Knowledge of Counterfactual Interventions Through Cognitive Models of Mechanisms

Jonathan Waskan
Department of Philosophy
University of Illinois at Urbana-Champaign
waskan@illinois.edu

Abstract: Here I consider the relative merits of two recent models of explanation, James Woodward’s interventionist-counterfactual model and the model model. According to the former, explanations are largely constituted by information about the consequences of counterfactual interventions. Problems arise for this approach because countless relevant interventions are possible in most cases and because it overlooks other kinds of equally relevant information. According the model model, explanations are largely constituted by cognitive models of actual mechanisms. On this approach, explanations tend not to represent any of the aforementioned information explicitly but can instead be used to produce it on demand. The model model thus offers the more plausible account of the information of which we are aware when we have an explanation and of the ratiocinative process through which we derive many kinds of information that are relevant to the evaluation of explanations.

Biographical Sketch: Jonathan Waskan is an assistant professor of philosophy at the University of Illinois, Urbana-Champaign. He also holds an appointment with the Beckman Institute for Advanced Science and Technology. He received his degree in 1999 from Washington University's Philosophy-Neuroscience-Psychology program. He recently published a book, *Models and Cognition*, with the MIT Press, and he has...
published articles in such journals as *Cognitive Science, Connection Science,* and *Southern Journal of Philosophy.*
Knowledge of Counterfactual Interventions

Through Cognitive Models of Mechanisms

1. INTRODUCTION

No recent model of explanation has been able to garner anything close to the level of support once enjoyed by the deductive-nomological (D-N) model. This may be due in part to the ever-growing litany of demands now placed upon models of explanation. Models must, to start with, supply reasonable and principled analyses for a wide array of cases (e.g., those involving flagpoles, syphilis, ink bottles, legionnaire’s, and so forth). They may also be judged for their ability to make sense of the role played by explanations in our lay and scientific lives. They may, for instance, be extolled or denigrated depending upon how well they answer questions such as these:

(a) In virtue of what, precisely, do explanations render events and physical regularities intelligible?
(b) From whence originates the information on the basis of which explanations are, directly or indirectly, assessed?

I gather that it is in no small measure because of its ability to answer questions like these that James Woodward (2002, 2003, 2004) favors the interventionist-counterfactual (I-C) model of explanation, the central claim of which is that explanations convey information about what would have happened to the explanandum had we altered the factors constituting the explanans in various ways (2003, 11). I favor, on similar grounds, the model model, according to which one has an explanation for an event or physical regularity only if one possesses an intrinsic
cognitive model (ICM) – roughly, a non-sentential mental representation akin to a scale model – of mechanisms that would produce it.

I argue here that certain epistemic principles cast doubt on the I-C model’s answer to (a) and, thereby, on its answer to (b). I contest neither Woodward’s claim that explanations in some sense embody information about counterfactual interventions nor his claim that this information is highly relevant to the evaluation of explanations. I do claim, however, that problems arise for the I-C model because of the sheer volume of this information and because it overlooks other kinds of equally relevant information. According to the model model, explanations do not, in and of themselves, represent any of this information explicitly, though they can be used to produce it on demand. For this reason, the model model offers a more credible account of both the information of which we are aware when we understand a happening and the ratiocinative process by which other kinds of evaluatively relevant information are derived.

2. EXPLANATION AND COGNITION

In Making Things Happen, James Woodward (2003) claims that, in the most general sense, explanations are psychological in nature; they are 'in here' rather than 'out there' (23). Woodward and I agree on this important point, so I will not defend it here. However, in order to flesh out certain important details of this proposal, I will point out that we are by no means alone. Wright and Bechtel argue, for instance, that "explanation is through-and-through an epistemic practice of making the world more intelligible …" and that it proceeds through the deployment of mental representations (2007, 51; also see Horst 2007, 56). Similarly, Jaegwon Kim claims that explaining is
an epistemological activity – it is a process of alleviating puzzlement – and that

*explanations* are products of this activity constituted by knowledge of, or at least

beliefs about, the world (1988, 1993). Thus, says Kim, explanations, these things that

we *have* that alleviate puzzlement and render happenings intelligible, are "on the side

of the "subjective" rather than that of the "objective," on the side of "representation"

rather than that of reality represented" (1988, 225). Constituted as they are by mental

representations, explanations have ‘factual content’; they assert something about the

world. Explanations can therefore be accurate or inaccurate (or somewhere in

between) depending upon the extent to which what an explanation ‘says’ is the case is

in fact the case (226).

