence of a storm, it does not reduce uncertainty about the causes of the storm. Nor does the fact that barometer readings and storms are correlated reduce uncertainty about the causes of the storm (barring application of Reichenbach's controversial Common Cause Principle). The closest fact in this area that does provide explanatory information is this: that the barometer reading and the storm have a common cause. This reduces uncertainty about the causes responsible for the storm; it says that whatever the cause of the storm, it must be such that it also causes certain barometer readings. It excludes from the space of possible causes those that do not also cause certain barometer readings. This is explanatory information according to **IAE**, although obviously very limited.

An informational account of explanation has been explicitly advocated by, e.g., Jackson and Pettit (1990), Lewis (1986), Railton (1981), and Skow (2014). Obviously, there are differences between these philosophers' accounts. For my purposes, the relevant similarity is their emphasis on information. It is clear from their texts that they are not merely using the term "information" informally but have in mind a notion like uncertainty- or possibility-reduction that generates the puzzle with which I am concerned. For example, Lewis (1986: 217) writes that "to explain an event is to provide some information about its causal history". Although he also writes that an explanation is "a proposition about the causal history of the explanandum event" (218), thereby dropping the term "information," it is clear from his examples that a proposition about the causal history of the explanandum is one that provides information in the relevant sense (i.e., uncertainty- or possibility-reduction) about it. For

[10] The informational version of the epistemic conception that Salmon (1984) attacked says that scientific explanations are "ways of increasing our information about phenomena of the sort we are trying to explain" (97), whereas **IAE** says that scientific explanations are ways of increasing our information about *the causes of* explananda. Salmon objected to the lack of causation in the former, not its reliance on the concept of information (1984, 101).

[11] I switch between talk of facts or events providing information and talk of representations of facts or events providing information. I take it that both can provide information in the relevant sense, but if one wishes one can take the former to be shorthand for the latter. I thank an anonymous reviewer for this suggestion.

example, explanatory information consists not only of positive information about what was in the causal history of the explanandum, but also negative information about what was not in the causal history of the explanandum (220). Lewis explicitly states that "the test" of whether something provides explanatory information "is that it suffices to rule out at least some hypotheses about the causal history of the explanandum" (221). This is precisely the sense of information as uncertainty- or possibility-reduction that generates the problem. What hypotheses about the causal history of the explanandum a proposition rules out or is incompatible with depends on auxiliary propositions. For example, although alone the proposition that Alice was not at the scene of the crime suffices to rule out all causal histories in which Alice was at the scene of the crime, it rules out different causal histories depending on whether it is conjoined with the proposition that Alice or Bob was at the scene of the crime or with the proposition that Alice or Bob or Carol was at the scene of the crime. Relative to the former, it rules out all causal histories in which Bob was not at the scene of the crime; relative to the latter, it rules out all causal histories in which neither Bob nor Carol were at the scene of the crime (see Section 4 for more discussion of the relativity of information to background knowledge and note the similarity of this case to the shell game presented there). Jackson and Pettit (1992) refer back to Lewis when they write, "We endorse the view that the job of explanation is to provide information on causal history" (13; Lewis is cited approvingly at p.12). The same idea is expressed when Skow (2014), explicitly following Lewis, writes that explanatory information "narrows down the list of possible causes (or possible causal histories) of the event being explained" (448) or "rule[s] out some hypotheses about what caused E [i.e., the explanandum]" (450). Finally, Railton (1981) also has uncertainty- or possibility-reduction in mind when he writes, "On the analysis given here, a proffered explanation supplies explanatory information (whether we recognize it as such or not) to the extent that it does in fact (whether we know it or not) correctly answer questions about the relevant ideal text" (243). Here, though, it is not uncertainty about the causal history of the explanandum that is reduced in explanations, but uncertainty

about the ideal explanatory text. That Railton has the relevant sense of information in mind is especially clear when he writes that “the amount of information carried by a 'message' is proportional to the degree to which it reduces uncertainty. [...] Hence, information is a kind of *selection power* over possibilities” (244; original emphasis).

I have described an instance of an informational account of explanation and provided textual evidence that accounts relying on the relevant notion of information have been prominent in the philosophy of explanation. Next, I show how an informational account is arguably implicit in some core arguments of the mechanistic research program.

