Inference to the Best Explanation

Richard Johns
revised October 2008

Before we can discuss inference to the best explanation, we must be clear on what an *explanation* is. That involves the notion of causation, and also of inference.

1. Causation and Inference

There is an important difference between physical causation and rational inference. Consider, for example, the following two sentences, that both involve the word 'because'.

> Charlie is sick because he ate lobster yesterday
>
> Charlie is sick, because it is noon and he is still in bed.

The first sentence makes a claim of cause and effect, namely that Charlie's present sickness was *caused by* his eating lobster yesterday. (Charlie may be allergic to seafood, or perhaps the lobster had sat in the sun for too long, and gone bad.) The second claim, however, does not concern cause and effect. The speaker doesn't claim that Charlie's remaining in bed longer than usual *caused* him to become sick. Rather, his remaining in bed is *evidence* for his being sick. From the fact that Charlie remains in bed, at time when he would usually be up and about, we are invited to *infer* that Charlie is sick.

Causation is a physical process that happens in the world, whereas inference is a mental process that occurs in the mind of a rational person. There is a big difference between saying:

(1) The physical event A produced the physical event B, and

(2) From knowing that A occurred, I can infer that B also occurred.

Surprisingly, perhaps, one of the most common confusions in thinking about causation, even among philosophers, is to get it mixed up with inference. *Do not make this mistake!*

2. Explanation

Section 1 says that the fact that A causes B has nothing to do with any inference from A to B. However, when we *claim* that A causes B, we typically back that up with an inference from A to B. The distinction here is subtle. The mere existence of a causal relation doesn't require us to be able to infer anything. But if we *say* that there's a causal relation, people will expect us to back that up with an inference. This is clear in the case

of trying to figure out what caused an observed event. (When one tells a story about what caused an observed event, one is trying to *explain* that event.)

Definition An *explanation* is a story about what caused an object to exist, or an event to occur.

Suppose a heavy log, weighing about 1200 pounds, moved about twenty feet up a slight incline during the night. Why did it move? (I.e. what caused it to move?) Now, suppose I saw a small child playing around the log yesterday, so my explanation is that the child moved the log. Do you accept this as a likely explanation?

I guess not, on the grounds that a small child wouldn't be strong enough to push such a heavy object up a slope. The purported cause just isn't *sufficient* for the effect. If you know some physics, you can even prove this insufficiency. The child can exert a force of only 30lb at the most, and it would take at least 800lb of force to get the log moving, from which it *follows* that the log will not move. You see what's happening? We *infer* from the supposed cause that the effect will not occur. So that cannot be the true cause.

Another explanation might be that a pickup truck dragged the log to its new position. We do the math: a truck engine has about 350lb.-ft. or torque, which with a low gear of 4:1 and a tire radius of 14" means about 1200lb of force. That should be enough.

In general, we only accept an explanation if the observed effect can be inferred from the purported cause.

[3. Aside: Stochastic Causes?]

I have been deliberately vague in saying that, in offering an explanation, one shows that the effect can be "inferred" from the alleged cause. You may be asking yourself whether I mean that the effect can be inferred with *certainty*, or merely with high *probability*. Good question.

Before I answer this, let me make something more precise. When I talk of inferring the effect from the cause, this is really a bit sloppy. Causes and effects are physical events, whereas the premise and conclusion of an inference are thoughts, or propositions. I should really say that we infer a *description* of the effect from a *description* of the cause. This clarification leads to a further question, however. How detailed is the description of the cause that we use? In general, our knowledge of the physical world is incomplete, so we are not able to give a complete description of the purported cause. We have to make to with a rough description.

Let me now answer the question above. With such a rough description of the alleged cause, we often find that the effect follows only with high probability, not with certainty. A good example of this is the case of the gummy carburetor. We're told that an engine with a gummy carburetor *sometimes* stalls, when you stop the car. (In other words, with

this rough description of the engine "the carburetor is gummy" we can only infer that the engine will stall with a certain probability.) In cases like this we have a strong intuition that there's another factor at work that, if we included it in our description of the engine, would enable us to infer with certainty whether or not the engine will stall each time we stop the car. For example, that factor might be the RPM of the engine. The engine with a gummy carburetor might stall whenever the RPM falls below 900.

This example leads us to a hypothesis that some find very plausible, called *determinism*.

