REVIEW SYMPOSIUM

INVARINCE, EXPLANATION, AND UNDERSTANDING


*By Paul Humphreys*

Jim Woodward’s book *Making Things Happen* is the product of more than a decade’s thinking about issues in causation and explanation. It constitutes a model of careful philosophical analysis and I admire his book greatly. The positions he takes are never left as merely intriguing suggestions; instead, they are grounded in serious arguments and examples, in explicit definitions, and in consideration of and replies to possible counterexamples. He also has gone to great lengths to provide detailed comparisons with alternative positions. The various parts form a tightly integrated whole that makes an impressive case for an interventionist account of causation and explanation. Any serious effort to address issues in these areas will from now on have to take into account Woodward’s positions and arguments.

Rather than raising the usual sorts of disagreements with the author, a mode of philosophical discussion that can be less than productive, I want to raise one principal issue and two related but subsidiary issues that it would be useful to learn more about from the author. The main issue involves the notoriously slippery relationship between explanation and understanding. It is usual to maintain that in scientific contexts understanding comes via explanations. If one holds this view, then in virtue of having provided us with an account of scientific explanation Woodward’s book has also provided an account of scientific understanding. I want to suggest as an alternative way of reading Woodward’s book that his account of causal explanation can be reinterpreted as giving us an account of understanding
rather than of explanation. If one does so, then those of us who are committed actualists about explanation, and who find the emphasis on non-actual situations in his account of explanation more appropriate to understanding than to explanation, can bring to bear Woodward’s many insights on our own approach, despite the often deep differences in outlook. In no way am I recommending that readers should perform this kind of reinterpretation of Woodward’s theory, only that if, after having read many persuasive arguments in favour of his approach, they remain unpersuaded by his emphasis on explanatory ‘w-questions’, they can still learn much from construing such questions as a mode of increasing our understanding. I certainly did, so the first issue concerns how Woodward views the relationships between his theory of scientific causal explanation and the more general issue of scientific understanding.

Let’s start with one of the key elements of Woodward’s treatment of causation and explanation, the concept of invariance. Informally, invariance can most easily be understood in this way: Suppose that there is a relationship between two variables that is represented by a functional relationship $Y = f(X)$. If the same functional relationship $f$ holds under a range of interventions on $X$, then the relationship is invariant within that range. The intuition underlying this kind of dependence is that if we can exploit this invariance to control the effect, then we have a causal relation. Invariance lies at the heart of causation and explanation, according to Woodward: ‘‘Invariance under at least one testing intervention (on variables figuring in the generalisation) is necessary and sufficient for a generalisation to represent a causal relationship or to figure in explanations’’(p. 250).

Why is it important for the relation to be invariant? The answer lies in a second core feature – how Woodward views explanations. These are answers to ‘w-questions’ – answers to questions of the form ‘What would have happened if things had been different?’. It is not enough in Woodward’s view to know that a relationship tells us that when $C$ occurs, $E$ follows, as happens with generalisations such as ‘All ravens are black’. Sentences of the form ‘All As are Bs’, generalisations that have often been taken to represent the logical form of laws cannot, unlike the kind of functional forms given in Coulomb’s Law, provide the information needed to answer w-questions because they do not tell us what would have happened if the subject in question had not been an A. The ravens generalisation provides, according to Woodward, a shallow explanation that tells us nothing about
what would have happened if the bird in question has not been a raven, whereas the functional form embedded in Coulomb’s law allows us to answer many questions about what would have happened had the values of the initial and boundary conditions had been different (pp. 187 ff.).

An important consequence of the emphasis on w-questions is that the greater the range of invariance of the relationship that supports such counterfactuals, the better the explanation. The counterfactuals that occur in Woodward’s account of explanation thus have a different function than do similar counterfactuals in *sine qua non* analyses of causation. One of the primary roles of true counterfactuals in those analyses is to guarantee that the stated cause did indeed play a role in the effect’s occurrence. Although that motivation is also present in Woodward’s account, the ability to answer w-questions goes well beyond the ability to ensure overdetermining and pre-empting causes are absent:

According to the DN model, derivational structure matters because it involves the exhibition of a nomologically sufficient condition for the explanandum phenomenon or, more broadly, because it involves the exhibition of nomic grounds for expecting. On the contrasting view I wish to defend, derivational structure is relevant to causal explanation because it is used to show how an explanandum phenomenon would have been different if the conditions cited in an explanans had been different in various ways (p. 201).

This indicates a major difference in outlook between two approaches to explanation. Actualist approaches consider explanatory content to be contained in an account of what actually happened by way of causation in the system that contains the explanandum event or process. Actualists do not deny that by exploring the solution space of a set of equations, for example, one can come to have a deeper understanding of the system covered by those equations nor do they deny that computer simulations, in virtue of making it easy to alter the form of laws, can provide considerable theoretical understanding of how things would have been different had the form of various laws been different. But the actualist’s core commitment is to the position that explanations require truth, and that truths about what could have happened, but did not, are at most relevant to deciding whether a factor is a cause. These counterfactual truths do not play a role in the explanation itself. In contrast, for Woodward, these alternative possibilities are central to the explanatory power of an invariant generalisation.
Woodward, of course, has much to say throughout his book about the role of counterfactual situations in explanation, but I can perhaps sharpen the issue by considering a specific topic in this area. An important aspect of Woodward's position is that interventions, as characterised by his condition ‘IN’ (p. 98) are allowed to change not just values of variables that occur in regularities but can change the mechanisms that produce those regularities (p. 109).

Now, although there is an argument, addressed by Woodward, that we should not alter the mechanism on the grounds that if the mechanism is broken ‘we may be misled into thinking that there is no causal relationship between $X$ and $Y$’, this is not the reason, in my view, why we should disallow such mechanism-breaking interventions. The important objection from the actualist perspective is that if in the actual case the mechanism played an essential causal role in bringing about the explanandum, what might happen when that mechanism is not present is not explanatorily relevant.

Although most of the examples of interventions given by Woodward involve changes only in the values of variables, at one point (p. 15), he allows that the value of the gravitational force at the Earth’s surface could be changed by an intervention. This makes sense, in that we know that by going to a higher altitude, or by going to another planet, the value of $g$ will change. But although it may increase our understanding of the general theoretical framework to explore what will happen when the value of $g$ is changed in non-actual circumstances, what might happen were $g$ to change does not enter into an explanation of why this block slid down that inclined plane with the acceleration that it did under the gravitational force that actually applied.