In keeping with this basic picture of what it is to have an explanation,

Woodward claims, "It is a plausible constraint on what an explanation is that it must

be something that provides understanding" (2003, 179). This gives way to a further

constraint:

explanations work by conveying information that provides understanding, and

this means that such information must be epistemically accessible to those

who provide the explanation or are enlightened by it. If some of the

information allegedly conveyed by an explanation cannot be grasped or

recognized by those who use the explanation, it is not through recognition of

that information that the explanation provides understanding (308).

To reiterate, explanations provide understanding by conveying information, so the

information must be such that it can be recognized or grasped. After all, as

Woodward intimates in the final clause of the above passage, *it is only through*
recognition of the information conveyed by explanations that they provide understanding. Turning now to the first sentence of the above passage, I suggest that we take things a step further than Woodward: Surely we should replace the weak claim that the information must be epistemically accessible with the stronger claim that the information must be epistemically accessed.² It is not because information is able to be called to mind that it produces understanding; it is because the information actually is called to mind. For instance, information tucked away in long-term declarative memory is epistemically accessible, but so long as it remains tucked away it will no more produce understanding than if it were bound up in procedural knowledge. Likewise, tacit knowledge – by which I mean the sort exemplified by your knowledge, which I presume has hitherto not entered your thought processes, that airships are capable of preventing flagpoles from casting shadows – is epistemically accessible, but unless it is accessed it cannot produce explanatory insight, enlightenment, or understanding. In short, information regarding a happening that we could think of but that we have not thought of cannot be what generates the feeling that we understand why the happening occurred. It seems appropriate, then, to say that in order to produce understanding the information conveyed by an explanation must be contained in occurrent, as opposed to dispositional (i.e., stored in long-term memory) or tacit (i.e., stored nowhere but producible on demand) mental states.

If the foregoing is correct, then a promising way to gain deeper insight into the nature of explanations is to determine what is going on, cognitively speaking, when we have an explanation. We can do this by figuring out what the contents of
explanations are, or what it is that they assert about the world. We can also investigate the nature of the vehicles bearing this content. Whereas Woodward opts for the former strategy, my focus is largely, though by no means exclusively, upon the latter. Despite our somewhat diverging interests, there is considerable tension between our views.

3. WOODWARD'S I-C MODEL

Woodward claims that explanations provide understanding by conveying information about causal relationships, which, he maintains, just is information that will "enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways" (italics mine, 2003, 11). Woodward adds that genuine explanations must carry information not only about conditions under which their explananda would be different, but also about "the range of conditions under which their explananda hold" (191). The first requirement is what makes his a counterfactual model of explanation. What makes it interventionist is that the differences in the state of the explanans are the direct result of imagined (though not necessarily physically possible) interventions.

It is clear that Woodward is – as the subtitle of Making Things Happen: A Theory of Causal Explanation suggests – offering up a theory of explanation and not just a set of criteria that any theory must satisfy. Woodward's main concern is with the contents of explanations, or with what they assert about the world in virtue of which they provide understanding (Woodward 2003, 7, 11; 2004, 57, 63). He states, "The theory of causal explanation that I have been sketching is … one that ties
explanatory import very closely to the provision of certain kinds of counterfactual information: it might fairly be described as a counterfactual theory of causal explanation" (italics mine, 2003, 196). His central thesis is that explanations enlighten by making assertions about how the state of an explanandum would and would not change in light of various alterations to an hypothesized cause. We must, once again, bear in mind that the contents in question are the contents of mental states (also see ftn. 10), presumably beliefs, about the cause of the explanandum and that insofar as they provide understanding such mental states must be occurrent rather than dispositional or tacit.

4. ICMs AND THE MODEL MODEL

The model model might be considered the psychologized offspring of mechanistic approaches towards explanation advanced in the philosophy of science (Salmon 1984; Glennan 1996; Machamer, Darden, and Craver 2000). Broadly speaking, according to the model model, to have an explanation is to have the belief that a certain mechanism is, or may be, responsible for producing some happening, where such beliefs are constituted by mental representations of those mechanisms. It is largely in virtue of our awareness of the information conveyed by these representations that events and physical regularities are rendered intelligible. According to the model model, the representations in question are ICMs. The rationale behind this proposal will become clearer following a brief, partial review of the justification for the wider claim that humans harbor and manipulate ICMs (Waskan 2003, 2006).
4.1 Computational, Scale, and Cognitive Models

Certain computational systems harbor what may be termed *intrinsic computational models* (InCoMs), examples of which include the virtual reality models used by the entertainment industry and the finite element models used in nearly all of the engineering disciplines and in most of the sciences. Although the representational media used for constructing these models are best understood in terms of sentence-and-rule based representations, at the higher level of the models realized by such media one finds non-sentential representations that share with scale models several of the properties that distinguish them from sentence-and-rule based representations. The most important for present purposes is that these models support, through finite means, boundless inferences concerning both the consequences of alterations to the objects they represent and the countless defeaters of those consequences; they exhibit *inferential productivity*. Put differently, InCoMs and scale models are immune to the frame problem of artificial intelligence (A.I.), which is the challenge of getting a representational system to predict what will change and what will stay the same following alterations to the state of the world (McCarthy & Hayes 1969; Bechtel, Abrahamsen & Graham 1998).