3. Information in the mechanistic research program

As far as I know, no prominent mechanists have explicitly endorsed an informational account of mechanistic explanation[12]. However, one is arguably implicit in some of their core arguments. I illustrate this with Piccinini and Craver's (2011) argument that functional analyses are mechanism sketches and Zednik's (2011, 2015) argument that dynamical systems models can be mechanistic explanations.

3.1 Functional analyses as mechanism sketches

A functional analysis explains a system's ability or capacity in terms of the functional properties of the whole system or of its parts. Functional analysis is thought to proceed relatively autonomously of consideration of the structural components that realize the functional properties, or play the functional roles, given the multiple realizability of such properties (Fodor 1968; Weiskopf 2011a,b). This provides prima facie reason for thinking that functional analyses are not mechanistic explanations of any kind. However, Piccinini and Craver (2011) argue that functional analyses are mechanism sketches, that is, incomplete mechanistic explanations (Craver 2007). Mechanism sketches lack relevant details because

[12] Note that according to such an account, two kinds of information could be explanatory: (causal) information about the mechanistic causal process that produces an explanandum result and (constitutive) information about the components, activities, and organization that constitute the mechanism that maintains, underlies, or produces the explanandum.

they contain black boxes and filler terms.

Piccinini and Craver (2011) distinguish three types of functional analysis: task analysis, functional analysis by internal states, and boxology. I only briefly describe these and give one example, since Piccinini and Craver's argument is basically the same in each case. A task analysis decomposes a capacity into subcapacities and their organization (Cummins 1975). A functional analysis by internal states explains a capacity in terms of a system's internal states and their interaction (Fodor 1968). Boxology analyzes a system in terms of its functional components or black boxes and their (often informational) interactions (Fodor 1968).

The reason that Piccinini and Craver (2011) claim that all three kinds of functional analysis are mechanism sketches is that each puts constraints on the possible mechanisms that implement the functions (or subcapacities) identified in the analysis. Function constrains structure. Similarly, structure constrains function: not just any structural component can perform any function. For example, take the boxological case of belief and desire boxes. The functionalist may claim that belief and desire boxes place no constraints on mechanisms. After all, they need not even be implemented or realized by two separate memory components. However, Piccinini and Craver argue, even if belief and desire boxes are implemented by the same memory component, this still places constraints on possible mechanisms. Implementing belief and desire boxes in the same memory component requires storing beliefs and desires in a single memory component, while ensuring that the system can accurately distinguish and keep track of which representations are beliefs and which are desires. Thus, even if belief and desire boxes are implemented by the same memory component, performing the functions of belief and desire boxes requires a mechanism(s) that is able to distinguish between and keep track of those two types of representation and transform them in relevant ways (Piccinini and Craver 2011: 303). Although this is consistent with the multiple realizability of functions, it does put some limits on what could possibly implement belief and desire boxes.

When a functional analysis constrains mechanisms, it limits the space of possible mechanisms that could implement the identified functions.[13] To describe as system as having belief and desire boxes is to reduce the space of possible mechanisms it contains. It excludes from the space of possible mechanisms those that cannot distinguish between and keep track of those two types of representation and transform them in relevant ways. But to reduce the space of possible mechanisms just is to convey information about mechanisms. Therefore, Piccinini and Craver's argument is that functional analyses are mechanism sketches because they provide information about mechanisms. I do not intend here to support Piccinini and Craver's argument; I only intend to show that it relies on an informational account of mechanistic explanation.[14]

3.2. Representational form as irrelevant to explanation

An informational account is also implicit in, or at least implies the crucial premise of, Zednik's (2011, 2015) argument that dynamical systems models can provide mechanistic explanations.[15] Dynamical systems models employ the mathematical concepts of dynamical systems theory, such as differential or difference equations (Chemero 2009; Izhikevich 2007). These equations allow the modeling of the evolution of the target system's variables over time, which can be represented graphically as a trajectory through state space (or phase space). The state space of a system is a high dimensional space that represents all its possible states, i.e., all possible values of all the system's

[13] It is possible that Piccinini and Craver could be working with a more robust notion of constraint that is inconsistent with an informational theory, but nothing in their paper suggests this.