As I understand it, there is a difference for Woodward between changes to the parameters in a functional relationship that leave the underlying functional form unchanged and interventions that, per nomological impossibility, change the form of the relationship itself. We agree that although it is a law-like relationship that opposite charges attract, considering what would happen if some hypothetical intervention occurred so that opposite charges did not attract one another is to enter a realm of no relevance to an explanation of why these two charges attracted. Here, then, is perhaps where one part of the boundary between explanation and understanding lies. Although it can enhance our scientific understanding to explore models that violate the laws of our universe, such
models cannot be used in explanations. A well-known example involves the conditions under which life can emerge in the universe. The ‘how possibly?’ questions investigated in the neighbourhood of anthropic principles add to our understanding of how life might have emerged if the laws had been different, but answers to them cannot explain life as it arose in our universe.

When considering invariant relationships, it is important, as Woodward emphasises (p. 348), that laws of nature not be taken as the standard. Woodward’s ability to explain without appealing to laws is a significant advantage for his position in that he is not committed to the kinds of top-down laws of social science that in some areas of investigation are being replaced with rule-governed individual based models. However, policies, conventions, and rules do play an important role in many of the human sciences and, unlike fundamental laws of physics, social policies and conventions, they can be changed with relative ease. Despite this difference with physical laws, it is equally important not to consider as a possible intervention something that would change a social policy when providing an explanation. For if an individual is arrested under a law, if a recession occurs because of a restrictive monetary policy, or if a woman is stoned to death because of a cultural attitude, these laws, policies, and customs must be taken as givens – they were in place when the events occurred and they play a central role in explaining them. Removing them from the scene by an intervention is as inappropriate as suspending the laws of nature.

We can perhaps see more clearly the ability to explain using policies by adapting an ingenious argument originally due to Robert Cummins and artfully deployed by Elliott Sober. Sober’s example involves a group of children who have to pass a test to demonstrate that they can read at the third grade level in order to be a student in a certain classroom. The selection process explains why all of the students in the room read at the third grade level, but it does not explain why any individual child reads at that level. Let us call any such screening policy a filter. A simple physical example is the use of an abiotic sieve to explain why all the stones in a heap have a maximum dimension of 1 inch in terms of the heap being the result of a filtering process with a sieve of mesh 1 inch. We can use filtering policies to argue that only facts about actual processes and property instances are required to fully explain the situation that results from implementing the filter.
The causal path in these cases looks like \( X \rightarrow F \rightarrow Y \). \( X(p)=1 \) interacts with \( F \) to bring about \( Y(p)=1 \). Similarly \( X(p)=0 \) interacts with \( F \) to bring about \( Y(p)=0 \). Although one could introduce a dummy variable \( F \) such that \( F(w)=1 \) if the filter is in place and \( F(w)=0 \) otherwise, it is in considering the causal interactions between the stones and the sieve mesh that we reach an explanation of why certain stones ended up in the heap and others did not. Information consisting in the size of various stones arriving at the sieve and the size of the sieve mesh is sufficient to explain the composition of the heap on the other side of the sieve. Knowledge of what would happen if the filter was not there is irrelevant simply because it is part of the very idea of a filter that it acts to selectively influence objects that encounter it.

Consider a security screening policy at an airport. Do we need to know what would happen if the policy were to be changed or removed entirely in order to explain why no security related events occurred on flights from that airport? Not if we know the properties of the individuals subjected to the policy and what interactions happened at the filter as a result of the policy. There is, of course, one crucial difference between social policies of this kind and traditional laws of nature that is relevant. When rational agents are subject to rules, policies, and strategies, they will generally consider the range of possibilities open to them in a way that arational entities cannot, and they will often adapt their behaviour to exploit the policy and to gain an advantage. Yet this is again something that is relevant to rational agents’ general understanding of the system and not to an explanation of the outcomes in a particular instance.

And so my two subsidiary questions to Jim Woodward are the following. Can social policies and rules play a similar role in explanations as do the usually more invariant mechanisms of the natural sciences? Why are actualist explanations in filtering cases of the above kind not fully adequate as explanations, even though they are able to answer very few w-questions? I look forward to reading his replies as much as I enjoyed reading his book.

Corcoran Department of Philosophy
University of Virginia,
Virginia, USA
By Elliot Sober

This is a wonderful book. There are two main ideas and lots of interesting secondary themes. The first main idea is a non-reductive account of causation that is inspired by Pearl (2000) and Spirtes et al. (2000). The second is an account of causal explanation in which the idea of invariance is developed to account for what makes generalisations explanatory and why one explanation is deeper than another. With respect to the lesser themes, let me mention Woodward’s discussion of directed graphs, his critique of Lewis’s counterfactual theory of causation and of the Mill/Ramsey/Lewis theory of law. There also is valuable discussion of the theories of explanation with which Woodward disagrees. Woodward is right to bemoan the fact that philosophical work on causation has become so ‘Balkanised’. There are metaphysicians who think about counterfactual accounts and philosophers of science who think about probabilistic accounts, structural equations, and directed graphs. These two groups should be learning from each other, but there isn’t enough of that going on. In addition, economists and other scientists have developed a variety of ideas about causation that have considerable philosophical interest. Woodward’s book performs the valuable service of showing how these ideas are related. Furthermore, it isn’t just philosophers working on causation and explanation who ought to read this book. Philosophers who use these notions in their own work – philosophers of mind and philosophers of biology, for example – have much to learn from him.

Woodward provides a manipulationist theory of causation. He is entirely right in thinking that this approach can and should be stripped of its anthropomorphic associations. Intervention is the key idea, and can be expressed as follows:

\[ (*) \quad X \text{ causes } Y \text{ if and only if there is a possible intervention on } X \text{ that would be associated with a change in } Y \]

(or with a change in the probability of \( Y \)).

Intervention is a causal notion: an intervention on the variable \( X \) causes \( X \) to take on a particular value. In addition, if \( I \) is an intervention on \( X \) with respect to \( Y \), then, if \( I \) affects \( Y \), it does so only via its influence on \( X \) (p. 98); there is no other causal pathway from \( I \) to \( Y \). There are additional causal conditions that define
what an intervention is. Interventions, as Woodward defines them, have to be very delicate (p. 130). An intervention on $X$ sets the state of $X$ without simultaneously altering lots of other stuff. Thus, you have to know a lot about the causal situation to know whether or not something counts as an intervention on $X$ with respect to $Y$. And so you must have lots of other causal knowledge to figure out whether $X$ causes $Y$. Since the right-hand side of (*) uses causal concepts, the account is not reductive. However, as Woodward notes, it is not circular.