The frame problem can be broken down into two main component problems, the prediction problem (Janlert 1996) and the qualification problem (McCarthy 1986). The prediction problem, as it afflicts sentence-and-rule-based systems, is this: In order to embody what the average human knows about the consequences of worldly alterations, a sentence-and-rule-based system would have to contain countless rules (*frame axioms*) that provide explicit specifications of how objects, both familiar and
novel, will behave relative to one another following each of innumerable possible alterations. This problem, which is sufficiently worrisome, is compounded by the qualification problem (McCarthy 1986) in that in order to embody human knowledge of the consequences of worldly alterations, not only would an infinite number of rules be required, but nearly every rule would have to be qualified in an infinite number of ways. The general short-coming of sentence-and-rule-based representations is that they only support predictions concerning the consequences of alterations and the defeaters of those consequences if those alterations, consequences, and defeaters have all been spelled out *antecedently* and *explicitly*. Representations of this sort I call *extrinsic* representations.\(^5\)

InCoMs and scale models, on the other hand, exhibit immunity to the frame problem because one can determine on an as-needed basis the consequences of countless alterations to the represented system and the countless defeaters of those consequences simply by manipulating the model in the relevant ways. For instance, a finite element model of a new automobile design can be used to determine how the actual system will hold up under various conditions by subjecting the model to the virtual counterparts of as many of those conditions as one can think of. Put succinctly, the reason why modelers are able to provide an explicit specification of the principles operative at the level of simple modeling elements but not those operative at the higher level of, for instance, everyday objects is that the furniture of the former world is vastly less complex than the furniture of the latter. The challenge is thus not to come up with huge numbers of rules, but to crunch many applications of comparatively few rules. The upshot is that, just as with scale models constructed
from nomologically constrained media, with the intrinsic computational models constructed from primitively constrained media, the consequences of countless distinct alterations and the countless defeaters of those consequences are represented as 'automatic' outcomes of the corresponding alterations to the model (see Haugeland 1987, 91). All of the relevant information is therefore implicit in the representation and so there is no need to represent it explicitly through the use of innumerable distinct data structures. Representations of this sort I term intrinsic representations.

The intrinsic cognitive models (ICM) hypothesis is just the claim that humans also utilize, for many of their everyday truth-preserving needs, similar models of everyday objects, properties, and events. The hypothesis is lent tremendous credence by the fact that it is the only model of mental representation that explains why our own thought processes are immune to the frame problem. It is, in other words, the only model that makes sense of our boundless tacit knowledge (see section 2) of worldly alterations.6

4.2 The Consequences of Modeling Mechanisms

According to the model model, ICMs enlighten us as to the possible why’s and how’s behind a happening by enabling us to envision the mechanisms that might have produced it. 'Produce', in this context, is meant to be broader in scope than 'cause', since the former also applies to cases where one has an explanation for how certain low-level happenings produce, in a non-causal (synchronic) sense of the term, higher-level happenings. To get a good sense for what it means to have an ICM of how a mechanism produces, in the causal sense, a particular happening, let us first
consider what it means to have an InCoM of how a mechanism produces a particular happening.

As mentioned above, primitively constrained computational modeling media can be used to construct models of physical systems that are like scale models in that they make the representation of a subsequent event (the effect) an 'automatic' outcome of the representation of some prior event (the cause). This accords well with the picture of explanation painted by many naïve physics researchers to the effect that individuals create non-sentential cognitive models of the world and 'run' these models in order see how the explanandum quite literally follows from the model’s starting conditions (see Chi et al. 1982; Gentner & Stevens 1983; Schwartz & Black 1996; Schwartz 1999; Hegarty 2004). The use of intrinsic models of mechanisms for this purpose constitutes a fully monotonic *exductive* reasoning process, one that reveals to us a form of necessitation that differs qualitatively from that revealed by deductive reasoning (Waskan 2006, 157; cf., Pylyshyn 1984, 103; Lindsay 1988, 232; Glasgow & Papadias 1992, 373-4). Thus, at least with regard to cases where our knowledge of causes is sufficiently deep to render the effects thereof intelligible – that is, in uncontroversial cases of causal *explanation* – the proposal that humans reason on the basis of intrinsic models of mechanisms does seem to make significant headway towards an understanding of what is involved in believing that causes necessitate, or force into being, and are thereby relevant to, their effects.

