The Place of Explanation in Scientific Inquiry: Inference to the Best Explanation vs Inference to the Only Explanation.

James Woodward

HPS Pittsburgh

1.17.24

1. Introduction

This is a paper about the place of explanation in scientific thinking. (There will be some remarks about the role of explanation in more common sense contexts but this is not my main concern.) One of my main themes is that finding explanations is a distinctive, independent goal in scientific inquiry. It is not replaceable by the goal of finding hypotheses which have such features as simplicity, unifying power, or other supposed theoretical virtues. With an important exception, noted below, involving "tuning", the role of explanatory potential (understood as having to do with with how well a hypothesis would explain if true) is also not to guide us to hypotheses that are true or inductively well supported, as claimed by advocates of inference to the best explanation (IBE).

Philosophical discussion of explanation has long recognized the notion of a potential explanation-- a set of claims[1] that if true would constitute a successful explanation of some explanandum. This implies that we can evaluate how well such claims would explain if those claims were true without knowing whether these claims are in fact true. I will adopt this idea and argue that it leads to the following picture: When the goal is finding an explanation of *E,* we often proceed by assuming that there exists some explanation of *E* (usually of some relevant sort -- a crucial assumption and one that may be mistaken), then formulating a set of different competing potential explanations of E and then (in a separate investigation) testing these empirically. Ideally this testing will involve finding evidence that undermines all but one of these competitors, the remaining one alone being supported by the evidence. In many cases, this will require evidence in addition to *E.* On this picture, considerations having to do with potential explanatoriness do not themselves guide us to what the truth is -- rather they guide us in constructing or discovering alternative potential explanations (as well as elaborating these -- see Section 12) which we then attempt to discriminate among by finding *additional* evidence.

[1] For convenience I will often describe these claims as "hypotheses" but the reader should think of this as a catch-all term meant to cover theories, models etc. Also we should bear in mind that explanatory claims typically consist of claims about initial and boundary conditions and constraints in addition to generalizations.

Successful explanation thus does require a true explanans (or something in this neighborhood-- see below) but establishing that an explanans is true requires the kind of discriminating evidence just described and not just appeals to potential explanatoriness. This process is sometimes described as inference to the only (remaining) as explanation (IOE) as opposed to inference to the best explanation. This strategy is reflected in Sherlock Holmes' often quoted remark.

> When you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth.

In what follows much of my discussion will be structured around the contrast between IBE and OBE with the former serving as a kind of foil to bring out the alternative view of the role of explanation which I think more defensible.

I begin by fleshing out the general picture of the role of explanation gestured at above as well as some background assumptions that will motivate my discussion. I will then turn to a more detailed treatment of the implications of some standard models of explanation, including the what-if- things- had- been -different account in Woodward, 2003, Hitchcock and Woodward, 2003 for IBE and OBE[2].

2. Some Background and Motivation.

I assume that science sometimes aims at discovering truths or at least claims that are, so to speak, in this neighborhood-- that is, claims that are good approximations or claims that are "effectively" valid within some domain of interest. Hereafter I will often use "truths" to encompass all these possibilities--- See Section 4 for more on this. This is not to claim that such truth -related goals are the only goals of science but merely that they are among its goals. This assumption ought to be common ground with defenders of IBE since these generally accept a truth requirement on explanation (or at least something like this, as above) and IBE is an inference method that is supposed to connect explanatoriness with evidence for truth. If successful explanation does not require the truth (or something similar) of what does the

---

[2] The literature on IBE is vast and there is much that I lack the space to discuss, even in this overly long paper. I do, however, particularly want to note a series of papers by Sober and Roche (e.g., 2013). These authors formulate (and then argue for) the following anti- IBE thesis: If *H* is a hypothesis, *O* an observation, and *EXPL* the proposition that if *H* and *O* were true, then *H* would explain *O*, then Pr(*H* | *O*&*EXPL*) = Pr(*H* | *O*)-- i.e. O "screens off" EXPL from H, so that *EXPL* adds nothing to the support for H in addition to what is provided by *O*. I don't adopt this formulation in what follows but I concur with the spirit of their claim: Whatever support there is for *H* comes from the ordinary non-IBE based evidence for H, including what is provided by *O*. That *H* would if true explain *O* adds nothing to the inductive support for *H* beyond this.

explaining, it is hard to understand how IBE is supposed to function. Of course similar remarks apply to IOE.

But even when science might aims at truth it seems obvious that not all truths are of much scientific interest. Consider a description-- as exact as possible-- of the positions of all of the fallen leaves in a certain neighborhood of Pittsburgh at 3pm on November 7, 2023. Even if possible, this is not (absent some very special circumstances) among the kinds of truths that science generally aims to discover. So the claim that science aims at discovering truths is (at best) incomplete; it needs to be accompanied by an account of *which* truths (among all truths) science aims to discover. My suggestion is that among the truths science aims to discover are truths that can figure in explanations-- discovering explanations (and hence the true claims that figure in these) is one of the goals of science. The goal of finding explanations thus provides much more specific guidance (and guidance that is needed) than the very general goal of just finding truths.

So far this may strike many as uncontroversial but notice that there is nothing in this picture that requires us to suppose that discovering explanatory hypotheses (or even hypotheses that would if true explain well) is somehow itself a means to discovering truths or that the role of explanatory considerations in science is to serve as a guide to truth or to what is inductively well supported. Instead, what I said above about the role of explanation is consistent with the picture adumbrated above: when explanation of some set of phenomena is among our goals, we often proceed by exploring possible or potential explanations of those phenomena-- that is, we try to discover or construct hypotheses or theories which *if true* would explain those goals. But finding such potentially explanatory hypotheses is by itself no reason to suppose that any particular one (even one that supposedly best explains by some standard) *is* true. To establish that we have to provide additional evidence that distinguishes among these potentially explanatory hypotheses, supporting one and ruling out or undermining alternatives. This of course is the IOE view.