Woodward discusses the distinction between causal generalisations that describe types of events and singular causal statements that describe token events. Consider, for example, the following two statements:

\[
\begin{align*}
\text{(type) } & \text{Short circuits cause electrical fires.} \\
\text{(token) } & \text{This short circuit caused this electrical fire.}
\end{align*}
\]

For Woodward, the type-level claim is true because if oxygen were present in a system, intervening on whether there is a short circuit would be associated with a change in whether there is a fire (or a change in the probability of a fire). The type-level claim could be true even if short circuits never occur and neither do electrical fires. The token-level claim is different. There has to be a short circuit and a fire. Moreover, the claim won’t be true unless oxygen is actually present. If it is not, there is no counterfactual dependence. Woodward denies that these differences between the two statements show that there are two concepts of cause. His view is that the type-level claim describes a possible cause of fires, while the token-level claim describes an actual cause of a fire, with ‘cause’ understood univocally (p. 76).

I once defended the idea that type-level and token-level claims involve different concepts of cause (Eells and Sober 1983; Sober 1985). I did so because I was thinking of the type-level claim from the point of view of a probabilistic theory of causation, and it seemed to me that token events are sometimes related causally even though the probabilistic requirement is not satisfied. The basic idea of a probabilistic theory is that a positive (negative) causal factor must raise (lower) the probability of its effect when other facts about the situation are ‘held fixed’. (The point of this last requirement is to distinguish cause from mere correlation.) I was struck by cases in which a token event $C$ doesn’t raise (or even change) the probability of $E$, but $E$ still
traces back to C. For example, consider a randomly mating population in which the parental generation has two alleles (A and a) at a given locus, each with a frequency of 0.5. This means that there is a prior probability of 0.5 that an individual in the offspring generation will be an Aa heterozygote. Now consider one of these heterozygote offspring, O, and suppose that both of O’s parents were themselves heterozygotes. Then the posterior probability that O will be Aa is also 0.5. That is, \( \Pr(O \text{ is } Aa) = \Pr(O \text{ is } Aa | O's \text{ parents are both } Aa) = 0.5 \). So the parents’ genotype didn’t change the probability that the offspring would be Aa. It therefore appears that the genotype of the parental pair isn’t a type-level cause of the offspring genotype. But still, the offspring’s genotype traces back to the genotypes of the two parents, so this seems to be a case in which there is token-causation without type-causation.

After reading Woodward’s book, I find this example and others of its ilk much less persuasive. Woodward emphasises, as does Hitchcock (1993), that multiple comparisons of probabilities need to be made. In the present example, \( \Pr(O \text{ is } Aa | O's \text{ parents are both } Aa) = \Pr(O \text{ is } Aa | O's \text{ parents are not both } Aa) = 0.5 \). So if we consider just two states of the parental pair – both parents are heterozygotes versus the condition that at least one of them is not – it appears that the parents make no difference. However, other, more fine-grained, comparisons also are possible. For example, if both parents were AA, or both were aa, the probability of O’s being a heterozygote would be zero. So there is a counterfactual dependence between parental and offspring genotype in the sense that there are some changes in the parental genotype that would be associated with a change in the probability of the offspring’s genotype. This is a point that can be accommodated within a probabilistic theory of causation if the theory is not restricted to comparisons of ‘C occurs’ and ‘C does not occur’. Woodward’s manipulationist account handles this case with ease.

Although Woodward wants to offer a unified account of type- and token-level causal claims, he still maintains that there is a difference between the two. This comes up in his discussion of Lewis’s theory of singular causal claims (pp. 213–214, 373). Lewis holds that ‘Harry’s smoking caused Harry’s cancer’ has two subjunctive implications. (1) If Harry had not smoked, he would not have got cancer, and (2) if Harry were to smoke, he would get cancer. Woodward accepts the first of these, but not the second. For
Woodward, token causes are necessary in the circumstances for their token effects, but they need not be sufficient. He rejects the requirement of sufficiency because he recognises that causation can occur in situations that are irreducibly chancy. I agree with Woodward about this, but it seems to me that the same considerations should lead him to reject the requirement of necessity. My suggestion is to bring Woodward’s accounts of type- and token-level causal claims closer together by abandoning the claim that token causes are necessary in the circumstances for their token effects. If Harry hadn’t smoked, the cancer could still have arisen, by chance.

Invariance is the key idea in Woodward’s account of explanation; and some of the ingredients are already present in his account of causation. To see why, let’s examine how (*) handles associations between events that are due to their having a common cause. Consider the much-cited barometer whose readings on one day are correlated with the occurrence of storms on the next. The correlation is due to the fact that barometric pressure causes both. There are two generalisations that describe this system, but only one of them is causal. Suppose that ‘low barometer reading on day i iff storm on day i + 1’ is true, and that ‘low barometric pressure on day i iff storm on day i + 1’ is too. The difference between these generalisations is revealed by intervention. If we intervene on the barometer, the barometer/storm generalisation changes from true to false. But if we intervene on the barometric pressure, the pressure/storm generalisation stays true. The causal generalisation is invariant (it stays true) under interventions, but the generalisation that relates the two effects of a common cause is not. A generalisation must be invariant under some interventions if it is to be causal. Woodward declines to address the question of whether all explanations of singular occurrences must be causal, but says that his account of explanation will be limited to causal explanation (p. 189). Given this, the idea that a generalisation must be invariant under some interventions if it is to be explanatory simply follows from his account of causation.