Although intrinsic models contain a tremendous amount of implicit information, it is important to realize that they are also fully capable of representing information explicitly. This because the alterations we make to intrinsic models need
not occur at the outset, but may occur even as events in the model unfold. This kind of explicit representation of information comes at a cost (e.g., inference processes merely about what is represented explicitly may not be inferentially productive), but that it can occur is important for various reasons. One is that working memory or knowledge limitations may prevent us from modeling both a set of mechanisms and the deeper mechanisms that enforce regularities in the behavior of that set of mechanisms.\(^9\) For instance, when understanding the behavior of a small neural network, it may be important to represent to ourselves the fact *that* ions move towards regions of opposite charge even if we are unable to, or uninterested in, simultaneously representing *why* this regularity holds (*cf.* Schwartz & Black, 1996). Because an ICM of a given mechanism can be constituted in part by such explicit representations of happenings, an explanation can 'bottom out' (Machamer, Darden, & Craver 2000), in particular places, at just about any activity or regularity that one can think of. These activities or regularities may be hypothesized brute activities, they may be induced from experience, or they may be derived from knowledge of the deeper mechanisms sustaining them. Indeed, bottom-out regularities may even (*pace* Woodward 2002) include regular connections, or correlations, between events that are the common cause of some third factor. As with finite element models, the thing about ICMs (even those constituted in this way by explicit representations) is that they enable us to see that *as long as the bottom out regularities hold*, certain events will follow automatically. Thus explanations partly constituted by explicit representations can still be "fully adequate" insofar as they specify the conditions that would suffice for occurrence of the happening in question (Hempel & Oppenheim 1948, 138). It is
merely that such explanations will be, as perhaps all in fact are, shallow in some respects and deep in others.

5. WHAT EXPLANATIONS DO NOT (EXPLICITLY) ASSERT

One thing Woodward and I agree upon is that a model of explanation should say something about the origins of the information about counterfactuals that figures, directly or indirectly, in the evaluation of explanations (Woodward 2003, 22; 2004, 57). However, given the epistemic constraints on explanation introduced above, it seems it intuitively implausible that explanatory information just is information about counterfactual interventions, or that we are able to derive this information from explanations because it is their very content (2003, 7, 11, 14, 191). In fact, there is a good deal more information associated with most explanations, information that is equally relevant to their assessment but that often has little or nothing to do with either the consequences of hypothetical interventions or any other counterfactuals. As we shall see, this forces proponents of the I-C model to make some hard choices insofar as the contents of explanations are concerned. The model explains the origins of all of this information, and without singling out any one or any few pieces of this information as that which the explanation asserts in virtue of our awareness of which the happening is rendered intelligible.

5.1 Explanatory Implications: An Inventory

Consideration of the killer asteroid explanation for the mass extinction that occurred during the Cretaceous period brings these points out clearly (see Woodward 2003, 10-11). As we have seen, Woodward claims that to have an explanation is to have information that provides understanding. We also saw that in order to provide
understanding the information must be recognized or grasped; it must be contained in 
occurrent mental states. Woodward central thesis, once again, is that the information 
in question specifies how various alterations to the factors constituting the explanans 
would and would not give rise to changes in the explanandum (2003, 11, 191).

To be clear on what the factors are in this case, the basics of the killer asteroid 
explanation are something like the following: When a massive asteroid crashed into 
the Earth a tremendous amount of material was kicked up into the atmosphere and 
ev eventually blocked out the Sun throughout the globe, and this led to the progressive 
cooling of the Earth. As a result, plants and animals that required the previous level of 
sunlight or warmth for their survival died off in droves, as did creatures that primarily 
feed on those plants and animals, and so on. Here, then, is a list of various changes to 
these explanatory factors and their consequence and non-consequences for the 
explanandum, the mass extinction:

If the asteroid had been diverted away from the Earth, then (ceteris paribus) 
the chain of events leading to the extinction would not have occurred.

If the asteroid had been made of a different kind of rock or metal of the same 
total mass, then (ceteris paribus) the mass extinction would still have 
happened.

If the asteroid had been obliterated by a barrage of nuclear devices prior to 
impact, then (ceteris paribus) the chain of events leading to the extinction would not have occurred.
If the impact could have been slowed down so that it was more of a landing than a crash, then (ceteris paribus) the amount of material kicked up into the atmosphere would not have sufficed to alter the climate enough to cause the mass extinction.