[14] Piccinini and Craver's (2011) argument might not follow if Kaplan and Craver's (2011) model-to-mechanism-mapping (**3M**) requirement is true. According to **3M**, the variables in an explanatory model map onto specific structural components and activities of the explanandum phenomenon's mechanism, but you can constrain a mechanism without referring to it or its components and their activities, for example, by describing what the mechanism is *not* like. (Unless, of course, one has a liberal conception of properties according to which everything has an infinite number of negative properties.) Like mapping and reference, similarity is similarly too strong (Weisberg 2013) – accounts of explanation in those terms would not count as explanatory models that in fact are. This is because the information a model conveys – what can be learned from a model – and, so, its explanatory power, can outstrip what it explicitly represents. Yet, importantly, information about that on which an explanandum depends is all that is necessary to answer explanation-constituting w-questions. This contradicts **3M** even if it is consistent with Piccinini and Craver's argument.

[15] For arguments that dynamical models are not mechanistic explanations, see, e.g., Chemero 2009, Chemero and Silberstein 2011 and Stepp, Chemero, and Turvey 2011.

variables. Such graphical representations allow intuitive analysis of state space topology and reveal abstract, dynamical features such as attractors (states into which the system tends from surrounding states) (Izhikevich 2007).

Zednik's (2011) argument distinguishes between mechanistic explanations, on the one hand, and the tools used for constructing and representing them, on the other. As a mathematical and conceptual framework, dynamical systems theory can be used to represent anything to which its concepts apply. If dynamical concepts can apply to the components, activities, and organization of mechanisms, then, according to Zednik, dynamical systems theory can provide mechanistic explanations. Zednik (2015) has extended this point, using evolutionary robotics and network science to show how new tools for mechanism description and discovery can go beyond the traditional mechanistic explanatory methods of decomposition and localization (Bechtel and Richardson 1993). What matters for an explanation to be mechanistic, according to Zednik, is not how it was constructed or its representational form – what matters is only that it represents a mechanism.

For example, take Beer's (1996, 2003) dynamical model of perceptual categorization (or categorical perception) (see Zednik [2011] or Povich [forthcoming] for more discussion). The model is a simulated system consisting of a 14-neuron continuous-time recurrent neural network (CTRNN), inside an evolved model agent (i.e., its network architecture was constructed with an evolutionary algorithm), inside a two-dimensional environment. The agent moves horizontally as circles or diamonds fall from above. It 'categorizes' these objects by catching the former and avoiding the latter. The agent perceives with an eye consisting of seven rays, each connected to a corresponding sensory input neuron. When a ray hits an object, its input neuron receives a signal inversely proportional to the distance from the object – the closer the object when 'seen' by a ray, the greater its input signal. The agent with the best performance evolved a strategy of active scanning (Beer 2003). First, the agent centers the object in its field of view, then it moves back and forth, scanning the object. The scan

narrows to home in on circles, while breaking to avoid diamonds. Beer (2003: 228-9) explains this active scanning as follows. First, he decomposes the agent-environment dynamics into the effect of the relative positions of agent and object on the agent's motion, and vice versa. Then, for both circle and diamond trials, he superimposes the motion trajectory of the object through the agent's field of view onto a steady-state velocity field, which represents, for each point in the agent's field of view, the agent's steady-state horizontal velocity in response to an object at that point (228). Finally, he notices from an examination of the agent's motion trajectories that it consistently overshoots the midline of its visual field, due to the lag in time for the neural network to respond to sudden changes in sensory input. Therefore, according to Beer, active scanning is explained by the dynamic interaction of the steady-state velocity fields and the neural network's lag.

Zednik (2011) argues that Beer's model describes interactive components in an extended mechanism that spans brain, body, and environment. The explanandum is the behavior of one component in this mechanism, the agent's active scanning. The model describes how interactions with the environment, along with the time lag in responding to stimuli, result in active scanning. This example shows how, for Zednik, what matters for whether an explanation is mechanistic is not whether the model is particularly machine-like or modular but only whether it describes a mechanism.