On this picture, successful explanations must appeal to assumptions that are true but it gets things the wrong way around to suppose that (for the most part) the role of explanatory considerations is to guide us to truth. Instead, finding correct explanations is valuable in its own right. I will say more below about what this involves but I take it to be consistent with the obvious point that explanations are also valuable because of their connection with other goals-- explanations can provide information relevant to manipulation and control and to prediction, can facilitate learning and so on.

Here an analogy may be helpful. Finding hypotheses that successfully predict is also a goal of science and one that is distinct from explanation. Of course a necessary condition for a predictive hypothesis being "good" is that makes reasonably accurate predictions much of the time or that it has good error rates-- call this an accuracy requirement. But we can also evaluate candidate predictive hypotheses on how good or valuable they would be *if* they were accurate-- i.e., there are independent dimensions of evaluation beyond accuracy. For example, a hypothesis that makes precise quantitative predictions that are accurate will often reasonably be preferred to one that makes accurate but vague and imprecise predictions (as in the fortune teller's "Something important will happen to you in the next year"). A predictively successful hypothesis that requires relatively little information as input can reasonably be preferred to one that requires a

huge amount of difficult-to-get information. Among hypotheses that make accurate predictions, some will be regarded as more scientifically valuable than others-- an hypothesis/algorithm that accurate predicts the three dimensional shape of proteins is a more valuable scientific achievement than, say, a rule for successfully predicting the outcomes of certain high school football games.

But although candidates for predictive hypotheses with the above characteristics would, if accurate, may be more valuable or worth knowing than others, it seems obviously misguided to argue as follows: Candidate predictive hypothesis *h* would if accurate be highly valuable to know and would lead to many successful predictions about important matters; therefore this is evidence that *h* is in fact accurate. (Call this Inference to the Best Predictor, in analogy with IBE) That it would be highly valuable to discover a hypothesis that successfully predicts various outcomes is certainly a good reason to try to discover or formulate such a hypothesis but when we do so, this by itself is no reason to conclude that this hypothesis is in fact an accurate predictor. Instead that needs to be established via an appeal to independent evidence for predictive success (perhaps accompanied by supporting formal analysis of when and why the predictive hypothesis works.) It would be convenient and valuable if we could predict the onset of various diseases like diabetes, cancer and schizophrenia from the presence of just a few genes but of course it does not follow that this is a reason to think that we can make accurate predictions on this basis-- as a matter of empirical fact, the genetic influences on these diseases typically involves very, very large numbers of genes, each of individually small effect. Successful prediction-- e.g., in the form of polygenic risk scores-- if possible at all, may require a great deal more genetic information that is more difficult to discover and interpret and may have other limitations. However, such less than ideal predictors rather than those that, if accurate would be ideal, may be the ones that are empirically grounded.

Prediction is not explanation but I suggest that a similar point holds for inference to the best explanation. Just as we can't infer that a hypothesis is an accurate predictor merely on the basis that it would be a particularly useful predictor if it were true, so also we can't infer from potential explanatoriness to truth or strong inductive support. As the analogy shows, hypotheses can have potential "informational" value[3] in the sense that if true or accurate they would tell us things we would like to know, without this value contributing to inductive support. Successful explanation is one such informational goal.

If we think about the role of explanation in this way it belongs to what is sometimes called the context of pursuit rather than the context of acceptance-- a claim that has been recently advocated by other writers (e.g. Nyrup, 2015, Carbrera, 2021, Wolff and Duerr, forthcoming) . That is, the discovery that some hypothesis *h* would if true explain evidence *e* is a reason for investigating whether *h* is true by getting additional evidence that distinguishes it from alternative potential explanations of *e*. And, as argued below, this in turn requires formulating those potential alternative explanations and discovering their testable implications, a process

[3] The idea that there are informational virtues in addition to confirmational ones and that explanation is a informational virtue is bruited in Salmon, 2001 and is discussed at greater length in Carbrera, 2017

which is also guided by explanation-based considerations. But the results of this sort of inquiry does not in itself provide reasons for accepting *h* or any of these alternatives as correct.

3. Successful Explanation as an Independent Goal

This independent goal (IG) picture has a number of appealing features. First, it accounts for many of the features of scientific practice to which defenders of IBE appeal. Like IBE-based views, IG recognizes that the discovery of explanations is central to a substantial amounts of science and that explanations need to appeal to assumptions that are true or truth-like. Moreover, it agrees with IBE -based views that assessment of competing potential explanations is central to science. However, IG differs from IBE views regarding the basis on which such assessments should be made-- competing potential explanations are not ruled out on the basis of which explains best but rather by getting evidence showing that these competitors are false. (Indeed, as argued below, the IG picture need not assume that it is even possible to make the comparisons of explanatory goodness to which IBE appeals.) When presented with cases of apparently legitimate inferences which IBE-based views interpret as comparisons of explanatory goodness the advocate of IGE will instead claim that these are instead more plausibly interpreted as case of IOE[4]. As an illustration, consider the following oft - quoted passage from Darwin (1876, p. 421) in support of the theory of natural selection :

It can hardly be supposed that a false theory would explain, in so satisfactory a manner as does the theory of natural selection, the several large classes of facts above specified.

This is often taken to be an illustration of IBE (e.g., Lipton, 1991/2004). However it seems just as plausible to treat it as an instance of IOE: Darwin is claiming that there are no other candidate theories that explain the facts he describes and that are true. For example, the alternative hypothesis of Divine Creation is refuted by such facts as the existence of functionless traits-- that is, this alternative hypothesis is rejected because it is false, not because if true it would explain less well.

As another illustration consider the following remarks of J. J. Thomson in support of the claim that cathode rays consist of negatively charged particles:

As the cathode rays carry a charge of negative electricity, are deflected by an electrostatic force as if they were negatively electrified, and are acted on by a magnetic force in just the way in which this force would act on a negatively electrified body moving along the path of these rays, I can see no escape from the conclusion that they are charges of negative electricity carried by particles of matter. (Thomson, cited in Achinstein 2001, 17)

Douven (2024) claims that this passage shows Thomson reasoning in accord with IBE but Thomson's own words ("I can see no escape..") suggest that he is instead reasoning in accord

[4] Put differently, when presented with a putative case, we need to ask whether there are good reasons to interpret this as a case of IBE rather than a case of IOE.

with IOE-- that the rays are negatively charged particles is the only explanation consistent with his evidence[5].