Definition If an effect can be inferred, with certainty, from a sufficiently-detailed description of its cause, then the cause is said to *determine* the effect. The doctrine of *determinism* says that every effect is determined by its cause.

In the carburetor example, it is an intuition of determinism that lead us to believe that there is some other factor which, when known, will enable us to predict with certainty when the engine will stall.

But determinism may not be true. Indeed, according to most physicists, quantum mechanics shows us that determinism is false. Consider (for example) the "decay" of an atomic nucleus, when the nucleus emits a number of smaller particles. Physicists tell us that, while some types of nucleus tend to decay more quickly than others, it is impossible to predict exactly when a particular nucleus will decay. Even the best possible description of the nucleus at one time (the quantum state) does not allow you to infer the time of decay with certainty, but only provides a (physical) probability for each time.

Some philosophers interpret the physicists as saying that the nuclear decay has no *cause*. But is this right? After all, what the physicists actually say is that the decay event cannot be *inferred* from the quantum state. These are the philosophers I warned you about, the ones who confuse causation with inference!

The right thing to say is that, while the event is caused, (at least we have good reason to think it is) the event cannot be inferred with certainty from even the best possible description of its cause. This is what is meant by a stochastic cause.

Definition Event A *stochastically causes* B if and only if A causes B, but the occurrence of B cannot be inferred with certainty from any description of A.

Definition An event with a stochastic cause is said to be *random*.

Note that, when some philosophers talk of random events, they mean events that are uncaused. In other words, the event is not produced by anything at all; it simply appears, spontaneously, *from nowhere*. The existence of such events is extremely dubious, however. I know of no evidence that such events exist.

4. What is a *good* explanation?

We have seen that an *explanation* of an event is a story about what *caused* the event to occur. We also saw that, for a purported cause (explanans) to be accepted as plausible, one must be able to *infer* the occurrence of the explained event (or explanandum) from the supposed cause. In other words, one must show that, if the purported cause is supposed to exist, then one *expects* the explanandum to exist as well.

A good explanation has to do more than this, however. It is not enough that the explanandum be inferable from the alleged cause. Consider again the 1200lb log that moved uphill during the night. I might offer the explanation that, during the night, a troupe of about 20 trained baboons passed by, with a strong rope. Cleverly they tied the rope securely to the log, and then spread out along the rope. Gripping the rope tightly with their furry hands, they dragged the log uphill, before melting away into the night.

No doubt you find this explanation weak, unlikely, or implausible. But what's wrong with it? For there is no doubt that the effect can be inferred from the cause here. We supposed that the log requires a force of about 800lb to slide it along, which is only 40lb per baboon. But baboons are strong animals, and could easily manage that. Under the supposition that the cause exists, the effect would certainly occur.

Once you start thinking up this sort of explanation, there's no end to it. Maybe a passing flying saucer caught the log in its 'tractor beam' and dragged it that way. Maybe …

The weakness in such explanations, I would say, is that the purported cause is not *plausible*, within our background knowledge. Explanations involving troops of trained baboons, or tractor beams from passing spaceships are rightly rejected despite the fact that they posit a cause of the explanandum, from which the explanandum can be inferred.

So there are three conditions for H to be a good explanation of E:

1. *Causation* Condition	H makes a claim about something that caused E. It may describe the nature of a known cause, or posit the existence of a previously unknown cause.
2. *Inference* Condition	E can be inferred from H, to a high degree. Bayesians say that E has a high *likelihood*, given H, relative to background knowledge K, i.e. that $P_K(\text{E} \mid \text{H})$ is high. (Note that $P_K(\text{E} \mid \text{H})$ need not be close to one, but rather must be a large multiple of $P_K(\text{E})$.)
3. *Plausibility* Condition	H is relatively likely to be true, compared to competing hypotheses, given our background knowledge. This background knowledge may include the results of other experiments that have been used to test H.

A good explanation is not always a true explanation, of course, but it usually is, as will be argued below. (Also, a true explanation is not necessarily a good one.) For a (fairly) good explanation that turned out to be false, consider the planet Vulcan.

The background to this story is that, after powerful telescopes were developed in the nineteenth century, astronomers noticed that Uranus's orbit departed slightly from predictions based on Newtonian mechanics. Rather than reject this well-established theory, some argued that Uranus's motion was being affected by the gravitational field of another, unknown planet. The orbit of this unknown planet was calculated, using Newton's theory, and the planet Neptune thus discovered, just where it was predicted to be.