Although Zednik's argument may not directly imply an informational account of mechanistic explanation, because representing a mechanism may be different than merely conveying information about it[16], it is consistent with and implied by an informational account. This is because models of many different forms, constructed by many different methods, can provide information about mechanisms. Therefore, an informational account of mechanistic explanation would imply that neither the form of a model nor the methods used to build it are relevant to whether it is (or provides) a mechanistic explanation. And Zednik is not alone. Distinctions such as his are increasingly common

[16] See Dretske (1988) for an informational, teleological theory of representation.

among mechanists. Hochstein (2016) makes a somewhat similar distinction between what he calls the "representation-as" and "representation-of" accounts of mechanistic explanation. According to the former, for an explanation to be mechanistic, it must represent a mechanism *as* a mechanism, i.e., the form of the representation itself must mirror the mechanism. According to the latter, any representation *of* a mechanism is a mechanistic explanation, regardless of the form of the representation. Levy (2013, 2014; see also Andersen 2014a,b), for example, makes a similar distinction between what he calls 'explanatory mechanism' and 'strategic mechanism'. Explanatory mechanism is the thesis that 'to explain a phenomenon, one must cite mechanistic *information'* (100, emphasis added). On the other hand, strategic mechanism 'articulates a way of doing science, a framework for representing and reasoning about complex systems,' using modeling methods such as decomposition and localization (104-5). Obviously, explanatory mechanism does directly imply an informational account of mechanistic explanation.[17]

4. A puzzle for informational accounts

Informational accounts of explanation have often been proposed and are implicit in some core arguments of the mechanistic research program. However, they seem to have some heretofore unseen counterintuitive consequences because information is most commonly thought to be relative to an agent's background knowledge. The relativity of information to background knowledge can be illustrated using Dretske's example of the shell game (1981, 78). There are four shells, and a peanut is under one of them. Alice, but not Bob, knows that the peanut is not under shells 1 and 2. Alice and Bob then both turn over shell 3 and see that it is empty. This gives different information to Alice and Bob: Alice learns that the peanut is under shell 4, while Bob only learns that it is not under shell 3[18].[19]

[17] I thank an anonymous reviewer for pressing me on this section.

[18] The difference in information conveyed to Alice and Bob can be precisely quantified (see Dretske 1981, 78). For Alice, 2 possibilities are reduced to 1, so she receives 1 bit of information; for Bob, 4 possibilities are reduced to 3, so he receives .42 bits of information.

Background knowledge, then, affects not only how much information is received, but *what* information is received (Dretske 1981, 81). In the most obvious and extreme case, an agent may not be able to extract any explanatory information from a model because she does not have the background knowledge to be able to interpret what the model says. Informational accounts make it possible that a model could be explanatory to Alice, but not Bob, because it conveys information to Alice that it does not convey to Bob, due to differences in their background knowledge. It seems that the explanatoriness of a model can only be assessed relative to particular agents, contradicting commonsense judgments and traditional philosophical goals of the theory of explanation. In lay and scientific practice, we often assess the explanatoriness of a model *simpliciter*, not just relative to some particular agent. Similarly, one of the things a philosophical account of explanation should do is tell us when a model is an explanation *simpliciter*, not just when a model is an explanation relative to some particular agent. One desideratum of an account of explanation should be to capture this "objectivity" of explanation. Thus, an account's reliance on the concept of information generates a *prima facie* conflict between the claims that 1) explanation is the conveyance of information, 2) information is relative to background knowledge, and 3) some models are explanations not relative to any particular agent. Note that this conflict arises even if one's account of explanation is not intended to be an account of explanation qua communicative or speech act, but merely from the account's reliance on the concept of information.

5. A solution

Informational theorists can avoid the above unwelcome consequences by relativizing to different sets of propositions and distinguishing scientific explanation *simpliciter* from explanation-to

[19] [Redacted] (personal communication) objects that Alice and Bob receive the same information but can draw different inferences because of their different background knowledge. Scarantino (2008, 633-4), citing the enhanced explanatory power of the concept of information, provides reason for resisting this way of thinking about the case (and information generally): "For instance, we won't be able to invoke the different information carried by the signal to the two players to explain why A[lice], unlike B[ob], made a bet that the peanut is under shell 4. The explanation of all differences in the behavioural impact of a signal owing to differences in background knowledge will have to be handed out to a separate and yet to be specified non-informational theory connecting information, learning and behaviour".