As these examples illustrate, to provide convincing examples of legitimate inferences that are instances of IBE one must show that these cannot plausibly be interpreted instead as case of IOE. A second appealing feature of independent goal view is this: IBE faces the problem of justifying the claim that there is a connection between potential explanatory goodness and truth: why is it reasonable to assume that the best potential explanation is also the one that is most likely to be true? A similar problem of course arises for more specific candidates for explanatory virtues such as simplicity. I think it is fair to say that there is no generally accepted answer to this question. By contrast, IG does not face this problem because it does not assume that there is any general connection between potential explanatoriness and truth of the sort envisioned by IBE. Here a brief digression may be helpful. Philosophers often draw (i) a contrast between epistemic and other sorts of virtues, which are regarded as in some way "external" to science. For example, my research may have the "virtue" of helping me to get rich but this is not an epistemic virtue internal to science. It is rather an external one. Furthermore, (ii) it is often assumed that a defining characteristic of epistemic virtues is that they must be signs of (or even a means to the end of discovering) truths. Thus simplicity may be regarded as an epistemic virtue to the extent that simpler hypotheses are more likely to be true and similarly for potential explanatory goodness. IG accepts the idea (i) that there is a contrast between internal (epistemic) and other virtues but rejects the idea that the epistemic virtues are confined to those that are truth-linked in the manner described in (ii) above. In particular, explanation is regarded as a genuine epistemic virtue or internal goal in science but not because potential explanatoriness is a means to discovering truths. This is part of what I mean by saying that the discovery of explanations is an independent goal that is valuable in its own right[6].

The restriction of epistemic virtues and goals to those that can be understood as something like signs of (or means to) truth is difficult to defend, in part because it requires a connection between those virtues and truth that does not seem to exist[7]. Furthermore the demand that

---

[5] That Thomson's remarks are plausibly interpreted in terms of IOE is also noted by Nyrup (2015).

[6] van Fraassen (1980) is naturally interpreted as claiming that the single goal or aim of science is the discovery of theories that are empirically adequate-- additional considerations such as those having to do with explanation are "pragmatic" rather than "epistemic" and not part of the aim of science. My contrary view is that the goal of explanation often requires us to move beyond theories that are merely empirically adequate-- successful explanation often requires the postulation of unobservables because explananda of interest depend on these. I see the acceptance of such theories as involving the satisfaction of an epistemic goal that is not "pragmatic" in van Fraassen's sense of this term.

[7] There is also the awkward fact that informativeness is inversely related to probability with the consequence that, as many have observed, within frameworks that measure inductive support probabilistically, an hypothesis that is strictly more informative than an alternative must be assigned a lower (both prior and posterior) probability.

everything valuable in science be linked to truth in this way (so that truth is the overarching goal to which everything else "epistemic" is a means) runs immediately into a problem noted earlier: that not all truths are scientifically valuable. If only some truths are scientifically valuable there must be additional, independent goals or constraints (like explanation) that help to characterize the valuable truths. As conditions on the kinds of truths for which we are looking, these cannot (on pain of incoherence) be valuable merely because they are a means to truth.

The argument in the preceding paragraph supports the claim that some additional goals besides a generic concern with truth need to be acknowledged as internal to science. However, one might still wonder why the focus should be on explanation rather than other goals (such as finding simple hypotheses) that might also serve to pick out valuable kinds of truths.

To explore this question, consider the specific alternative suggestion that simplicity can play the role described in the previous paragraph-- that is, simple truths (as opposed to other sorts of truths) are valuable to discover for their own sake. One obvious problem, already noted, is that no one has been able to provide a clear characterization of simplicity that has the right features to play this role. But in addition, simplicity just doesn't seem like a good candidate for something that is valuable for its own sake in science. If we ask what is so great about the discovery of simple hypotheses, we seem inclined to appeal to further goals or considerations. The idea that simplicity is valuable because it is a sign of truth is one such possible candidate for such an "instrumentalist" justification and one we have found wanting. Alternatively one might hold that simpler hypotheses are (usually?) easier to reason or calculate with or to test and are valuable for that reason. Or perhaps the value of simplicity can be understood in terms of the idea that science aims at finding a small number of simple truths from which many other truths can be derived, thus assuming a conception of science according to which it aims at something like economical description or organization of truths and nothing more. Quite apart from the controversial character of this conception, when we ask why *that* kind of economical organization is so valuable, we are likely to be led back to ease of use or comprehension considerations. I think that a similar argument applies to such candidate virtues as unification-- arguments that these are valuable typically operate by claiming that these are means to something else.

I suggest that by contrast explanation is much better suited to the role of a goal that is valuable in its own right and that can act as a constraint on the kinds of truths science aims to discover . First, as already suggested and will become clearer below, finding explanatory truths is a different goal from finding truths that are simple or unifying. Moreover, if we adopt a suitably minimalist conception of explanation of the sort described in section 8 below we can avoid the mushiness and unclarity associated with judgments of simplicity and unification. Finally, the "why should we care about that? " question seems less pressing when asked about explanation. If someone claims to have discovered an explanation for the accelerating expansion of the universe or the extinction of the dinosaurs, one might legitimately wonder whether this information might serve some additional external goal but as far as goals internal to the scientific enterprise go, it does not seem that we are naturally led to ask for a further justification of why the discovery of such explanations is valuable.

4. Underdetermination as a motivation for IBE

Although it does not strictly require this assumption[8], the idea that IBE is a legitimate inference form flourishes against the background of a certain picture of science that was common in the past century but less so today, at least among philosophers of science. According to this picture, at any given moment there are usually (perhaps always) a large number of different hypotheses and theories such that the available evidence taken in itself (that is, independently of IBE- type considerations) does not differentially support one of these over the others. We thus face a massive "underdetermination problem". Nonetheless we think that we have some basis for choosing among these alternatives and since the evidence, considered in itself, is insufficient, the only alternative basis seems to involve the so-called trans-empirical or theoretical virtues--virtues that we can identify a theory or hypothesis as possessing, independently of our evidence. Considerations of potential explanatories are obvious candidates for such virtues, especially if we fold other candidates, like simplicity, into them. When this picture of science is assumed, it can seem virtually inevitable that we must be using IBE (or something in its neighborhood) extensively and indeed one finds frequent claims among advocates of IBE that this strategy is used extensively (and legitimately) in both science and common sense contexts[9].