So a precedent was set: when a planet deviates from its strict Newtonian course, the gravity from another (unknown) planet is to blame. Mercury was the next whose orbit was noticed to be a little off. The best explanation, of course, was to posit another planet, whose orbit was calculated to lie inside even that of Mercury, the closest known planet to the sun. The new planet would be very hot, so they called it Vulcan, after the Greek god of fire.

At this point, would you say that the Vulcan hypothesis is a good explanation of the deviation in Mercury's orbit? Note that the only (practical) alternative would be to say that Newtonian mechanics is false. And Newtonian mechanics was the greatest scientific achievement in human history, the sublime and wonderful theory that was the foundation of all physics. Kant even viewed it as *a priori*, a necessity of human thought about the world.

Some astronomers claimed to observe Vulcan, but most could not, so the Vulcan hypothesis languished. There was no viable alternative, however, until 1920 or so, when Einstein's general theory of relativity replaced Newton's theory of gravity. Einstein's theory predicted the observed orbit for Mercury, assuming no planet existed between Mercury and the sun. So Vulcan vanished.

5. Inference to the Best Explanation

As far as I am aware, something like the rule of Inference to the Best Explanation (IBE) was first stated by physicist Christian Huygens:

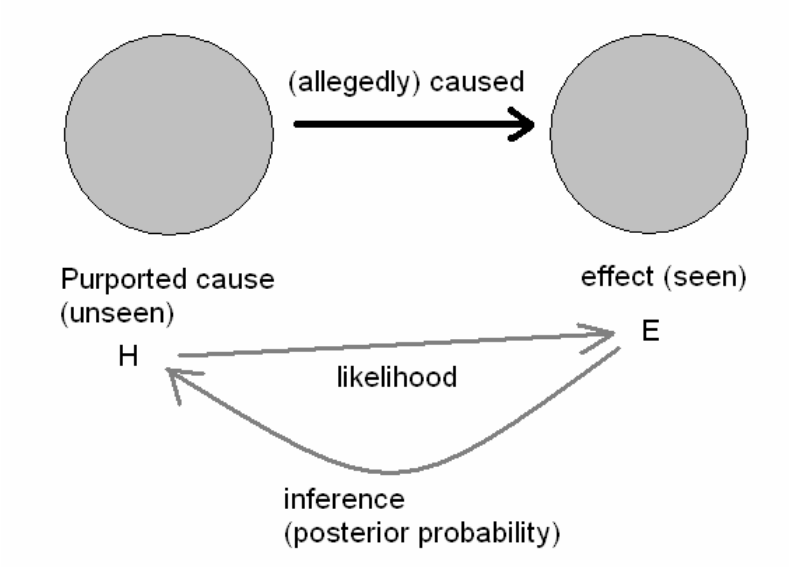
> *Treatise on Light* (1678)
>
> "One finds in this subject a kind of demonstration which does not carry with it so high a degree of certainty as that employed in geometry; and which differs distinctly from the method employed by geometers in that they prove their propositions by well-established and incontrovertible principles, while here *principles are tested by the inferences which are derivable from them*. The nature of the subject permits no other treatment. It is possible, however, in this way to establish a probability which is little short of certainty. This is the case when the consequences of the assumed principles are in perfect accord with the observed phenomena, and especially when these verifications are very numerous; but above all when one employs the hypothesis to predict new phenomena and finds his expectations realized.

The "principles" Huygens refers to here are hypotheses about the nature of light, which are used to explain the results of optical experiments. A hypothesis that makes correct predictions is a good explanation (as long as the hypothesis is plausible, although Huygens says nothing about that).

So we can summarise this general rule as "From a given set of data, infer the best explanation of those data". This rule is called "inference to the best explanation".

Definition The method "Inference to the Best Explanation" (IBE) tells you which hypothesis H to infer from the available evidence E. It says you should infer the hypothesis H that *best explains* E.

Inference to the best explanation is, in a sense, reasoning backwards. Let us focus on the second (inference) condition of a good explanation, that H explains E when E can be



inferred from H. Then IBE says (roughly) that we should infer H from E in those cases where E can be inferred from H. From E, we infer the hypothesis from which E is best inferred. (Confusing?)