(i.e., explanation *to* a particular agent). For scientific explanation *simpliciter*, I propose to relativize information to "total scientific background knowledge" (TSBK), the current total store of propositions that are known the scientific community (cf. Kitcher [1989] on the explanatory store over *K*, and Craver and Kaplan 2020).

To determine the content of TSBK, any account of scientific social knowledge will do, though I am partial to Bird's (2010; 2014)[20]. Bird conceives of social knowledge as performing social functions analogous to the functions of individual knowledge. So, for p to be socially known, p must be true, accessible to relevant members of the community (e.g., other scientists, but not necessarily the lay public), propositional in nature, the product of social mechanisms whose function promotes truth, and available as an input into social action or social cognitive structures (Bird 2010, 42–4). Unlike Bird, I remain agnostic as to whether there exists a single social agent that knows all the propositions that make up TSBK. Notice that while I have been speaking of information as relative to background knowledge, information is simply relative to a set of propositions – it need not be the case that that set of propositions is known by an agent (which is why Scarantino [2015] prefers the term "background data").

When a model explains *to* an agent (or a particular group of agents, an audience), I relativize to that particular agent's background knowledge. Explanation-to is where most pragmatic concerns are likely to arise. A model needs to be presented in a particular way, tailored to the intended audience's presumed background knowledge, in order to maximize the probability that they will extract the appropriate information from it.

This distinction allows for it to be the case that a model is a scientific explanation *simpliciter*, but not an explanation *to* an individual agent, because their background knowledge does not allow them

[20] Bird's account has been criticized by Lackey (2014). See Povich (2018b) for a response.

to extract the appropriate information from the model[21]. Metaphorically, we might say that even if a model does not inform an individual agent, it could inform the scientific community as a whole. If Bird (2010; 2014) is right, this would not just be metaphor – explanation *simpliciter* would be a form of explanation-to where the agent involved is a social agent.

Let me motivate why I am using Bird's account to determine the relevant set of propositions to which information is relativized in explanations *simpliciter* and not some other set of propositions, for example, the set of all true propositions or the set of propositions known at the "end of inquiry". To understand why I propose relativizing to TSBK, consider why it makes sense to relativize to background knowledge K^A, rather than, say, the set of all true propositions, in explanations to an agent A. It is only relative to A's own background knowledge that she is able to learn something from the information conveyed by a signal. On any particular occasion, A might not actually extract information from a model, but she would be able to. An analogous thought is at work in my TSBK proposal. It is helpful to think of the scientific community as something like a group agent with group knowledge, though, again, I do not want to commit to any robust notion of group agency. Although no single individual knows TSBK and although the scientific community might not actually extract information, and so learn something, from a model on any particular occasion, it would be able to, via the division of cognitive labor (Bird 2014). Bird's functional account of social knowledge ensures that social knowledge is useful to, and actually able to be used by, the scientific community. Explanatory *simpliciter* models inform the scientific community about, or increase the scientific community's knowledge of, causes (or mechanisms or the ideal explanatory text or whatever). If information were

[21] [Redacted] ([personal communication]) objects that a model can contain explanatory information for A without being an explanation to A, because A might not understand the model. To accommodate this intuition, I could distinguish between a model's being an explanation to A and a model's actually explaining to A. In the former, a model contains the relevant information relative to A's background knowledge, but A might not extract it. In the latter, A actually extracts the information. An analogous distinction seems possible with respect to scientific explanation *simpliciter*. For the scientific community to actually extract some information from a model means that that information becomes part of its knowledge, TSBK. Obviously, how a piece of information becomes part of TSBK is a controversial question. Bird (2010, 32) suggests publication as the relevant process by which a piece of knowledge becomes scientific knowledge.

relativized to the set of all true propositions or the set of propositions known at the end of inquiry, we would be left with a conception of explanation that would never be of use to us, individually or collectively. Relativizing to TSBK captures just enough objectivity of explanation; any more would be inappropriate, given that models are interpreted representational structures used by us for our purposes.[22] I am willing to concede, therefore, that, for example, the Hodgkin-Huxley model of the action potential was not explanatory in the Middle Ages because no one, individually or collectively, had the ability to extract any information from it. We must either make this concession or, if we relativize to all true propositions or propositions accepted at the end of inquiry, concede that there are explanatory models from which it is impossible for us, individually or collectively, to learn anything. I prefer the former course, but the latter is also a way out of the inconsistent triad, and my formal solution below can be adapted to fit it. (Simply replace TSBK below with whatever set of propositions you want to relativize to.)

