Essay #2, Topic #2

On Hempel's Account of Scientific Explanations

What should a scientific explanation consist of? This is a question that Carl Hempel tackles in his essay "Laws and their role in scientific explanation." His prescription has played an important role in the history of the philosophy of science and arguably in science itself. However, as I will discuss in this paper, his theory does not address some important concerns that can be raised with this criterion of explanatory relevance: the role of post hoc explanations and the role of human psychology. As I shall address, Hempel's model is not explicit about these topics, but if it is the case that the model excludes post hoc explanations and explanations where human cognition does not allow the prediction of the explanandum based on the explanans, we would apparently be left with no explanation at all.

Hempel proposes two models of scientific explanations, both of which require the use of laws. The first type Hempel discusses is the deductive-nomological explanation. Under this model, a scientific explanation is a deductively valid argument, where the fact to be explained – called the explanandum – is the conclusion. The premises consist both of laws of nature, and particular facts about the world. These statements are called explanans. An example of such an argument-cum-explanation would be:

1 *A gas exerts more pressure as its temperature increases, all else equal.*

2 *The temperature today is higher than it was yesterday.*

3 *My tire pressure is higher today than it was yesterday, but I have not added air.*

Statement 1 is a law of nature, or in Hempel's terminology a "covering law," as it subsumes the explanandum 3. Statement 2 is a particular fact about the world. Statement

3 is the deductively valid conclusion of premises 1 and 2 (given some additional implied premises such as “my tire contains a gas,” “my tire pressure gauge is accurate,” and the like).

The inductive-statistical explanation also models explanation as an argument, with laws of nature and particular facts as the premises and the explanandum as the conclusion. However, in an I-S explanation, the argument does not have to be deductively valid; rather, it must be inductively strong – that is, that the premises must make the conclusion highly likely. The I-S model allows for probabilistic laws to be used as premises.

Hempel proposes two conditions that must be met for an explanation to be satisfactory. The first of these criteria he calls “the requirement of explanatory relevance.” By this, he means that the explanation must make one expect or believe that the fact to be explained did or would occur. Hempel calls his second criterion “the requirement of testability,” meaning just that the statements used as premises in the explanation (the laws and particular facts) must be able to be empirically tested.

According to Hempel, laws are a necessary part of explanations because laws fill the requirement of explanatory relevance. That is, laws provide the link whereby the explanandum is to be expected based upon the other explanans. In a sense, scientific laws could be thought of as rigid, well-verified if-then statements. *If* a gas is heated, *then* it will expand. *If* a magnet is broken, *then* the two pieces will be magnets as well. Hempel puts it this way “laws….are...statements of universal form [which] asserts a uniform connection between different empirical phenomena or between different aspects of an empirical phenomenon” (p. 309).

Hempel's model of explanations seems quite simple and perhaps too straightforward to bring any major criticism against. Upon reflection, however, it has some serious and problematic implications. One question that can be raised against Hempel's view of explanation is, why are laws needed as premises/explanans rather than simply universal generalizations? Indeed, what makes a law different from a universal generalization in the first place? This is a fairly titanic question that Hempel hands off to other philosophers (at least in his essay "Laws and their Roles in Scientific Explanations). I will do the same, except to note that Hempel fails to justify the inclusion of laws over universal generalizations. He comments that laws are thought of as different from other universal generalizations because laws can be used to support counterfactuals. However, he does not address why this is the case, nor why the property of supporting counterfactuals is necessary for explanation.

Beyond this, we must ask, what is unique about Hempel's model in terms of description or prescription? One feature seems the most prominent: his requirement that people must expect (be able to predict) the explanandum based upon the explanans – that is, the requirement of explanatory relevance. This leads to two problems that I will address in turn: it does not seem to allow for post-hoc explanations, and it apparently does not allow for explanations where human cognitive limitations prevent the prediction of the explanandum from the explanans.