Woodward has more to say about invariance and this is where his views on explanation go beyond what he says about causation. For example, he argues that greater invariance is a source of greater explanatory depth. The rough idea is that explanations answer ‘what if things had been different?’ questions and that explanations are better the greater the range of such questions that they
correctly answer. Woodward discusses an idea from the econometrician Tyge Haavelmo (1944), who contrasts two generalisations about the performance of a car (p. 258). Suppose the car is tested on a highway that is flat and straight. From repeated trials, you can identify a functional relationship between pressure on the gas pedal and the car’s top speed. The second generalisation involves a more complex engineering analysis of the car’s components, including the gas pedal but also various other components, as well as the effect of the highway surface.

\[
\text{Top speed} = f(\text{pressure on the gas pedal})
\]

\[
\text{Top speed} = g(\text{pressure on gas pedal, properties of other car components, properties of the highway surface})
\]

Haavelmo says that the second, more complex model, has greater ‘autonomy’. In Woodward’s terminology, it is more invariant. The first model ceases to be true if you take the car to a new highway or instal a new type of spark plug. The second model remains true under those changes. Woodward emphasises that the changes considered here are not restricted to interventions on the input variable that the models share. The first model, he says, does not mention the highway surface or the kind of spark plug. However, if this model is not taken to be restricted to a certain type of highway surface, it presumably ranges over all such surfaces, and therefore is false. In any event, Woodward separately argues that a generalisation need not be true to be invariant and explanatory (p. 304). For example, if two models are both false because both involve idealisations, but both are good approximations in some appropriate sense, both can be explanatory, and the one that is more invariant will be the deeper explanation. In this case, what is invariant is that the model remains a good approximation under various interventions and changes.

Woodward provides several persuasive examples in which degree of invariance seems to track degree of explanatory depth (pp. 260–261). Van der Waals’ force law is deeper and more invariant than the ideal gas law. Van der Waals’ law holds in any circumstance that the ideal gas law holds, but not conversely. Similar relationships obtain between general relativity and Newtonian gravitational theory and between Newtonian gravitational theory and Galileo’s law of free fall. Despite these examples, I am sceptical that greater
invariance is a source of greater explanatory depth. A different type of example illustrates why.

Consider two quantitative variables, each of which causally influences the state of another. For example, how tall a corn plant of a given genotype grows is influenced by the amount of water and the amount of fertiliser it receives. We might study the influence of these two factors by cloning the plant and putting the clones in a variety of different treatments, each characterised by a given amount of water and a given quantity of fertiliser. At the end of some period of time, say sixty days, we measure the plants and calculate the average height in each treatment group. I want to consider two models, each of which tells part of the story. The water model says height = f(water) + E and the fertiliser model says height = g(fertiliser) + E. Since each model fails to describe at least one causally relevant variable, each includes an error term E. Now suppose that the water model correctly describes the fact that water has a big effect on height, while the fertiliser model correctly describes the fact that fertiliser has a very modest effect on height. This comparison is made by considering the effect on height of a standard deviation’s change in water with the effect of height of a standard deviation’s change in fertiliser.

Let’s suppose that our study is carried out with all plants experiencing the same range of temperature. Call this domain D. Suppose that water has a large effect on height and that fertiliser has a small effect, not just in the domain D, but in lots of other temperature ranges as well. You can visualise this by imagining, for each temperature domain, one regression line that represents the effect of water on height and another that represents the effect of fertiliser on height. In each domain, the line for water has a steep positive slope, while the line for fertiliser has a very modest positive slope. But now let me add a further detail. Suppose the fertiliser lines for different domains always have exactly the same modest slope, but the water line’s slope changes a bit from domain to domain. It is always steep, but sometimes it is steeper than others. Woodward’s idea about invariance seems to entail that the fertiliser model for the population in domain D is deeper than the water model because the fertiliser model is more invariant. This doesn’t sound right to me. I feel confident in saying that the water model is the better explanation, because it describes the cause that has the stronger effect. Whether it is deeper I do not know. But it seems clear to me that it isn’t less deep.
In the favourable examples that Woodward describes in support of his claim that invariance is a source of explanatory depth, the deeper generalisation explains why the shallower generalisation is true or approximately true in the circumstances in which this is so. Perhaps this is what makes for greater explanatory depth in these examples. But the idea that one generalisation explains another is not easily captured by a theory of causal explanation in which the focus is on explanations of singular occurrences. Woodward is right that there is something good about generalisations that remain true or approximately true across a range of circumstances. We often want our tools to be robust, not fragile. But whether this virtue of a generalisation is what makes an explanation deep, when the generalisation is applied to a specific example, is another matter, and it is here that I am sceptical.

Another element in Woodward’s account of explanation is the thesis, also defended by Garfinkel (1981) and Putnam (1975) as the basis for their argument against reductionism, that a macro-event at \( t_2 \) is explained by macro-information about the system at \( t_1 \), not by micro-information about \( t_1 \) (231–233), if the macro-state at \( t_1 \) can be realised by multiple micro-states. Woodward illustrates his point by describing a chamber that contains a mole of a gas; its pressure declines from the value \( P \) to \( P' \) when the volume is increased while the temperature is held constant. Why does the pressure change in this way? A macro-story might cite the values of the macro-properties just mentioned and the ideal gas law. A micro-story might describe the positions and momenta of the \( 6 \times 10^{23} \) molecules and how each impacts the walls of the container. Woodward says that the micro-story ‘would fail to provide the explanation for the macroscopic behaviour of the gas’ because even if the molecules had a different set of positions and momenta (consistent with the container’s temperature, pressure, and volume at \( t_1 \)), the pressure still would have gone to \( P' \) at \( t_2 \) (except if the positions and momenta had fallen in a set of measure zero).

I see three problems here. First, it is not correct that the system’s micro-state made no difference; there exist changes in micro-state that would have prevented the pressure change, just as there are changes in macro-state that would have done so. The macro-facts at \( t_1 \) screen off the micro-facts at \( t_1 \) from the macro-state at \( t_2 \) (except for the set of measure zero), but why does that show that the micro-facts are not explanatory? Second, the requirement
that the *explanans* contain only information that is necessary for the *explanandum*. Leads to highly uninformative disjunctive explanations. Why did Moriarty die? Holmes' shooting him is not the explanation, according to Woodward; rather, the explanation is simply that a certain disjunction is true, where each disjunct describes a possible cause of Moriarty's death and says that it occurred. Third, the claim that explanatory factors are necessary for their effects can clash with Woodward's other thesis that generalisations provide deeper explanations the more invariant they are. For example, what should we say if the theory of ideal gases is less invariant than the theory about the individual molecules? Is the micro-story better because it is more invariant or worse because it cites factors that aren't necessary for the macro-effect? In discussing Friedman (1974) and Kitcher's (1989) unificationist account of explanation, Woodward criticises their winner-take-all approach, according to which there is just one story that counts as *the* explanation of a given occurrence (p. 367); Woodward prefers a greater pluralism, one that grants that different stories can all be explanations. In this light, it is odd that he says, of the micro-account of the change in pressure, that it is not an explanation. Shouldn't the pluralist recognise that generality and detail are both explanatory virtues, that they conflict, and that the trade-off we regard as optimal for a particular explanatory problem is simply a function of our interests (Jackson and Pettit 1992; Sober 1999)?