The discussion that follows is premised on the assumption that this is a distorted picture of our epistemic situation. Instead we are often in a position to get evidence that rules out all but one of the competing potential explanations without relying on IBE type considerations or other trans-empirical criteria such as simplicity. So we don't have to rely on IBE to provide inductive support and are instead in a positin to implement IBE. Moreover, although we are able to distinguish hypotheses that are potentially explanatory from those that are not, there is usually no basis for the more fine grained ranking required to license identification of a "best" potential explanation that can be used in an IBE.

There are several reasons why the strategies available for excluding alternatives and generating inductive support are far more constraining than many philosophers recognize. One has to do with the best way of understanding the content of successful theories. The standard examples used to motivate the existence of an underdetermination problem usually come from "fundamental" physical theories-- e.g., particle and gravitational physics . But the modern way of understanding these is that they are "effective" theories, holding to some suitable degree of approximation within some limited regime or domain, characterized by an energy or length scale[10]--e.g., Newtonian gravitational theory holds for relatively weak gravitational fields and velocities that are small in comparison with those of light, General Relativity perhaps holds for energies up to the Planck length, the standard model of particle physics holds up to some unknown energy scale but not beyond and so on. Moreover a natural interpretation (or at least one that I will assume) is that what matters for the effective truth or validity of these theories is the effective correctness of the dependency relations that they postulate, as captured by, e.g.,

---

[8] The argument that follows about the underdetermination problem being exaggerated is intended to remove one important motivation for IBE. However, even if this argument is mistaken, it does not follow that IBE is legitimate. The objections to IBE that I discuss in later sections still remain.

[9] See, e.g, McCain and Poston (2023) for arguments along these lines, including the claim that IBE is widely used in science.

[10] See e.g., Weinberg, 1999.

the Newtonian gravitational force law and the field equations of GR, rather than other commitments associated with these theories, such as their supposed ontologies[11].

When the content of these theories is understood in this restricted way (rather than, say, as claims about exact and literal truth at all possible energy scales), it is easier to see how they can be strongly confirmed by available evidence. This is because they claim less than they are sometimes taken to claim, in the sense that they are not claims about truth at all energy scales and because we take them to be claims about dependency relations as opposed to other sorts of claims (e.g. about ontology. Moreover it is simply an empirical fact that there are no known alternatives to the above theories, when understood as effective claims of the sort described above. For example, there is no known alternative (alternative in the sense of making claims about dependency relations leading to substantively different predictions) to Newtonian gravitational theory in the domain of applicability of that theory that is consistent with known evidence. Similarly for GR and the standard model. Indeed, in a number of cases there are powerful arguments based on generally accepted evidence and generic theoretical considerations that such alternatives do not exist. For example, Weinberg (1996) argues that that the combination of non-relativistic quantum mechanics (really just the assumption of unitarity), Poincare invariance and cluster decomposition lead virtually inexorably to quantum field theory-- at scales at which Poincare invariance holds there are no alternatives that are consistent with the available evidence, hence no underdetermination problem of the sort bruited above. When we move to theories or hypotheses outside of physics the point becomes even clearer-- there is no evidentially credible alternative to the hypothesis that genes are composed of DNA or that a great deal of human visual processing occurs in the occipital lobe.

Of course this is not to deny that there are many cases in which we do not yet know which of several competing explanatory hypothesis is correct. There are no doubt cases in which we never know which is the correct hypothesis because the needed evidence will never be available. However, the considerations just described do suggest that the claim that underdetermination is ubiquitous (or even the general rule) is misguided. To this we may add that if underdetermination was extensive and dealt with via IBE, one would expect there to be (formulated) alternatives to the above hypotheses and theories, consistent with the available evidence, but which explain less well and are taken to be inductively unsupported for this reason. This is not what one sees in many areas of science.

---

[11] Philosophers often focus on what they take to be the "ontological" commitments associated with theories-- Newtonian theory is committed to the existence of gravitational "forces" and a notion of absolute spatial position while GR denies the existence of these. They then infer that since GR is a successor theory Newtonian theory must be fundamentally false because of these ontological commitments. I take the commitments of Newtonian theory that do explanatory work to have to do with the dependency relations it postulates rather than its ontology. This fits with the w-condition account of explanation discussed below (which focuses on such dependency relations and not ontological correctness) and also with my claim that Newtonian theory is effectively valid with its domain, since the ontology of Newton's theory is not even approximately correct.

A second set of considerations has to do with the power of the available inductive strategies themselves: Philosophical folklore to the contrary, it is sometimes -- perhaps often-- possible to systematically generate a set of alternative explanatory hypotheses that are plausibly be taken to be exhaustive (again when interpreted in the effective and domain specific way described above) and then to search through these in a systematic way, finding experimental results which exclude whole subsets of these at once[12]. Such generate and search strategies can minimize the inductive risk posed by unconceived alternatives, which are sometimes thought to pose a fatal objection to OBE. The use of the parameterized post Newtonian formalism (PPN) is a well-known example of this strategy-- see Will, 1981/1993, and Earman 1992. This formalism characterizes the space of alternative gravitational theories to GR in terms of a small set of measurable parameters and principles -- e.g., theories that obey the equivalence principle and those that do not. When stringent tests confirm the equivalence principle, this excludes in one fell swoop all theories that imply that this principle is false. Sometimes this procedure can be iterated in such a way that there is only one remaining explanatory hypothesis that is consistent with the evidence, which is what happens with GR. Machine learning of causal relations such as Spirtes et al. (2000) proceed via a broadly similar strategy.