The first issue is that the Hempel's criterion of explanatory relevance means that the model does not account for certain situations in which explanations are taken as acceptable (e.g., by the scientific community) but where the explanandum cannot be predicted from the explanans. One example of this could be evolutionary explanations.

Suppose that one wants to explain how a species of moths came to have very long tongues. A D-N style explanation would require premises (laws and particular facts) that would predict this. But, though we can have extensive knowledge about the environment of these moths and the mechanisms of evolution, it does not seem possible to predict a particular adaptation. It seems that in such cases, if we did not already know what sort of adaptation had occurred, the premises available to us would be something along the lines of:

4 *Having a trait that makes survival and reproduction more likely causes that trait to become more common in the species. (Law)*

5 *Moths of species X live in an environment where having a longer tongue would make survival and reproduction more likely. (Particular fact)*

6 *Moths of species X live in an environment where having black wings would make survival and reproduction more likely. (Particular fact)*

7 *Moths of species X live in an environment where having short antennae would make survival and reproduction more likely. (Particular fact)*

8 *Etc., ad infinitum.*

It is apparent that we could not predict the explanandum (that moths of species X have long tongues) based upon these premises. However, we do find certain evolutionary explanations to be acceptable. For the moths in question, the evolutionary explanation would go something like this:

9 *Moths of species X evolved in an environment where having a longer tongue would make survival and reproduction more likely. (Particular fact)*

10 *Having a trait that makes survival and reproduction more likely causes that trait to become more common in the species. (Law)*

11 *Moths of species X randomly mutated such that some moths had longer tongues. (Particular fact)*

12 *Moths of species X have very long tongues.*

Here, the explanandum does follow from the explanans. But in order to make this happen, we had to add a particular fact (11) post hoc. The only way we know fact 11 is through an argument with the explanandum as a premise:

13 *Moths of species X have very long tongues*

14 *Random mutation leading to a trait that makes survival and reproduction more likely causes that trait to become more common in the species.*

15 *Moths of species X evolved in an environment where having a longer tongue would make survival and reproduction more likely.*

16 *Moths of species X must have randomly mutated such that some moths had longer tongues.*

If Hempel's model of explanation bars this sort of post hoc reasoning (as it seems it should, since one should not be able to *assume* the conclusion in order to *predict* the conclusion), it would bar the argument we find acceptable as an explanation. It is unclear how Hempel would respond to this problem. He may conclude that the argument we take as an explanation isn't properly an explanation, and further, evolutionary explanations in general are nearly impossible to provide – the random nature of mutations prevents us from having knowledge of particular facts of mutations. However, it seems that the post hoc explanation given is better than nothing, and does contribute to our understanding.

If Hempel's model does not accommodate our post hoc explanations, this would be a problem for his model. The model requires explanation to involve prediction, but prediction and explanation are very different animals, with the key difference here being that facts learned post hoc can be (and regularly are) used for explanation. Consider this scenario: you are in a casino with a friend. Suddenly, near the slot machines, bells and whistles and lights go off, and a woman jumps up and down, whooping and yelling, "hooray!" Your friend asks, "Why is she yelling like that?" You take in the sights and deduce from the explanandum (the woman's yelling), and other particular facts (the bells and lights, your location in a casino, the woman's location near the slot machines, the nature of slot machines) that she must have won the jackpot. You reply to your friend, "I think she won the jackpot." Your friend is satisfied with your explanation, but you would not have been able to predict the yelling without knowing she won the jackpot, a fact you deduced using the yelling itself.

Many other scenarios can be given where the explanandum is required to deduce some fact used as part of the explanans, and where people take these as perfectly satisfactory explanations. In order for Hempel to defend his model, he would need to show either how such accounts fit in his model, or why we should not accept such accounts as explanations when they seem to contribute to our understanding of how things happen.