If the tremendous amount of material kicked up into the atmosphere were immediately collected by a huge number of scrubbers, then (ceteris paribus) not enough solar radiation would have been reflected back into space to produce the mass extinction.

If the atmospheric temperature were maintained throughout the globe by a system of thermostat-controlled, nuclear-powered radiators, then (ceteris paribus) …

If we grant the above epistemic constraint and follow the above formulation of the I-C model, we find that one does not have the killer asteroid explanation, and thus (barring possession of some other explanation) the explanandum has not been rendered intelligible, unless one explicitly thinks of various ways (such as those listed here) in which changes in the character of the explanatory factors mentioned above would and would not change (e.g., lessen or eliminate) the mass extinction. It is, therefore, an implication of this proposal that one can read and comprehend the above description of the chain of events leading from the asteroid impact to the mass extinction and yet still not understand the possible basis for the mass extinction. One must, in addition, explicitly think of various ways, such as those listed here, in which
changes in the character of the explanatory factors mentioned above would and would not change (e.g., lessen or eliminate) the mass extinction. Thus, a further consequence of the I-C model is that anyone in possession of the killer asteroid explanation should, with hardly any effort at all, be able to produce a list like this one.

Once these implications of the I-C model are brought to light, it starts to look far less plausible. Surely one can gain considerable insight into the possible basis for the mass extinction merely by reading and comprehending that description of what may have been the actual chain of events. At the same time, I doubt many would go on record as claiming that in order to comprehend the above description of the chain of events leading from the asteroid impact to the mass extinction, one must have thoughts either about diverting the asteroid, different kinds of materials, nuclear explosions, scrubbers, about the mass extinction failing to occur altogether (or about any other relevantly similar set of counterfactual circumstances). As for the task of producing a list of counterfactuals like the one above, it is not accomplished without a certain amount of effort. If only explanations wore their implications on their sleeves!

Before we consider how a proponent of the I-C model might respond to these charges, we should first bear in mind that one who possesses an explanation like the one described above also knows about the consequences of alterations that have nothing to do with the explanandum. One who has the killer asteroid explanation, for instance, should be able to infer that had extraterrestrials painted "Earthlings Suck" on the asteroid, the impact would have obliterated all evidence of this fact. Moreover, explanations can also be used to make inferences that have nothing to do with alterations to either the explanans or explanandum. Such inferences play a big role in
the evaluation of explanations whose hypothesized explananda are beyond our spatial or temporal reach. For instance, from the combination of the killer asteroid explanation and the layering explanation for the formation of sedimentary rocks, one can infer that one should find, throughout the globe, in the layers of sediment that were laid down at around the time of the cataclysm an unusual concentration of materials (e.g., ash or nickel) that would have been thrown up by the asteroid impact.

It is also no small matter that this inference is, like the others we have been considering, tacitly qualified in innumerable ways, for scientists often invoke these qualifications if things do not pan out as they expect. For instance, if an initial inspection of sedimentary rock at a few different locations fails to turn up any ejecta material, a scientist who favors the killer asteroid explanation can justify his decision to hang onto that explanation by pointing out that the inference had the following, tacit qualification: It must not be the case that both the ejecta was initially evenly distributed in the upper atmosphere and, as it filtered down into lower layers, it was concentrated (e.g., by low pressure systems or atmospheric currents) and deposited into relatively discrete regions. In that case, we would expect to find ejecta material in the sedimentary rocks of some regions, but not others.

Assuming, on the other hand, that the initial search is successful, another important implication of the killer asteroid explanation – an implication that, once again, is relevant to the assessment of the theory but has nothing to do with alterations to either explanans or explanandum – is that at layers above the ones where the foreign material is found, one should in general find far fewer fossils of cold-blooded animals than at lower layers. At the same time, there should be a less pronounced
decrease in the quantity of fossils of creatures that are more fit for climate change (e.g., omnivorous creatures with insulation such as fat or feathers). This is so provided (*inter alia*) it is not the case that massive or violent geological processes (e.g., landslides, colliding plates, volcanic eruptions) have mixed up the layers of sedimentary rock in certain regions.

We have seen that the I-C model limits the information provided by explanations to counterfactuals, specifically to miscellaneous

(i) information regarding how alterations to the factors constituting the explanans would and would not give rise to changes in the explanandum.

It should now be clear, however, that explanations provide far more information than this. They can be used to produce

(ii) information regarding the consequences of alterations to the explanans that have nothing to do with changes in the explanandum,

(iii) information regarding the implications of an explanation that have nothing to do with changes in either the explanans or explanandum, and

(iv) information regarding conditions that might prevent the consequences specified by (i) – (iii) from obtaining.