In addition, whole classes of alternative hypotheses also can sometimes be excluded on the basis of design-based considerations -- that is, considerations having to do with the nature of the data generating process. If our evidence is merely that X and Y are correlated, there are many alternative hypotheses besides the claim that X causes Y that might explain this correlation-- it might be due to a single common cause Z, two common causes W and U and so on. But if the correlation is the result of a properly randomized experiment, we can exclude all of these alternative common cause hypotheses as very unlikely.

Finally, in many areas of science, the generally accepted explanatory theories are supported by many disparate forms of evidence and argument which converge to provide inductive support for a single result. For example, one form of evidential reasoning for theory T may take the form of comparing predictions derived from T with observed results but this may also be supplemented by a so-called deduction from the phenomena in which T is derived from observed evidence and generally accepted theoretical principles. Newton followed this strategy when he derived his gravitational law from Kepler's laws and more general assumptions about the motions of the planets being due to a forces of some kind centered on the sun. Newton also used additional complicated iterative and confirmatory procedures, as described in Smith (2014) and Harper (2011). The upshot is that many scientific theories or hypotheses are so strongly connected to many different forms of converging evidence that their "effective" correctness is greatly overdetermined.

I conclude that when a number of alternative hypotheses are consistent with the known evidence it is often possible to find additional non-IBE evidence that supports one of these hypotheses

[12] Stanford (2006) emphasizes the role of unconsidered alternatives as a source of underdetermination. I agree that when there are unconsidered alternatives this undermines claims of inductive support. But there are solutions to this problem-- as emphasized above it is often possible to systematically generate and test alternatives or to produce general considerations showing that they do not exist.

and excludes the others. IOE is often a realistic goal. Of course this does not show that IBE is a mistaken strategy, but it does show that there is an alternative to it

5. Formulations of IBE

Turning now to a more detailed look at IBE, there are a variety of different formulations in the philosophical literature. (See, e.g. Carbrera, 2020, and the various formulations quoted in Douven, 2017.) These differ mainly in the strength of the conclusion that is taken to be warranted when the potentially best explanation is identified-- it may be contended that this explanation should be "accepted" or that we should "infer" to its truth but it may instead be claimed, more weakly, merely that this it has stronger inducive support (from the evidence that it would explain if true) than the alternatives. In order to streamline discussion I will generally adopt this weaker formulation and to avoid unnecessary verbiage will often just use the locution "strong inducive support" to describe the conclusion of an IBE. Also I will follow recent discussion in adding the requirement that for an IBE to be justified, the best explanation must be one that is satisfactory or good enough, thus avoiding van Fraassen's "best of a bad lot" objection. "Best explanation" should thus be interpreted in a way that incorporates this requirement.

Finally my focus in what follows will be on IBE understood as a normative thesis-- that is as a claim that a certain kind of inference is justified, where the standards of justification are (roughly) those that are generally accepted in science. I do not doubt that as a descriptive matter, people sometimes reason in accord with IBE[13].

One possible version of IBE (arguably endorsed by Harman, 1965) claims that *e* inductively supports *h if and only if h* is the best potential explanation of *e*. It is widely recognized, however, that the "only if" part of this biconditional (that is, *e* inductively supports *h*--> *h* explains *e*) is implausible: the occurrence of an effect can be strong evidence for the occurrence of a cause but the former does not explain the latter, the occurrence of one effect of a common cause can be evidence for the occurrence of another effect of that cause without the former explaining the latter and so on[14]. I will thus assume in what follows that IBE requires only a best potential explanation--> inductive support connection; that is, as indicated above, if *h* is the best

[13] Arguably this is sometimes part of what goes on when people adopt conspiracy theories. As a mischiveous aside, I note that some academic disciplines may be more susceptible to IBE-type reasoning than others-- one thinks of evolutionary psychology and portions of economics.

[14] See Lipton 2004 for similar observations. A number of other common strategies for assessing inductive support--e.g calibration of measurement devices-- seem to have little to do with IBE and to be more naturally viewed in terms of establishing reliability. If I measure the length of a table once with a measuring stick and get 57.3 inches as a result (R), it would be naive, to say the least, to argue that the best explanation of R is that length of the table really is 57.3 inches and hence that this conclusion is true. (This argument could be used to "show" the correctness of pretty much *any* aribitrary measurement.) Instead what matters is the reliability of my measurent and this is something that can be checked by re-measuring, using a different measurement device and so on.

potential explanation of *e* among some set of competitors and also a sufficiently good explanation, then *e* provides strong inductive support for *h*.

Given this conception of IBE one might think that a natural way of evaluating it is to formulate clear (ideally formal) criteria for inductive support and also for when a potential explanation is "best" and to then investigate the connections, if any, between these two. There are a number of accounts of inductive support (or of notions in the same neighborhood having to do with hypothesis testing[15], learning strategies from machine learning etc.) that might be employed in this way and of course there are also many "models" of explanation in the philosophical literature.

However, for the most part this has not been the path taken by defenders of IBE. Early formulations (such as Lipton, 1991) did not appeal to explicit accounts of inductive support at all. More recently (including the revised version of Lipton, 2004) the default assumption for the "inductive" side of things has been some version of Bayesianism with researchers worrying about whether this is compatible with IBE and, with some exceptions, concluding that it is. On the explanation side, advocates have largely avoided explicit discussion of the relation between IBE and standard models of explanation, opting to bypass these and instead to formulate the notion of best explanation in terms of a list of explanatory virtues, where (depending on the author) these include simplicity, unification, "mechanism", agreement with background knowledge.

For example, Cabrera writes (2017) :

> in my view, whether H1 constitutes an explanation according to one of the extant philosophical models— e.g. the Deductive-Nomological model (Hempel and Oppenheim 1948), the Statistical Relevance model (Salmon 1971), the Unificationist model (Friedman; 1974; Kitcher 1989), the Causal model (Salmon 1984; Woodward 2003), etc.— does not seem to do any real justificatory work. Rather, the feature that justifies any application of IBE is that the hypothesis does well with respect to the various virtues listed above. Presumably, any hypothesis that does well with respect to those virtues will be confirmed in accordance with IBE.