This solution can be expressed formally as follows. Think of the space of all possible causal histories of an explanandum phenomenon as a disjunction or set of those causal histories. Since this set contains all possibilities, its prior probability is 1. When a model M conveys information about an explanandum phenomenon's causal history, it reduces the space of possible causal histories, thereby excluding some possibilities, i.e., decreasing the probability of a subset of initial possibilities to 0. This has the effect of increasing the probability of the subset of *remaining* possibilities to 1, since we now know that the actual causal history is in this subset. If we call a proper[23] subset of possible causal histories C_R, then a model M explains *simpliciter* when and only when $P(C_R|M\&TSBK) = 1$, but <1 given TSBK alone (that is, without M; see below) (Dretske 1981). A model M explains to an agent A

[22] Contessa (2007, 53 fn. 6) notes that one of the few agreements in the literature on scientific representation is that representation is a triadic relation between a representational vehicle, a target system, and a user base. It should be unsurprising that explanation is similarly implicitly triadic.

[23] C_R must be a proper subset or no information is conveyed.

when and only when $P(C_R|M\&K^A) = 1$, but <1 given K^A alone (without M), where K^A is the agent's background knowledge. To be a bit more precise, and in order for these equations to make sense, we need to make sure that M and TSBK are of the right ontological categories. TSBK is the current total store of propositions that are known the scientific community. M might not itself be a set of propositions. Models come in many varieties, including but not limited to mathematical models, computational models, and concrete models (Weisberg 2013). Thus, "M" in the equations above should be interpreted as the *content* of a model. The content of a model is how the world is according to the model, what the model "says" about the world, or how it represents the world as being. I will not give an account of this here, but I think I can appeal to any of the extent accounts of scientific representation. For example, on Suárez's (2004) inferential conception of scientific representation, what a model says about the world is determined by its associated inference rules.

I include the "without M" condition above because without it, a model would cease to provide information and, so, cease to be explanatory, once it becomes part of the current store of scientific knowledge or an individual agent's background knowledge. We want to say that such a model is still an explanation because it provides information relative to the relevant set of propositions, the model itself (i.e., its content) *excluded*.

If the probability of unity above is worrisome because it implies certainty, note that C_R is a *set* of causal histories (or mechanisms or propositions in the ideal explanatory text or whatever), not a particular causal history. The probability of unity here only implies that we can reduce the probability of some other possibilities (C_R's complement) to zero. If one is skeptical of this, we could instead require that M only probably exclude C_R's complement. This would reduce the probability of C_R's complement, but not to 0, increasing the probability of C_R, but not to 1 (see Scarantino 2015). Then, a model M explains *simpliciter* when and only when $P(C_R|M\&TSBK) > P(C_R|TSBK)$. A model M explains to an agent A when and only when $P(C_R|M\&K^A) > P(C_R|K^A)$.