There are actually three "arrows" involved, for each possible explanation of E. One is the (alleged) causation of E by its purported cause. The other two are inferential arrows. Suppose the hypothesis H described the purported cause. Then we must first consider the degree to which H entails E, i.e. the probability of E given H (this is the *likelihood* of E under H.) The higher this likelihood is, and the more plausible H is within *K*, the more strongly we can infer H from E (this is the lower, backward grey arrow). The inference from E to H is known as an *inductive* inference, since H is probable to some extent, but not certain, given E.

One important consequence of IBE is that science is essentially *competitive*. One cannot know whether to infer H from E simply by looking at H, E, and relations between them like $P_K(E \mid H)$. One needs to other what *other* explanations of E are possible, and compare them to H. Sometimes a hypothesis seems very probable, as it accounts for the data pretty well. But then you think of an even better explanation, and suddenly the first one seems weak. The reverse can happen. A hypothesis may seem like a rather poor explanation of the data. But when you realise that all the alternative theories are far worse, that hypothesis becomes highly probable.

There is an analogy here with sport. In one tournament a weak team might win easily, since all the other teams are far worse still. In another tournament a strong team might not win, since another team is even better.

Let's look at some examples that illustrate this point. Consider first the problem of explaining the origin of life on earth. These are the main hypotheses have been proposed to explain this.

1. Abiogenesis

   Laboratory simulations of the early earth show that various amino acids can form under those conditions. And amino acids are the building blocks of proteins, DNA, and the rest. These amino acids became concentrated in little pools of water, forming a "pre-biotic soup". Then complex, self-replicating molecules (perhaps some form of RNA) spontaneously formed in the soup, as amino acids bonded together. These molecules were the first life on earth.

2. Pan-Spermia

   Life originated on some other planet (earth being unsuitable). Simple living organisms (such as bacteria) were somehow transported to earth. (Perhaps they were in a comet, whose tail the earth passed through? Perhaps an advanced civilisation brought them here on a space ship?)

3. God did it

   The universe was created by a living, intelligent being. That being wanted to create other living organisms, and chose earth as a place to do it.

As you can see, this is an extremely weak field. The first explanation, that of self-assembly by pure chance, is generally acknowledged to be extremely problematic. (It's worth noting that there are many different versions of it.) Even the simplest self-replicating molecules contain thousands of just the right amino acids, arranged in just the right way. The chance of this happening spontaneously is vanishingly small. The second theory hardly deserves to be called an explanation at all, as it is hopelessly vague. The third theory is widely held, but not among scientists. If we routinely invoked God to explain physical phenomena, then science would be finished. So we'd much rather not appeal to God here, either.

If we compare 1 and 3, we see that they have opposite defects. For 1, it is $P_K(\mathrm{E} \mid \mathrm{H})$ that is very small. For 3, it is $P_K(\mathrm{H})$ that is small.

What do we do in a case like this? Well, perhaps another (stronger) hypothesis will be found. If these three options are all there is, on the other hand, then one of them must be true. It may even be that one of them, while weak, is much less weak than the other two, in which case it will be highly probable, and reasonable to believe. Note that advocates of all three views argue on this basis. Each says, in effect, "While my view might seem laughable, you should see the competition …"

An example of the other kind, where there are many strong explanations of a phenomenon, is the case of the missing cod in the North Atlantic. It's well known that cod stocks in the North Atlantic near Newfoundland have collapsed, and have failed to recover even after the cod fishery was completely closed some years ago. Here are some possible explanations:

1. The expanded seal population is eating more cod, keeping the stocks low.

2. The warmer ocean temperatures are hurting the cod somehow.

3. The cod stocks were driven so low, by fishing, that the few cod fry that are now produced are mostly eaten by predators.

4. Some disease is killing off the cod.

Now I'm not a marine zoologist, so I don't know how strong these explanations really are, but they all look fairly reasonable to me. Each is certainly far superior to all of the explanations of the origins of life.

The moral of the story is that some things, like the origin of life, are devilishly difficult to explain, so that there are no good candidates. Others, like the missing cod, are much easier, so that there are many good candidates. So you cannot reject an explanation simply because it is weak, or accept one simply because it is strong. You have to look at the competition.

6. Bayesian IBE

Bayes's theorem captures the idea of IBE very neatly. It actually goes a little beyond IBE in being precisely quantitative. It tells you exactly how strongly you should infer each of the possible explanations of an observed phenomenon.