Notice also that the type of information specified by (iii) does not, and the type of information specified by (iv) need not, concern interventions or counterfactuals of any sort.

### 5.2 Privileged Information

Let us consider now what options are available to proponents of the I-C model insofar as the contents of explanations are concerned, assuming, once again, that it is
in virtue of our awareness of what explanations assert that they produce understanding. To begin with, they might claim that what explanations assert about the world are all of the counterfactuals falling under (i). However, typically associated with any given explanation will be a tremendously long, potentially infinite list of such counterfactuals, and in either case it is quite implausible that explanatory insight stems from awareness of all such counterfactuals.

Another approach would be to say that just one, or – to stick with the formulation of the I-C model taken from Woodward (2003) – perhaps various counterfactuals constitute the content of the explanation. However, I can find no principled reason for giving any one, or any few, of the possibly infinite number of counterfactuals the privileged status of the content of the explanation, or that which it asserts in virtue of our awareness of which the happening is rendered intelligible. Indeed, take any counterfactual of the above sort that you wish, one can arguably have a particular explanation without ever becoming aware of it. More pressing still, there seems no principled basis on which to privilege members of (i) over members of (ii) – (iv), not all of which involve counterfactuals and all of which are equally relevant to the evaluation of explanations.

One could, perhaps, modify the I-C model so that it encompasses some or all of these types of information, thereby opening up a wider assortment of potential responses. However, all such options seem just as arbitrary as the choice of some one or few counterfactuals from (i). Indeed, such options seem more arbitrary still once one recognizes that there is a model of explanation waiting in the wings that, unlike the I-C model, forces no such choices, provides an intuitive and tractable account of
what explanations assert in virtue of our awareness of which explananda are rendered intelligible, and gives an elegant portrayal of the origins of all of the information listed in (i) – (iv) without requiring awareness of any of it.

One other response proponents of the I-C model might consider is the claim that pragmatic considerations limit the range of relevant interventions to some graspable subset of (i). Indeed, like many others, Woodward does maintain that pragmatic considerations, broadly construed, constrain what counts as a genuine explanation for some happening. For instance, what counts as a legitimate answer to the question of why *homo sapiens* have subcutaneous fat will vary depending upon the interrogator’s interests. An appeal solely to specific genetic processes will probably not answer the evolutionary biologist’s version of the question, nor will an appeal to a particular sequence of random mutations and environmental pressures answer the geneticist’s version of the question.11 Accordingly, supporters of the I-C model might claim that in any given case interests will limit the range of relevant interventions to a far more manageable size than what I have allowed. Thus, for one interested in ontogeny, relevant interventions are not likely to include those that involve blocking early hominid migration to a particular semi-aqueous environment or raising the water temperature in said environment, but they might include those that involve knocking out the adipogamma gene, bathing that gene with gamma 1 protein in vitro, altering the consequent chemical cascade in any number of ways, and so on.

Although such pragmatic considerations do limit which interventions will be relevant in any given case, they do not limit them to the extent that would be required in order to defuse the concerns raised above. As A.I. researchers have known for
some time, there is typically no straightforward way (and seldom any way at all) to provide an explicit specification of the consequences of all possible interventions even for simple ‘toy worlds’, let alone for the vastly more complex systems that interest scientists. In other words, even granting pragmatic restrictions, the number of relevant interventions will generally still be so expansive as to raise anew the problem of which counterfactual interventions constitute the content of a given explanation, or that which it asserts in virtue of our awareness of which the event is rendered intelligible.

Nor, it is also important realize, is there anything in the appeal to pragmatics that justifies privileging members of (i) over the members of (ii) – (iv), or vice versa. Admittedly, interests will restrict which pieces of information falling under (ii) – (iv) count as relevant in any given case. However, just as with (i), pragmatic restrictions do not go far enough.

6. TACIT KNOWLEDGE OF EXPLANATORY IMPLICATIONS

Getting back to the killer asteroid explanation, on the alternative that I prefer, anyone who possesses the aforementioned explanation for the extinction should be able, as was I, to use that explanation to make inferences about both its innumerable implications and the qualifications thereto. On this view, although one might use the explanation to generate information about counterfactual dependencies, the explanation is itself something distinct from the assertion of such counterfactual dependencies. Instead, at the heart of the killer asteroid explanation lies a model of how the event actually came about, a model that one can, but need not, use to generate information regarding any of countless counterfactual circumstances and a
good deal more besides. The model is at once part of an occurrent belief about the mechanisms that produced the mass extinction and the wellspring of boundless tacit knowledge (see section 2, ftn. 10) of the implications of this belief. The information (i.e., concerning mechanisms) directly conveyed by such a model will be of sufficiently limited scope that we might 'grasp' it all and, thereby, find that an event or regularity has been rendered intelligible (see section 4.2, ftn. 9).