One likely reason for this focus on explanatory virtues is that, as we shall see, many of the standard models do not contain the resources needed to make the kinds of discriminations regarding explanatory goodness that IBE seems to require. It might be thought that this is a limitation/defect of the standard models but, as I will argue, it is far from clear that this is the case. On the contrary, it appears that IBE requires assumptions about how to rank potential

[15] Recall that the orthodox interpretation of standard frequentist statistics is that it is *not* a theory of inductive support but rather a theory that makes recommendations about acceptance and rejection of hypotheses on the basis of information about error characteristics. Formal learning theories of the sort described in Schulte, 2022 are also not theories of inductive support but rather theories about learning strategies. On the other hand, Bayesianism and likelihoodism, as well accounts of HD confirmation, are theories of inductive support. Despite these differences, in order to be ecumenical I will treat all of these as possible explications of IBE.

explanations that are difficult to motivate, quite independently of whether these are connected to inductive support in the way that IBE claims. Indeed, in some cases, it is unclear that the claimed virtues are explanatory virtues at all, at least always or even for the most part. Moreover, as I will argue, it is not clear that we need to appeal to the virtues to make defensible comparisons among candidate explanations. I will suggest instead that a much "thinner" notion of explanation and explanatory goodness (based on the ability of a potential explanation to provide answers to a range of what if things had been different question, as described in Woodward, 2003, Hitchcock and Woodward, 2003) gives us all that can reasonably required for such comparisons. With the qualification concerning tuning mentioned above , this conception of explanation does not license anything like IBE. In other cases, although proposed virtues are indeed virtues it is implausible that they are connected to inductive support in the way described by IBE.

Although a number of the standard models of explanation do not not provide resources that support IBE, it is also true that several of the standard models naturally connect to explanatory virtues invoked by defenders of IBE. Perhaps the best known example involves the unificationist models developed by Friedman (1974) and Kitcher (1989) which connect to the proposed explanatory virtue of unification. However, extensive discussion of these models has shown that it is difficult to formulate a notion of unification that is connected to explanation in a plausible way. This isn't just the usual difficulty of formulating precise necessary and sufficient conditions for some concept of philosophical interest; rather there are much deeper difficulties. Some candidate theories/ hypotheses that meet the criteria in unificationist models arguably do have explanatory import but other realistic examples meeting those criteria do not and we have no clear story about the difference[16]. There is also substantial disagreement in different areas of science about what unification involves and the the extent to which it is a virtue-- many economists tout the unifying explanatory virtues of a adopting a rational choice framework for all of social science; while other social scientists (e.g. many sociologists) are do not regard unification as an important desideratum at all. So unification is a contested notion, both in terms of its value and how it should be characterized. For this reason, the strategy of invoking unification as an explanatory virtue while attempting to bypass the standard models of explanation in which this is discussed is a highly problematic strategy. Instead there are compelling reasons for advocates of IBE who regard unification as an explanatory virtue to engage with philosophy of science literature on this topic and extract lessons from its successes and failures. A similar point holds for the vast literature on simplicity as a virtue of hypotheses, explanatory or otherwise.

Turning now to the "inductive" side of IBE, here matters are also less than satisfactory. As noted above, currently Bayesianism is the most common framework for thinking about support, both within the literature on IBE and more generally. Although there are alternative frameworks whose connection with explanatory considerations might be explored in order to avoid making my discussion even more complicated than it already is, I will largely follow the literature in assuming that inductive side of IBE should be understood within a Bayesian framework. I will assume, though, that this framework is most helpful when certain conditions obtain: when one has an exhaustive list of mutually exclusive suitably specific alternative hypotheses (thus no reliance on a "catch-all", or "none of the above" alternative, since this is

[16] For discussion, see Woodward, 2003 and Morrison (2000).

necessary for clearly defined likelihoods and also when results of the analysis do not depend on arbitrary assumptions about priors.

6. Some additional background assumptions

As should be clear from my discussion in Section 4, I assume, along with most of the literature on explanation, that the explanans of a successful explanation must be true or at least possess some truthlike property-- e.g it must be, in relevant respects, a good approximation, or "effective" within its domain of application etc. This excludes views according to which the explanans of a successful explanation can be radically false [17]. This is of course an assumption that defenders of IBE should accept. If radically false hypotheses can successfully explain, it will not make sense to infer to the truth of such hypotheses or regard them as inductively well supported on the basis of their explanatory credentials.

I take it to follow from this assumption that a theory or hypothesis which is a good approximation or effective within a domain can be used to explain, even if there is a successor theory which is even more accurate. (In effect I assumed this in Section 4). Again this is an amendment that advocates of IBE should be happy to accept. For one thing, the literature on IBE makes detailed use of examples in which the theories inferred to (e.g., Newtonian theory) are merely effective in this sense. Indeed, given that most or virtually all known theories and hypotheses are good approximations rather than exact truths, an exact truth requirement on explanation would render IBE in applicable to almost all cases.

A closely related issue has to do with what counts as alternative (or "competing") potential explanations for the purposes of IBE. (This will turn out to be important.) Suppose one thinks that General Relativity (GR) provides "better" explanations than Newtonian Theory (N) (or at least that GR explains phenomena that N does not explain. Consider explananda that we ordinarily think are explained by N (the motion of the planets, the tides etc.) I assume that the advocate of IBE will not want to argue that if GR provides a better explanation than N of these and other explananda, it follows that these explananda provide evidential support for GR but not for N. In other words, the advocate of IBE should avoid concluding that N and GR are potential explanations that compete in such a way that IBE-mediated inferences lead to the conclusion that there is evidence supporting GR but not N (since GR rather than N should be inferred to as the "best explanation"). I think that the most natural way of achieving this goal is to say that when considering evidence and explananda that are within the scope of two theories, one of which is a special or limiting case of the other, as is the case with N and GR, we should not treat

[17] As sometimes seems to be suggested in Rice, (2021). But Rice also emphasizes the role of extracting true counterfactuals about the target system in successful explanation which sounds much closer to my own view. From my point of view the crucial point bearing on explanation is that a model or set of hypotheses can capture correct information about dependency relations (true countefactuals) even if not everything in the model is literally true-- it is the appropriate counterfactuals and dependency relations that do the explanatory work and that need to be true or approximately so. I agree with Rice that some elements of a model can be false and known to be false consistently with this requirement, as when a model talks about an infinite population of rabbits.

these theories as competing alternative explanations *with respect to those explananda.* In other words we should agree that both N and GR can explain some of the same explananda/phenomena and that IBE should not be understood in such a way that it licenses inference to GR only but not N[18].