(I extend my thanks to Daniel Hausman and Jim Woodward for helpful comments on previous drafts of this review.)

**REFERENCES**


Department of Philosophy
University of Wisconsin, Madison
WI, USA

Author’s response

By James Woodward

I am very grateful to both Paul Humphreys and Elliott Sober for their generous, thoughtful, and constructive discussions of *MTH*. They raise a large number of interesting questions and problems. For reasons of space, I am able to focus on only a few of these in what follows.

1.

As Humphreys observes, my account of explanation assigns a central role to counterfactual information about how the explanandum phenomenon would have been different if the conditions cited in its explanans had been different in various ways. In contrast, Humphreys urges that counterfactual information is relevant to understanding, but not to explanation. Instead, Humphreys favours an ‘actualist’ account of explanation, according to which “explanatory content [is] contained in an account of what happened by way of causation in the system that contains the explanandum event or process” and “counterfactual truths do not play a role in the explanation itself”. One of the examples that he uses to motivate these
claims involves the use of an abiotic sieve “to explain why all the stones in a heap have a maximum dimension of one inch in terms of the heap being the result of a filtering process with a sieve of mesh one inch”. Humphreys contends that:

Information consisting in the size of various stones arriving at the sieve and the size of the sieve mesh is sufficient to explain the composition of the heap on the other side of the sieve. Knowledge of what would happen if the filter was not there is irrelevant simply because it is part of the very idea of a filter that it acts to selectively influence objects that encounter it.

Humphreys’ actualist approach to explanation reflects a fundamental split between, on the one hand, those (e.g. Salmon 1984; Dowe 2000) who think that explanation (and perhaps causation as well) has to do just with what actually happens and those, like me, who think that causal and explanatory claims must be understood (at least in part) in terms of the counterfactual commitments that they carry. I cannot possibly settle this deep issue in the space allotted here. Instead I shall confine myself to reminding the reader of some of the considerations that led me in MTH to favour the counterfactualist approach.

First, consider Humphreys’s sieve example. On my view, in describing this device as a sieve or filter that successfully sorts objects according to whether or not they fit through the mesh, we commit ourselves to a number of counterfactuals about what would happen if, contrary to actual fact, the sieve were removed or its dimensions or various other characteristics were changed. Moreover, these counterfactuals seem highly relevant to the explanation of how (and even if) the sieve works. To take the simplest possibility, suppose we were to remove the sieve entirely from any contact with the stones (we dump them directly on the ground). If, under these conditions, we still find ourselves with a pile of stones all of which are less than one inch in dimension, we would be justified in concluding that whatever the explanation for the sizes of the stones in the pile, it did not have to do with the action of the sieve. Similarly if the sieve works ‘as advertised’, it is crucial to the explanation of its sorting ability that the mesh be not only of the appropriate dimension but that it also be relatively rigid. Suppose that if stones that are two inches or more in dimension were introduced into the one-inch mesh, they would readily pass through (because the mesh is highly flexible and elastic or tears easily). Then it is not just the
fact that the mesh has one-inch dimensions that is explanatorily
relevant to which stones get through. In general, then, to say
that it is the action of the sieve that explains the size of the
stones in the pile is to commit oneself to a set of counterfactual
claims about what the sizes of stones in the pile would be if the
characteristics of the sieve were to be varied or if it were to be
removed.

This example, as well as a number of others discussed in MTH,
illustrate a more general motivation for adopting a counterfactual ra-
ther than a purely ‘actualist’ account of explanation: any acceptable
type of causal explanation needs an account of causal and explana-
tory relevance. Moreover, it seems uncontroversial that relevance in
this sense is a notion that has to do with explanation, rather than
with some broader category of understanding. I claim that appealing
to counterfactuals provides a natural and straightforward way of
providing such an account. In simple structures of the sort presently
under consideration, 1 X is explanatorily relevant to Y if and only if
the value of Y would change under some interventions that change
the value of X. We spell out how in detail X is causally or explanato-
rily relevant to Y by spelling out in detail how the value of Y would
be different under different values of X. Thus, in Salmon’s well-
known example, the ingestion of birth control pills by the male
Mr Jones is causally and explanatorily irrelevant to his subsequent
failure to get pregnant, precisely because there are no changes in
whether Jones takes birth control pills that will change whether he
becomes pregnant. In contrast, it is hard to see how to capture the
notion of relevance in purely actualist terms, without appeal to coun-
terfactuals. We cannot do this by appealing to facts about spatio-
temporally continuous processes and conserved quantities, for
reasons described in MTH, pp. 350 ff. Nor can we do this by appeal-
ing to facts about conditional probabilities (MTH, p. 357). Of
course, it is possible that there is some alternative way of capturing
the notion of relevance while remaining within the confines of an ac-
tualist account, but until we are presented with such an alternative, I
think that it is reasonable to take counterfactual information to be
relevant to explanation (and not just understanding) in the way de-
scribed in MTH.

In related remarks, Humphreys resists my suggestion that it is
relevant to consider what would happen if mechanisms connecting
causes and effects were to change, in characterising the notion of
an intervention and hence in the characterisation of casual and explanatory claims. He writes that we should “disallow mechanism changing interventions” and that:

the important objection from the actualist perspective is that if in the actual case the mechanism played an essential causal role in bringing about the explanandum, what might happen when that mechanism is not present is not explanatorily relevant.

In thinking about the idea of a mechanism changing intervention, it is useful to distinguish several different possibilities. First, there are cases in which by intervening on some variable \( X \), we break or disrupt the relationship between \( X \) and its previous causes, putting \( X \) entirely under the control of the intervention. Here what is changed or disrupted is the mechanism in which \( X \) figures as a dependent variable. Among the motivations for allowing this sort of possibility are (a) that it provides a natural way of modelling many real-life experimental manipulations; and (b) that interventions that do this are particularly normatively desirable since putting \( X \) entirely under the control of the intervention eliminates the possibility of certain kinds of confounding (cf. *MTH*, pp. 95 ff.)