Before concluding, let me briefly say something about the ontic conception, which I mentioned several times earlier, and the relevance to it of my solution to the above puzzle. The debate between the ontic and epistemic conceptions of scientific explanation has been evolving (Illari 2013; Povich 2018a). Salmon framed the debate in terms of what explanations do (though see Wright 2015). Many philosophers (e.g., Strevens 2008, Wright 2012, among others) since have framed it metaphysically, as a debate about what explanations are: the ontic conception was associated with the claim that scientific explanations are ontic structures, usually causes and mechanisms; the epistemic conception was associated with the claim that scientific explanations are epistemic states or representations. Craver's (2014) most recent formulation of the ontic conception backs away from the metaphysical claim and instead focuses on what he calls demarcatory and normative constraints on explanation: "in order to satisfy these two objectives [of explanatory demarcation and explanatory normativity], one must look beyond representational structures to the ontic structures in the world" (28). Thus, according to this construal of the ontic conception, an informational account like Lewis' (1986) seems to count as an ontic conception because according to that account, what demarcates explanatory from non-explanatory representations is that only the former provide information about certain "ontic structures in the world" (i.e., causes) and what distinguishes good from bad explanation (i.e., normative constraints on explanation) more relevant information about causes (Craver and Kaplan 2020 is also relevant here). In fact, Craver (2014) speaks of information throughout his most recent presentation of the ontic conception, most relevantly for my point when he writes, "If the philosophical topic of explanation is to provide criteria of adequacy for scientific explanations, then the ontic conception is indispensable: explanatory communications, texts, and representations are evaluated in part by the extent to which they deliver more or less accurate information about the ontic explanation for the explanandum phenomenon" (36-7). If proponents of the ontic conception, as construed by Craver (2014), think that "communications, texts, and representations" are explanatory to the extent that they provide

information about “ontic explanations”, then the question arises as to whom or what such information is relative. This is where my solution can come in handy. The ontic proponent can then distinguish between the ontic explanation (cause, mechanism, whatever), models that are explanations-to (which provide information, relative to particular individuals, about the ontic explanation), and models that are explanations simpliciter (which provide information, relative to TSBK, about the ontic explanation).

6. Conclusion

Informational accounts have an impressive history in the philosophy of explanation (e.g., Jackson and Pettit 1990; Lewis 1986; Railton 1981; Skow 2014). I have shown how an informational account is also implicit in core arguments of the mechanistic research program. An informational account is implied by Piccinini and Craver's (2011) arguments that functional analyses are mechanism sketches; an informational account is at least consistent with and implies the central premise in Zednik's (2011, 2015) argument that dynamical systems models can be mechanistic explanations.

Informational accounts, however, seem to have counterintuitive consequences, for they seem to imply that assessments of explanatoriness can only be made relative to individual agents. This conflicts with lay and scientific commonsense judgments about explanation and traditional goals of the theory of explanation. By distinguishing scientific explanation *simpliciter* from explanation-to and relativizing the former to TSBK, these counterintuitive consequences are avoided.

University of Rochester

Department of Philosophy

mapovich@gmail.com

References

Batterman, R.W. and C. Rice. 2014. Minimal model explanations. *Philosophy of Science* 81: 349–76.

Bechtel, W. and R.C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Beer, R.D. 1996. 'Toward the evolution of dynamical neural networks for minimally cognitive behavior', in P. Maes, M. Mataric, J. A. Meyer, J. Pollack, and S. Wilson (eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. Cambridge, MA: MIT Press. pp. 421–9.

Beer, R.D. 2003. 'The dynamics of active categorical perception in an evolved model agent', *Adaptive Behavior*, 11(4): 209–43.

Bird, A. 2010. Social knowing: the social sense of 'scientific knowledge'. *Philosophical Perspectives* 24: 23–56.

Bird, A. 2014. When is there a group that knows? Distributed cognition, scientific knowledge, and the social epistemic subject. In *Essays in Collective Epistemology*, ed. J. Lackey, 42–63. New York: Oxford University Press.

Chemero, A. 2009. *Radical Embodied Cognitive Science.* Cambridge, MA: MIT Press.

Chemero, A. and M. Silberstein. 2008. After the philosophy of mind: replacing scholasticism with science. *Philosophy of Science* 75: 1–27.

Chirimuuta, M. 2014. Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese* 191: 127–53.

Cohen, J. and A. Meskin. 2006. An objective counterfactual theory of information. *Australasian Journal of Philosophy* 84: 333–52.

Contessa, G. 2007. Scientific representation, interpretation, and surrogative reasoning. *Philosophy of Science* 74: 48–68

Craver, C.F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, C.F. 2014. The ontic account of scientific explanation. In *Explanation in the Special Sciences: The Case of Biology and History*, ed. M. Kaiser, O. Scholz, D. Plenge and A. Hüttemann, 27–54. New York: Springer.