As an additional illustration of this assumption, consider inferring to the electroweak theory of the electromagnetic and weak interactions in circumstances in which the available evidence *e* just supports quantum electrodynamics (QED). Assume QED is an implication of the electroweak theory, with this theory providing a potential unifying explanation of both forces . In other words, we are supposing that don't yet have evidence regarding whether the electroweak theory makes correct predictions regarding the weak force, although the electroweak theory is an excellent potential explanation of phenomena involving that force and we do know that the electroweak theory implies QED for which we have strong evidence. Presumably we we don't want to conclude that *e* provides inductive support for the electroweak theory but not for QED on the grounds that the electroweak theory if true would provide better explanations than QED and it is only this better explanation that is inductively supported. QED and the electroweak theory are not competitors in this way and in any case, by hypothesis, *e* supports QED but not the electroweak theory [19].

As yet another illustration consider the following two hypotheses: *h1*-- ingestion of aspirin relieves pain, *h2* ingestion of aspirin relieves pain via a mechanism that involves ... Suppose (as was the case for a long time), we have strong evidence *e* for *h1* but (at least as we ordinarily think about such matters) no evidence about the mechanism of aspirin. If *h2* would if true provide a better explanation of *e* than *h1* (the grounds being that explanations that describe mechanisms are better than those that do not), are we entitled to infer to *h2* via IBE? Are we entitled to infer that our evidence better supports *h2* than *h1*? The idea that *e* supports *h2* at all seems strange enough and the idea that *e* supports *h2 rather than h1* even stranger. We can avoid the latter conclusion (although not the formet) by not treating *h2* and *h1* as competitors.

What then is it for two potential explanations *h1* and *h2* to be competitors ? I will assume that a sufficient condition is that *h1* and *h2* make substantively different predictions about phenomena that fall within the intended domain of both. For example, N and MOND (modified Newtonian gravity-- a theory that proposes a modification to N to account for some observed properties of galaxies, in a domain where N is widely thought to be applicable) are competitors, since MOND proposes a different gravitational force law than N. A hypothesis *h1* according to which *X*

---

[18] Just to clarify: my own view is that potential explanatory considerations by themselves do not license inference to either GR or N-- my claim here is that insofar as there is a case for IBE it needs to be restricted in some way to avoid treating GR and N as competitors, only one of which can receive evidential support via IBE. If such a restriction is difficult to formulate and motivate within the framework of IBE, so much the worse for that framework.

[19] Indeed in the envisioned circumstances it seems odd to suppose that the electroweak theory is inductively well supported at all. But that is a different problem for IBE--- a problem for the view that the most potentially unifying explanation is the one that is best supported.

causes *Y* via a single direct route from *X* to *Y* is a competitor to a hypothesis *h2* according to which *X* does not cause *Y* and the correlation between them is entirely do to a third variable *Z*, since the first implies that some intervention on *X* will change *Y* and the second denies this. This notion of competition fits with my own view (suggested above) according to which when choosing among competing explanatory hypotheses one must provide evidence that favors one and excludes or undermines the other, but I believe it also should be congenial to advocates of IBE, for reasons described above.

7. IBE and some standard models of explanation

I turn now an exploration of the relationship between IBE and some of standard models of explanation (DN, IS, SR, causal mechanical) found in the philosophical literature. I noted above that advocates of IBE have frequently claimed that it does not require or depend on any particular model of explanation of the standard sort. I have already suggested that such claims are misleading. In fact, the different models stand in a variety of different relations to IBE, some potentially supportive, and others not. In particular some models fit rather naturally with IBE in the sense that they embody a close connection between potential explanatory goodness and some conception of inductive support. However, as we shall see, these models involve unsatisfactory conceptions of explanation or inductive support. Other models do not provide the resources to support IBE, either because they do they do not license the kind of ranking of potential explanations that IBE requires at all or because, to the extent they allow such rankings, the rankings either have implausible consequences or they cannot be linked to inductive support in the way that IBE requires.

Consider first the DN model. If *h* is a potential DN explanation of *e*, then *e* provides hypothetico-deductive (HD) confirmation of *h*. Thus, if the DN and HD models are accepted, we can reason from the fact *h* is a potential explanation of *e* to the conclusion that *e* provides at least some inductive support for *h*. On the other hand, the claim that *h* is a "better" explanation than other potential explanations of *e* plays no role in the inference just described. Indeed, the DN model seems to provide no basis for concluding that any one among the possible explanations that if true would satisfy the DN requirements is better than any other, just as the HD model does not support judgments of stronger or weaker inductive support. This is because the DN model does not discriminate further among explanations as long as the basic requirement of nomic derivability is satisfied. Of course it is also true that the DN model and the HD model are subject to well-known and arguably fatal difficulties. Nonetheless this is a case in which there is a straightforward connection between well-specified conceptions of explanation and of confirmation, such that there is no mystery about why a potential DN model is HD confirmed.

Hempel's IS model of statistical explanation (Hempel, 1965) introduces a comparative element that is missing from the DN/HD connection. Take the IS model to hold that *h* (assumed to include a statistical law) is a better potential probabilistic explanation of *e* the higher probability of *e* would be if *h* was true, with the restriction that *h* is no explanation at all if $Pr\ (e/h) < 1/2$. Within a likelihoodist framework $Pr\ (e/h)$ is also a measure of the inductive support that e provides for *h* and within this framework the hypothesis *h** for which $Pr\ (e/h^*)$ has the highest likelihood (maximum likelihood) is taken to the hypothesis that is best supported by *e*. Thus the better the IS explanation that *h* if true would provide for *e*, the better supported *h*

is by *e*. This provides a straightforward connection between a model of what makes an explanation "best" and a measure of inductive support, with the connection taking a form that supports IBE. Of course the IS model has attracted considerable criticism and it is restricted in scope since it applies only to hypotheses that provide statistical explanations of individual outcomes. Nonetheless it is worth noting that the IS model is a better fit with IBE than alternative models of the statistical explanation of individual events, such as Salmon's SR model (1971), as noted immediately below.