Suppose an event E has been observed, and there are three possible explanations of E, namely H_1, H_2 and H_3. Prior to observing E, our epistemic state was K. Bayes's theorem then states:

$$P_K(H_1|E) = \frac{P_K(E|H_1)P_K(H_1)}{P_K(E|H_1)P_K(H_1) + P_K(E|H_2)P_K(H_2) + P_K(E|H_3)P_K(H_3)}.$$

(If there were more hypotheses, we would simply add more terms to the bottom of the fraction.) The equation may look complicated, but notice that it's very repetitive. The expression $P_K(\mathrm{E} \mid \mathrm{H}_1)P_K(\mathrm{H}_1)$, for example, appears on the top of the fraction and again at the bottom. This expression measures how good an explanation H_1 is of E, since for it to be a high number, both $P_K(\mathrm{E} \mid \mathrm{H}_1)$ and $P_K(\mathrm{H}_1)$ have to be large. The other two expressions in the theorem are $P_K(\mathrm{E} \mid \mathrm{H}_2)P_K(\mathrm{H}_2)$ and $P_K(\mathrm{E} \mid \mathrm{H}_3)P_K(\mathrm{H}_3)$, which measure the quality of H_2 and H_3 as explanations of E.

For simplicity, let's define the *strength of H_1* (as an explanation of E) to be the product $P_K(\mathrm{E} \mid \mathrm{H}_1)P_K(\mathrm{H}_1)$, and the same for H_2 and H_3. Bayes's theorem then becomes:

$$P_K(H_1|E) = \frac{Strength\ of\ H_1}{Strength\ of\ H_1 + Strength\ of\ H_2 + Strength\ of\ H_3}.$$

Since $P_K(\mathrm{H}_1 \mid \mathrm{E})$ is a fraction, the actual values of these strengths are unimportant. What matters are their *relative* sizes. If they are all equal, for example, then $P_K(\mathrm{H}_1 \mid \mathrm{E}) = 1/3$, regardless of their actual value. The actual value could be 0.1 or 0.000001, for example. Similarly, $P_K(\mathrm{H}_1 \mid \mathrm{E})$ could be quite high (say 0.9) even if the strength of H_1 is quite low, say 0.000001. It only requires that the other strengths be even lower! This is the case when a weak explanation is strongly inferred, because the alternatives are much weaker still. Again we see the *competitive* nature of science.

7. The Subjectivity of IBE

Some philosophers, such as Karl Popper, have criticised IBE as being too subjective. After all, the third (plausibility) condition for a good explanation allows free rein for one's biases, prejudices, and so on. It allows neo-Platonists like Copernicus and Kepler to believe the heliocentric hypothesis on spurious grounds of mathematical harmony and economy, and for giving the sun (God's representative in the visible world) a fitting location. Scientific inference, by contrast, is (according to Popper) objective, allowing no place for such personal input.

(Popper thought that the scientific method is to propose bold, risky, hypotheses, and then to try to *refute* them empirically. At no point is a hypothesis regarded as probably true, or a good explanation. The most we can ever say is that the hypothesis is bold, and not yet refuted.)

Supporters of IBE often counter that the subjective element in real science is well documented, so that any account of science that failed to acknowledge it would be deficient. Furthermore, a limited role for subjectivity does not, it is argued, threaten the high degree of warrant that many scientific theories enjoy. Where the data are plentiful and accurate, for example, researchers whose initial biases are quite different will eventually "converge" on the same opinion, as those prejudices are overwhelmed by the empirical evidence. (This is, after all, what happened in the conflict between the heliocentric and geocentric theories.) This defence of scientific rationality is often called "the washing out of the priors".

Others argue, on the other hand, that even in the most favourable cases, the data never *determine* a unique hypothesis. (This fact is referred to as the "underdetermination of theories by the evidence".) There is always more than one possible explanation of any data set, and the judgement that one explanation is "best" is always highly subjective.

Finally, some philosophers claim that the rationality of science requires that we humans have some *a priori* knowledge. The underdetermination of theories by the evidence requires that additional, non-empirical, knowledge exists, if science is to be rational. (And, of course, science *is* rational, on this view.) Our *a priori* knowledge need not be very specific, nor need it be certain. It might be little more than a (warranted) tendency to think that the natural world is likely to be uniform in certain respects. But however modest such *a priori* knowledge might be, there remains the difficult question of how we might have acquired it.