The second issue I take with Hempel's model is that there seems to be a heavy psychological component in the requirement of explanatory relevance, but its importance is not addressed. Suppose we have a situation where the explanandum is theoretically predictable from the explanans, but this prediction would require much more computing power or mental capacity than any person is capable of. Is the account provided an explanation? What if there was a situation where only people with particularly large working memory capacity could predict the explanandum from the explanans? Is the account an explanation, but only for some people? Do we want our conception of an "explanation" to depend upon the mental abilities of those we are explaining to? The answer to this question is not readily apparent. Further, and perhaps more pertinent to modern science, what if a person could only predict the explanandum from the explanans within his lifetime with the aid of a computer? Does this count as an explanation?

One aspect of science that these questions seem particularly relevant to is connectionist models or neural networks in cognitive psychology. Scientists have attempted to understand, explain, model, or simulate (depending upon which scientist you ask) how the human brain processes information by using networks of idealized neurons instantiated on a computer. The networks consist of many "units" that mimic the actions of human neurons. Each unit is connected to many others and a unit can send signals to other units it is connected to. An individual unit performs only the following functions: it adds up the strengths of signals it is receiving from other neurons, weighting them based upon the strength of its connection with the transmitting unit. Then, the unit calculates (based upon an assigned function) what strength signal to send out based upon the inputs it received, and then sends out this signal. The strength of a connection can be changed through intervention (i.e., the experimenter changes the weight) or though a learning mechanism (e.g., back propagation) wherein the network changes its own connection weights based upon programmed rules. Some units in a network get input from the environment, and some output their activity into the environment.

A network built of these simple units, adhering to these simple rules can be made (through appropriate learning mechanisms and input) to do remarkable things and become quite complicated. For example, one network (discussed by Hinton, 1992) can recognize handwritten digits. A typical network may consist of thousands of units with hundreds of thousands of connections. But, if a person was made fully aware of all the parameters of the network (how the units work, all the connections between units and their weights, the rules by which the weights change, the input-output conversion function the units implement, and the input the network has received) she could not predict what the output would be – at least not without using a computer to implement the network – though theoretically, the explanandum is predictable from the explanans. So, should we consider the parameters of the network an explanation? Even more, should we by analogy take – as some scientists do – the features of a human's neurons to be explanations of how the brain works?

Some scientists (most notably McCloskey, 1991) have argued that appealing to the parameters of neural networks is no explanation, since one still does not really understand the inner workings of the network. He makes an analogy to a black box, saying that if you do not understand what happens between input and output, one cannot refer to the black box as an explanation. McCloskey's objection is that he does not *understand* how these networks work. He discusses a particular network (Seidenberg & McClelland's 1989 model of word recognition and naming) designed for word recognition. He argues that one cannot answer questions such as "why doesn't the network distinguish 'shin' from 'chin'?" by appealing to the parameters of the network, and it is not possible to give an explanation on appropriately understandable theoretical levels. For example, one cannot ask of a network, “does it perform orthography-phonology conversion in a single mechanism?” and get an understandable answer.

If we cannot appeal to the parameters of the neural network to explain why it does not distinguish ‘shin’ from ‘chin’ – since a person could not predict this outcome based upon the parameters – what else can we use to explain this? The apparent answer is nothing. Failing an explanation that fulfils Hempel’s explanatory relevance criterion, are we left with no explanation at all? This seems unsatisfactory. Again we have a circumstance where Hempel’s model of explanation may bar an account as explanation but fail to provide an alternative. Surely appealing to a neural network’s parameters to explain its behavior is better than no explanation at all. It is unclear how Hempel would accommodate these charges.

To sum, Hempel’s D-N and I-S models of explanation fail to explicitly address the two concerns raised: the role of post hoc explanations and the role of human cognition in explanation. The model’s apparent stance, though, would give us no explanation at all in situations where explanations not in accord with the model are typically given. This is a problem for the model and warrants further discussion.

References

Hempel, C.G. (1966). *Philosophy of Natural Science*. Englewood Cliffs: Prentice-Hall.

Hinton, G. E. (1992). How neural networks learn from experience. *Scientific American*, 267(3), 105-109.

McCloskey, M. (1991). Networks and theories - The place of connectionism in cognitive science. Psychological Science, 2(6):387-395.