Craver, Carl F., and David M. Kaplan. 2020. Are more details better? On the norms of completeness for mechanistic explanations." *The British Journal for the Philosophy of Science* 71.1: 287–319.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72: 741–65.

Demir, H. 2008. Counterfactuals vs. conditional probabilities: a critical analysis of the counterfactual theory of information. *Australasian Journal of Philosophy* 86: 45–60.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Fodor, J.A. 1968. *Psychological Explanation*. New York: Random House.

Hochstein, E. 2016. One mechanism, many models: a distributed theory of mechanistic explanation. *Synthese* 193: 1387–407.

Illari, P. 2013. Mechanistic Explanation: Integrating the Ontic and Epistemic. *Erkenntnis* 78: 237–55.

Izhikevich, E.M. 2007. *Dynamical Systems in Neuroscience*. Cambridge, MA: MIT Press.

Jackson, F. and P. Pettit. 1992. In defense of explanatory ecumenism. *Economics and Philosophy* 8: 1–21.

Kaplan, D.M. and C.F. Craver. 2011. The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. *Philosophy of Science* 78: 601–27.

Kitcher, P. 1989. Explanatory unification and the causal structure of the world. In *Scientific Explanation*, eds. P. Kitcher and W. Salmon, 410–505. Minneapolis: University of Minnesota Press.

Lackey, J. 2014. Socially extended knowledge. *Philosophical Issues* 24: 282–98.

Lewis, D. 1986. Causal explanation. In his *Philosophical Papers, Vol. 2,* 214–40. New York: Oxford University Press.

Meskin, A. and J. Cohen. 2008. Counterfactuals, probabilities, and information: response to critics. *Australasian Journal of Philosophy* 86: 635–42.

Piccinini, G. and C.F. Craver. 2011. Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese* 183: 283–311.

Povich, M. 2018a. Minimal models and the generalized ontic conception of scientific explanation. *The British Journal for the Philosophy of Science* 69(1): 117–37.

Povich, M. 2018b. Social Knowledge and Supervenience Revisited. *Erkenntnis* 83(5): 1033–43.

Povich, Mark (Forthcoming) 'Mechanistic explanation in psychology.' *The SAGE Handbook of Theoretical Psychology* (Eds.) Hank Stam and Huib Looren de Jong.

Railton, P. 1981. Probability, explanation, and information. *Synthese* 48: 233–56.

Rice, C. 2015. Moving beyond causes: optimality models and scientific explanation. *Noûs* 49: 589–615.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Salmon, W. 1989. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.

Scarantino, A. 2008. Shell games, information, and counterfactuals. *Australasian Journal of Philosophy* 86: 629–34.

Scarantino, A. 2015. Information as a probabilistic difference maker. *Australasian Journal of Philosophy* 93: 419–43.

Skow, B. 2014. Are there non-causal explanations (of particular events)? *The British Journal for the*

Philosophy of Science 65: 445–67.

Skow, B. 2016. *Reasons why*. Oxford University Press.

Sober, E. 1983. Equilibrium explanation. *Philosophical Studies* 43: 201–10.

Stepp, N., A. Chemero, and M. Turvey. 2011. Philosophy for the rest of cognitive science. *Topics in Cognitive Science* 3: 425–37.

Strevens, M. 2008. *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.

Suárez, M., 2004. An inferential conception of scientific representation. *Philosophy of Science* 71(5): 767–79.

Weisberg, M. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Weiskopf, D.A. 2011a. Models and mechanisms in psychological explanation. *Synthese* 183: 313–38.

Weiskopf, D.A. 2011b. The functional unity of special science kinds. *British Journal for the Philosophy of Science* 62: 233–58.

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

Wright, C. D. 2012. Mechanistic Explanation without the Ontic Conception. *European Journal of Philosophy of Science* 2: 375–94.

Wright, C.D. 2015. The ontic conception of scientific explanation. *Studies in History and Philosophy of Science Part A* 54: 20-30.

Zednik, C. 2011. The nature of dynamical explanation. *Philosophy of Science* 78: 238–63.

Zednik, C. 2015. Heuristics, descriptions, and the scope of mechanistic explanation. In *Explanation in Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences*, ed. C. Malaterre and P.-A. Braillard, 295–318. Dordrecht: Springer.