As an illustration of both points, consider first a situation in which people in population \( P \) decide voluntarily on the basis of unknown causes whether or not to take a drug. Differences in the rate of recovery from some disease among drug takers and non-drug takers might reflect the efficacy of the drug but they might also be entirely due to other differences among these two subpopulations – perhaps having a stronger immune system (measured by \( S \)) causes one to be more likely to (\( D \)) take the drug and \( S \) also causes the differential rates of recovery (\( R \)). As explained in *MTH*, if we conduct a randomised trial in which subjects are randomly assigned either to a group that receives the drug or to a group that does not and if we then compare rates of recovery across the two groups, we effectively rule out the possibility of this sort of confounding. One way to think of the randomisation is that it destroys the previously existing causal relationship (breaks the mechanism) between \( S \) and whether one takes the drug, putting who gets the drug entirely under the control of the randomisation process. At least in this sort of case, in contrast to Humphreys’ claim in the passage quoted above, it is highly relevant to the explanation of recovery to consider what would happen to the recovery rate under conditions in
which the previously existing mechanism linking $S$ to whether one takes the drug is changed or disrupted.

Suppose that we are interested in the causal or explanatory relationship (if any) between $X$ and $Y$. Is it ever appropriate to consider interventions on $X$ that change or disrupt the mechanism linking $X$ to $Y$? (Note that here, unlike the previous case, our focus is on changing the mechanism in which $X$, the variable intervened on, is an independent variable.) In exploring this question, we need to recognise a point that is discussed in more detail in *MTH*, Chapter 3: the notion of changing or disrupting a mechanism always must be understood as relative to a particular characterisation of the mechanism and the generalisation governing it. What looks like changing a mechanism described by one generalisation can always be re-described as instantiating a new set of values for some more general mechanism, described by some broader generalisation, with the latter being unchanged. Suppose that we have a spring that conforms to Hooke’s law \( F = -kX \) within a certain range of extensions but that we stretch it to a length \( x^* \) at which the restoring force is no longer linear, although not so much as to break the spring. In doing so, have we changed or disrupted the mechanism linking $F$ to $X$? If that mechanism is just what is described by \( (H) \), then it is changed. If instead we describe the mechanism between $F$ and $X$ by some more general and complex function \( F = G(X) \), which captures both the range of values of $X$ for which the relationship between $F$ and $X$ is linear and the non-linear relationship between $F$ and $X$ that holds for other values of $X$, then that mechanism and the generalisation $F = G(X)$ is not changed by extending the spring to $X = x^*$. A similar point holds even if extend the spring so far that it breaks. For this reason, among others, I would resist Humphreys’ suggestion that there is a fundamental distinction between interventions that change mechanisms and those that do not and that what happens under the former interventions are irrelevant to successful explanation, although possibly relevant to understanding.

One potential source of confusion here concerns the notion of ‘changing’ not just mechanisms like the one described by \( (H) \) but (more radically) laws of nature. It is standard practice in philosophy – a practice that I criticise in *MTH* – to describe both generalisations like Maxwell’s equations or the field equations of General Relativity and low-level local generalisations like Galileo’s law of freely falling.
bodies or Hooke’s law conceived as describing a particular kind of spring as ‘laws’. I agree with Humphreys when he writes that “if it is a lawlike relationship that opposite charges attract, considering what would happen if some hypothetical intervention occurred so that opposite charges do not attract is to enter a realm of no relevance to the explanation of why these two charges attracted”. But the reason why this is irrelevant, on my view, is that as best we know there is no intervention that will change the actual situation in which opposite charges attract to a situation in which opposite charges do not attract. It is not just that we lack the technological power to do this; rather, the laws of electromagnetism say there is no possibility of doing this and they certainly have nothing to say about what would happen if such a possibility were somehow to be realised. When I say in MTH that good explanations should provide counterfactual information about what would happen to their explananada under interventions that change the values of the variables in their explanans, I mean information about what would really in fact happen, as an empirical matter (where this information might be provided by physics or some other relevant science or by experimental manipulation), under such interventions. We either already know, or often at least can find out experimentally, what would happen under interventions that change the mesh of a sieve or the extension of a spring. There is thus, from my point of view, a principled reason why the latter sort of ‘what if’ information is explanatorily relevant while the ‘what if’ question posed by Humphreys in the electromagnetism case is not.

Humphreys also links his remarks about the explanatory irrelevance of mechanism-changing interventions to a set of claims about the explanatory power of appeals to norms, rules, and conventions in connection with social phenomena. He is impressed by the explanatory credentials of such appeals and also holds that in circumstances in which such norms, etc., are in place, information about what would have happened if they had not been in place is explanatorily irrelevant. He writes:

Policies, conventions, and rules do play an important role in many of the human sciences and, unlike fundamental laws of physics, social policies and conventions, they can be changed with relative ease. Despite this difference with physical laws, it is equally important not to consider as a possible intervention something that would change a social policy when providing an explanation. For if an individual is arrested under a law, if a recession occurs because of a restrictive monetary policy, or if a woman is stoned to death because of a cultural
attitude, these laws, policies, and customs must be taken as givens – they were in place when the events occurred and they play a central role in explaining them. Removing them from the scene by an intervention is as inappropriate as suspending the laws of nature.

While I do not claim that appeals to norms are always explanatorily empty, I hold, in contrast to Humphreys, that such appeals, when unaccompanied by information about the conditions under which the rules in question (and the behaviour they govern) would have been different, are explanatorily shallow. To provide any very satisfying explanation of the sorts of social phenomena Humphreys describes, it is not enough just to exhibit them as instances of generally observed rules, precisely because doing this fails to provide relevant ‘what-if-things-had-been-different’ information and because such rules tend to be relatively non-invariant. I think this assessment conforms to the judgments of many methodologically sophisticated social scientists and that it is a virtue, rather than a defect, in my account of explanation, that it reflects this. Within contemporary social science, a common way of providing such ‘what-if-things-had-been-different’ information involves the exhibition of social phenomena and the rules of which they are instances as equilibria in some underlying game. When explanatory, an account of this sort will also give us information about the conditions under which there would be different equilibria (different rules) – thus satisfying my ‘what-if-things-had-been-different’ condition. More specifically, accounts of this sort will provide a variety of different kinds of explanatorily relevant information. They both show us what the individual preferences and beliefs of the actors are, and, given the strategic structure of the situation they face and the pay-offs associated with different courses of action, how this leads to some observed bit of behaviour (which may or may not conform to some generally accepted rule) and how if those beliefs, or preferences, or the strategic structure of the game and the pay-offs associated with it were to change in various ways, this behaviour/rule would change.