As for the specifics of the killer asteroid explanation, I am certainly not claiming that it is solely in virtue of imagining a large asteroid striking Earth that information concerning every one of the subsequent events automatically follows. Indeed, the current example is very much like the ones that I used in *Models and Cognition* to illustrate how an explanation can, depending upon who holds it, be fully adequate in the aforementioned sense while being deep in certain respects and shallow in others. For instance, a paleontologist who supports the killer asteroid explanation might be capable of deriving information about many of the bottom-out events and principles constituting his version of the model (see section 4.2) from his beliefs about the mechanisms behind them. This kind of information might be utilized for anticipating exceptions to those principles, for generating predictions that could not be generated solely on the basis of a superficial understanding of the model, and so forth.

The layperson’s version of the model might, on the other hand, be hopelessly shallow in certain respects. For example, most of us who imagine the high-velocity impact of a massive asteroid will automatically imagine lots of debris being flung up into the air, but that this debris should actually blot out the sun for a long period may
only be represented explicitly by our model. We might, for instance, represent it to ourselves as a mere brute happening (e.g., as an event explainable, but not explained, by a deeper model of the sustaining mechanisms), in which case this aspect of our model will be lacking in inferential productivity. In any event, once we imagine that the sun has been blotted out by this cataclysm, the rest of the explanation follows quite naturally from the knowledge that most of us share (also represented explicitly one imagines) of the basic properties of plants and animals. Having formed a model of the mechanisms that may have produced the mass extinction, even one that is shallow in the respects just mentioned, one can use that model to make countless inferences of the sort specified by (i) – (iv), including each of the ones listed above.

7. **PRECEDENT FOR THIS CRITIQUE**

In order to situate the present discussion in its wider context, let me point out some of the precedent that one may find for the above critique of the I-C model. To start with, and as explained in greater depth elsewhere, there is a closely related set of problems that plagues, directly or indirectly, the D-N model of explanation and some of its would-be successors (Waskan 2006). The problem besetting the D-N model is that even if one is able to provide a deductive-nomological formulation of a particular explanans (e.g., the killer asteroid account) that implies its target explanandum, it is often exceedingly difficult to imagine how one might provide a formulation that will also imply the countless implications of the explanans beyond the explanandum (MacCorquodale & Meehl 1948; Greenwood 1999, pp. 5) and the innumerable ways in which each implication is qualifications (Hempel 1988; Fodor 1991). Bearing in mind that our present concern is to determine what is going on, cognitively speaking,
when we have an explanation, the overarching problem facing D-N and I-C approaches is that of elucidating how we humans are able to produce boundless information of the sort described by (i) – (iv) from finite means. Succinctly put, whereas the D-N model fails to supply the finite means, the I-C model is swamped by the boundless information. There are, in addition, models such as the covering-law (Hempel 1965) and unification (Kitcher 1989) models that, like the D-N model, are plagued by the inferential limitations of extrinsic representations. The model model marks an advance over these approaches in that it gives a deep account of our boundless tacit knowledge of the information in question by appealing to the inferential productivity of ICMs of mechanisms.

There is also precedent for the charge that the I-C model improperly eschews the actual in favor of the counterfactual. Bogen (2004) and Machamer (2004), for instance, have both claimed that whether one happening causes another hinges not, as Woodward would have it, upon what would happen under (often highly idealized) counterfactual circumstances, but rather upon what actually does happen under real-world conditions. Though these two researchers focus on different facets of the issue – the truth conditions of causal claims and the ontological underpinnings of causal processes, respectively – they agree that to overlook the manner in which a cause actually brings about an effect is to forsake the very thing that makes it a causal interaction.

Although there is a great deal of overlap between our respective views, the concerns just expressed primarily concern the character of something prima facie objective (i.e., causation), whereas my present interests lie with something far more
subjective (i.e., possession of causal explanations). The intimate relationship between causation and causal explanation cannot be explored to any great depth here.\textsuperscript{12} I do hope, however, that the present essay lays sufficient groundwork for a later discussion of Woodward’s own critique of the mechanistic approach towards explanation, a discussion wherein certain facets of this relationship arise. Woodward claims that the mechanistic approach has trouble making sense of (inter alia) causal interactions between parts of mechanisms, cases where we seem to have knowledge of causes without knowledge of mechanisms, and disinhibitory influences (2002, 2004). It will be important to determine whether or not the specific psychologized version of the mechanistic approach advanced here has the resources within itself to answer these charges or if, as some have suggested, mechanistic approaches must in general incorporate central tenets of the I-C model (Woodward 2002; Glennan 2002; Psillos 2004).