To take a simple example, it seems to me that it is not a very satisfying explanation of why people drive on the left hand side of the road in England simply to say that there is a rule, $L$, to this effect and people follow it. We get more insight into people’s behaviour if we see them instead as involved in something close to a pure co-ordination game. Assume that the players are (to a good approximation) initially (that is, prior to the adoption of
any convention) indifferent about which side of the road they drive on but recognise that it is very much in their interest to drive on the same side as everyone else. Once a convention about driving on the left is established, no one has an incentive to deviate from this choice and so this equilibrium is sustained – it is a Nash equilibrium. If instead driving on the right had been established as a convention, a similar analysis would show why that would be sustained as a coordination equilibrium. This analysis relies at a number of points on information about what would happen if the left-hand-side rule were violated – in describing what would happen if an individual driver were to unilaterally deviate from the rule and in describing what would happen if a right-hand-driving rule had been initially established. Indeed the whole notion of a Nash equilibrium – the central solution concept in both competitive and non-competitive games – appeals to such counterfactual information, since it has to do with what the payoff to each agent would be if that agent were to unilaterally adopt some alternative strategy, given the strategies chosen by the other players. Note also that, among other things, simply saying that the drivers follow the rule, \( L \), leaves it open whether they do so because they have some intrinsic preference for the left-hand side of the road or for some other reason. The analysis described above addresses this issue, among others.

As this example illustrates, in order to explain why people follow the rules (or exhibit the behaviour) they do in one set of circumstances, it is highly relevant to consider the (different) rules they would follow (or the different behaviour they would exhibit) in other circumstances. Among other things, social norms and rules are often pretty unstable – people vary in the extent to which they conform to them even when the rules are in some sense “in effect” and the norms themselves often change in response to changes in other factors. Grasping such larger patterns of dependence can be highly relevant to explaining why some particular rule is obeyed in a particular case.\(^2\)

2.

Turning now to Elliott Sober’s remarks, I want to focus on two aspects of his discussion: the connection between invariance and
explanatory depth and his remarks about explanation by macrotheories.

Suppose that both fertiliser and water affect plant height but that fertiliser has a modest but constant effect on height across a range of different circumstances while water has a larger but less stable effect on height across different circumstances. Sober worries that it follows from my view that a model of plant height that includes only fertiliser provides a deeper explanation of plant height (because the fertiliser/height relationship is more stable or invariant) than a model that appeals only to water. In contrast, Sober judges “the water model is the better explanation because it describes the cause that has the stronger effect”. Although I have reservations (see below) about Sober’s use of this example, I fully agree with the more general point behind these remarks. As I explicitly say in *MTH*, both ‘explanatory depth’ and ‘degree of invariance’ are “complicated and multidimensional notions” (p. 265). In saying that greater invariance was a source of explanatory depth, I did not mean to imply that this was the only consideration relevant to assessing the depth or goodness of an explanation, or that it always outweighed in importance every other consideration.

This having been said, I also must add that I have reservations about the notion of ‘causal strength’ as it figures in Sober’s example and his use of this notion to motivate the claim that the water model provides a ‘better explanation’ than the fertiliser model. The central problem is that the conclusion that the effect of water on plant height is ‘larger’ than the effect of ‘fertiliser’ (or that water is the ‘stronger’ cause) requires that we have some principled common basis on which to compare the per-unit change in height associated with each of these variables and it is hard to see how to choose this basis in a non-arbitrary way. Suppose that when both water and fertiliser are measured in grams, regressing height on water yields a larger coefficient than regressing height on fertiliser. If we instead measured water in kilograms and fertiliser in micrograms this relationship might well be reversed. In other words, assuming that fertiliser is measured in so-called non-standardised units (that is units like grams rather than units that are normalised to the standard deviation of the fertiliser variable – see below), whether regressing height against fertiliser yields a linear relationship with a very modest slope or a steep one will depend on the
units in which fertiliser is measured. Similarly for water. This problem becomes, if anything, even more acute in the common case in which the variables whose importance is to be compared cannot be measured in the same unit. Suppose one regresses a subject’s income $I$ measured in dollars on father’s income $F$ and a subject’s years of schooling $E$, measured in dollars. How do we go about comparing the size of the effects of these two variables on $I$? For these reasons, steepness of slope of the regression line as measured in non-standardised units looks like a non-starter as a measure of how well one variable explains another, or of ‘causal strength’.

A common strategy in the causal modelling literature for dealing with this difficulty is to use standardised variables – that is, to make the comparison in standard deviation units for each variable. The assumption is that this provides the needed ‘common currency’ for meaningful comparisons of effect size. However, it is well known that this strategy has its own difficulties – in particular, it has the consequence that the ‘strength’ of a cause or the ‘size’ of an effect attributable to a cause depends on the actual range of variation in the variables being measured. Interestingly, considerations of invariance can be used to show the limitations of this strategy – thus further illustrating the utility of this notion in explanatory assessment. Suppose that there is a fixed relationship between height, water, and fertiliser, as reflected in the regression coefficients in the regression equation relating these variables with non-standardised units. If we switch to standardised units and measure the size of effects (or the strength of causes) by comparing the regression coefficients with these standardised variables, then whether water is a stronger cause of height than fertiliser will depend on whether the population we are looking at is one in which there is a great deal of variation in the amount of water that individual plants receive and little variation in the amount of fertiliser, or vice-versa. Now in one sense this consequence seems unproblematic – as long as we take the explanatory task to be that of explaining the variation in $Y$ in a particular population and as long as we are clear that by causal strength all that we mean is something like how much of the actual variation in $Y$ is attributable to the actual variation in $X$.

In other respects, however, this way of measuring causal strength (and even more so explanatory goodness) seems poten-
tially misleading – exactly because it is population specific in the way described. Measured in this way, the causal strength with which water and fertiliser affect height is not, so to speak, an intrinsic feature of the causal relationships between these variables, which is stable across different populations. Instead it depends on accidental and extrinsic facts having to do with the distribution the values of these variables happen to take in different populations, even though the underlying causal relationships, as reflected in the non-standardised regression equation, are exactly the same across different populations. It is in part considerations of just this sort that lead many researchers to object to other well-known population-specific measures of ‘explanatory power’ such as $r^2$ or to the use of heritability coefficients to measure the strength of genetic relationships. The intuition underlying such objections can be viewed as an intuition about the role of invariance in explanatory assessment: measures of explanatory strength like $r^2$ or the comparative size of standardised regression coefficients are not invariant across changes in populations, changes in the variance of the independent variables, etc.; and a proper measure of explanatory power should be so invariant. Viewed in this way, it is not at all surprising that an approach like mine that links explanatory depth and invariance, and measures of explanatory goodness in terms of standardised regression coefficients should lead to conflicting results.

Let me conclude by emphasising again that I do not object to the use of standardised regression coefficients and measures like $r^2$ to assess ‘explanatory goodness’ in one sense of that notion, as long as we are clear about just what it is that we are measuring and as long as we are clear about the limitations of such measures. Sober is quite right to claim that there are notions of (or dimensions of) explanatory goodness (including whatever is measured by $r^2$) that are independent of my notion of invariance.

Sober is also sceptical of some of my claims about the role of macro-properties in explanation. Suppose that a mole of gas is confined to a container of volume $V_1$ and exerts pressure $P_1$ at time $t_1$. The gas is then allowed to expand isothermally to a new equilibrium state at which its volume is $V_2$ and its pressure is $P_2$ at time $t_2$. We want to explain why the new pressure is $P_2$. Consider first the following strategy for doing this: We establish the exact position and momentum of each of the $6\times10^{23}$ molecules
comprising the gas (call this $S_1$) and then derive the position and momentum of each molecule (call this $S_2$) at the new equilibrium state, using the laws of Newtonian mechanics and the force laws governing each of the molecular collisions leading to $S_2$. We then note that $S_2$ is sufficient for the new pressure $P_2$. Call this the micro-explanation for $P_2$. Of course it is completely impossible to actually construct such a micro-explanation but I claimed in MTH that, even if one could, “there are important respects in which it would fail to provide the explanation of the macroscopic behaviour of the gas we are looking for”. My reasoning did not appeal to any general preference for macro-explanations over micro-explanations when macro-states can be multiply realised, but rather to the ‘what-if-things-had-been-different’ conception of explanation. When we ask for an explanation of the new pressure $P_2$, we are interested not just in exhibiting a nomologically sufficient condition for this explanandum but also a specification under which this explanandum would have been different. As described, the micro-explanation fails to provide this information. In particular, it fails to tell us under what changes in the initial micro-state of the gas, the final pressure would have been the same as $P_2$ and under what changes in the initial micro-state the final pressure would have been different. By contrast, a derivation of the value of $P_2$ using the ideal gas law, and the initial volume, temperature, and pressure does provide this information: it shows us how, if these macroscopic variables had been different, a different pressure $P_2^* \neq P_2$ would have resulted.

I did not claim in MTH that no micro-account could provide this information. For example, a derivation that included a specification of all of the micro-states associated with $P_2$ and of the various micro-states under which alternatives to $P_2$ would be realised, would (modulo the above qualifications about calculational intractability etc.), provide an explanation of why $P_2$ rather than these alternatives were realised. My point was that just tracing the micro-evolution of the gas from $S_1$ to $S_2$ does not provide this information. Nor did I claim that the derivation of $S_2$ from $S_1$ is entirely unexplanatory. On the ‘what-if-things-had-been-different’ account, this derivation does explain something: namely why the particular distribution of molecular positions and momenta $S_2$ rather than some alternative to it was realised. The problem is that at least in typical cases, this is not the explanandum that we want explained: to
repeat, typically, what we are interested in explaining is why $P_2$ rather than some alternative pressure eventuated. In other words, the explanatory limitation of the micro-account does not have to do with its providing less good explanations than the macro-account in general, but rather in its failing to explain something rather specific that we want explained. This is what I meant by saying “there are important respects in which [the micro-strategy] would fail to provide the explanation of the macroscopic behaviour of the gas we are looking for”.

It does not follow from this argument (and I certainly would reject any general claim to that effect) that reductive or micro-theories are always less explanatory than macro-theories. Still less would I want to endorse any general argument for ‘anti-reductionism’. Statistical mechanics (SM) is a paradigm example of a reductive, micro-theory and the ‘what-if-things-had-been-different’ conception of explanation shows how SM explains many features of the behaviour of gases, including behaviour that is not explained at all by phenomenological thermodynamics. It is worth emphasizing, however (and it is relevant to the example under discussion) that statistical mechanics does not explain macro-behaviour by following the micro-strategy described above.

Having said this, let me also add that if I were writing this part of $MTH$ today, I would express myself somewhat differently. I think that Sober is entirely correct to criticise, as he has done in more detail elsewhere (Sober 1999) the sweepingly general arguments for the explanatory superiority of macro-level or non-reductionist theories that one finds in writers like Fodor (1974) and Kitcher (1984). Caricaturing only slightly, these arguments contend that when the properties $P$ and $Q$ related in some macro-level generalisation ‘If $P$, then $Q$’ have (or may have) multiple micro-realisers, the macro-level generalisation will provide better explanations than the micro-theories that describe the behaviour of the micro-realisers (because the macro-theory is more unified or more general) and is to be preferred for that reason. I did not mean to endorse such arguments in $MTH$, but I see now that the remarks that Sober discusses may well have suggested that I was sympathetic to them. In fact, although I lack the space to show this, I think that my account of explanation provides strong reasons for rejecting such arguments.
1. In more complex structures, a more complicated characterisation of causal relevance is required. Cf. *MTH*, Chapter 5.
2. What is sometimes called the New Economic History (or New Institutional Economics) offers a number of additional illustrations of this point in connection with the norms and rules governing property rights.
3. To put the point in a slightly different way, the micro-strategy is inappropriately specific. An analogy: a window will shatter if struck by a rock with momentum greater than 2 kg m/sec. It is at best misleading to (just) cite the fact that it was struck by a rock with momentum 3.98 kg m/sec to causally explain its shattering. This tells us nothing about the conditions under which the shattering would not have occurred.

**REFERENCES**