8. **SYNOPSIS**

The model and the I-C model offer competing accounts, both of the information that explanations convey in virtue of our awareness of which happenings are rendered intelligible and of the origins of information that bears on the evaluation of explanations. I have tried to show here that the model model better accounts for our vast knowledge of this evaluatively relevant information owing to its analysis of the structure and content of the mental representations that lie at the heart of all explanations. On this view, occurrent beliefs about the actual mechanisms by which a happening may have been produced endow us with tacit knowledge of many kinds of evaluatively relevant information.
In all candor, I do not pretend to expect an immediate groundswell of support for the model model or even for the present account of our vast knowledge of explanatory implications. I do hope, however, that those who share my interest in the problem of accounting for this knowledge will recognize its intimate connection with the frame problem of A.I., accept that there is a well-vetted solution to the frame problem in the ICM hypothesis, and at least keep an open mind to the possibility that the problem at hand admits of a similar solution. Perhaps some of those who allow this much will also deem it worth their while to consider the extent to which the model model is able to meet the rest of the ever-growing litany of demands now placed upon models of explanation.

ACKNOWLEDGEMENTS
I thank The British Society for the Philosophy of Science, The Society for the Metaphysics of Science, and the philosophy department at Western Michigan University for allowing me to present these ideas and for the valuable feedback that I received as a result. I am also indebted to James Woodward and most of all to the two anonymous referees at International Studies in the Philosophy of Science whose thoughtful critique resulted in many improvements to this article.

REFERENCES


FODOR, J.A. (1991) You can fool some of the people all of the time, everything else being equal; hedged laws and psychological explanation, *Mind*, 100, pp. 19-34.


---

1 But see Waskan (2006).

2 At times, Woodward appears to embrace this stronger claim. For instance, when discussing the information that sometimes goes unstated when *giving* an explanation, Woodard claims, "It must be information that those who use such explanations … [not only can, but do] recognize as relevant to the import of the explanations they offer" (2003, 180). Surely to recognize the relevance of information to an explanation, one must be aware of that information, or have a conscious thought or belief involving it.

3 The general idea that cognitive models of mechanisms underlie at least some explanations has also been endorsed by Brewer (1999), Horst (2007), and Wright & Bechtel (2007). Closely related is Bechtel and Richardson’s (1993) proposal that we discover mechanisms by relying on certain cognitive heuristics.
In Waskan (2006) I consider the possibility that some explanations are constituted by the combination of ICMs and mathematical formalisms as well as such objections as that probabilistic explanations and the explanations supplied by fundamental physics do not fit the model mold.

The intrinsic-extrinsic distinction was introduced by Palmer (1978) and modified by Waskan (2003, 2006).

There are, of course, differences between scale models, InCoMs, and ICMs (Waskan 2003, 2006).

This is similar to Glennan's (1996) analysis of the metaphysics of causal relations, but see Waskan (2006, 298).

One might object that there are cases where we believe that a causal connection exists but lack knowledge of the mechanisms that enforce the connection (Woodward 2002, 2004). This issue demands more extensive treatment than can be provided here (but see Waskan 2006, 211-2, 240-3).

For now I will merely say that if we have no beliefs about underlying mechanisms whatsoever, then the information we possess is not sufficient to render the effects of such (putative) causes intelligible. In other words, I would classify such cases, which are far rarer than one might think, as non-explanatory.

Also helping us overcome memory limitations is our ability to work through, in piecemeal fashion, interconnected chains of events (Hegarty 2004).

Woodward, who was gracious enough to comment on an earlier draft of this paper, has claimed that "many of the beliefs about counterfactuals that people entertain when they are in possession of causal explanations are not occurrent, but merely [tacit]." Again, however, it cannot be in virtue of our possession of tacit beliefs that the explanandum is rendered intelligible. It must, then, be the one or few beliefs about counterfactuals that are occurrent that play this role. If I read his further comments aright, in the killer asteroid case he does settle upon just one counterfactual – namely, the one concerning what would have happened if the impact had not occurred. However, saying that this is the content of the explanation leads to further difficulties discussed below. Notice also that this response supports my contention that on Woodward’s view to possess an explanation is to have certain beliefs.
‘Version’ may be the operative word here in that there appear to be multiple distinct *explananda* in such cases. This means that we need not view interests as actually infusing the explanatory relationship itself (Woodward 2003, 230). I readily confess that *Making Things Happen* has made me more consistent on this point (see Waskan 2006, 226, 241) and has advanced my thinking on many other matters besides.

Though section 4 of Machamer (2004) does seem to be a significant step in the right direction.