Consider next accounts of the statistical explanation of individual events which hold that the goodness of the explanation that some statistical hypothesis h provides for e is independent of the probability value h assigns to e. (Salmon, Jeffrey, Railton) According to these views, as long as h assigns the correct probability value to e, it explains e just as well if the assigned probability value is low than if it is high. Prasetya (2021) claims that such accounts are inconsistent with IBE. I don't think this is quite right, at least if inconsistency means that the accounts imply that there are cases in which *h* is a best explanation of *e* and yet inference to *h* on the basis of *e* is not warranted or *h* is not well inductively supported by *e*. Rather the problem is that the accounts deny that there is a notion of a best statistical explanation of a sort that allows us to apply IBE. Suppose that a radioactive decay event *e* is observed and consider two potential explanations: hypothesis, *h1*, according to which Pr (*e/h1*) is high and *h2* according to which Pr *(e/h2)* is low. According to the accounts under consideration, if *h1* was true it would provide an equally good explanation of *e* as *h2* would if it were true. Since neither *h1* nor *h2* is a best explanation in comparison with the other, we cannot apply IBE to argue that one hypothesis is better inductively supported than the other. What is true is that, as Prasetya puts it elsewhere, there is lack of "positive correlation" between explanatory goodness and inductive support in this example: If we assume that support is measured by likelihood, *e* provides stronger inductive support for *h1* than for *h2* even though *h1* and *h2* are equally good potential explanations of *e*. This is a counterexample to what we called above the "only if" formulation of IBE-- the formulation that says that if *e* provides stronger support for *h1* than for *h2*, *h1* must be a better or best potential explanation of *e*-- rather than the "if" formulation of IBE (best explanation--> strongest inductive support).

I turn next to Salmon's causal mechanical (CM) model (Salmon, 1984) which Prasetya suggests is "(merely) compatible with IBE" or "neutral" with respect to it. Of course Salmon explicitly rejects IBE (see, Salmon, 2021) but nonetheless one can ask whether some version of the CM model or perhaps an extension of it can be consistently combined with IBE. One immediate problem (paralleling the issue with the SR model) is that while Salmon's discussion lays out conditions for something to count as a CM explanation, it says little or nothing about what makes one potential CM explanation "better" than another. A defender of the CM model thus might hold that there is no basis for such judgments: while two proposed CM explanations can differ in that one makes only true claims and the other makes false claims and this provides grounds for preferring the former, there are no further grounds for saying that one would provide a better explanation if true. Thus there is no notion of a better/best potential explanation

available in the CM model which can be used by IBE[20]. This does not show the CM model to be inconsistent with IBE (for the same reason that the SR model is not inconsistent with IBE) but it would certainly suggest that the CM model is not friendly territory for IBE.

If one wishes to use the CM model to make comparative assessments of potential explanatory goodness, the move that seems most consistent with the spirit of that model is this: hold that potential explanations that if true would provide more rather than less causally relevant detail (with causal relevance understood in accord with the CM model) are to that extent better potential explanations. Guided by this idea, we might consider a version of IBE incorporating this "comparative" understanding of CM and according to which we are entitled to infer to the potential CM explanation which would, if true, provide the most CM relevant detail regarding the target explanandum *e*. However, this proposal seems a non- starter -- it apparently tells us to infer to the most detailed hypothesis possible or conceivable (assuming it makes sense to suppose that there is such a hypothesis) as long as that this hypothesis would if true provide more CM relevant detail for e than alternatives. For example, in a context in which we have observed via an RCT that a drug is effective in producing recovery--this is our evidence/explanandum *e*-- from an illness but know nothing more, the proposal recommends (if it recommends anything at all) inference to some highly speculative but extremely detailed hypothesis about the unknown mechanism of action of the drug as long as if true that the hypothesis is the most detailed potential CM explanation of e. Needless to say, this is not considered good scientific methodology. Given the observed efficacy of the drug, figuring out its mechanism of action is certainly valuable for a number of a reasons, but discovering this is a separate, additional problem, requiring additional evidence that distinguishes among alternative hypotheses regarding the drug mechanism.

It is one thing to claim (although this is controversial and subject to different interpretations) that among explanations with true assumptions, those that provide more relevant detail are (always) better[21]. It is quite another matter to claim that we should regard hypotheses which if true would provide more relevant detail as better inductively supported in virtue of this fact. It is no wonder that, given his views about explanation, Salmon was not a fan of IBE.

Very similar conclusions apply to the role of mechanistic information more generally in connection with IBE. There is of course a large philosophical literature on mechanisms and mechanistic explanation and "mechanism" (that is the provision of mechanistic information) is on many lists of explanatory virtues that reference IBE. Information about mechanisms is unquestionably valuable (and often explanatory) but again it seems misguided to infer from this that a hypothesis that if true would supply mechanistic information about an explanandum is for that reason better inductively supported than an alternative hypothesis that does not provide such information . Information (that is truths) about mechanism is a virtue (at least in part because

[20] This illustrates our general point that for a model of explanation to fit at all with IBE the model must at least allow us to make sense of the idea that some potential explanations of can be better than other explanations of the same explananda

[21] See e.g. Craver and Kaplan, 2020 and the references discussed there.

mechanistic information is or often is explanatory) that has to do with valued content but which does not directly connect to inductive support.

8. The w-condition account of explanation and its relation to IBE.

I turn now to a discussion of the relation between the account of explanation (the w-condition account) developed in Woodward, 2003, Hitchcock and Woodward 2003 and IBE, as well as the implications of that account for some candidate explanatory virtues. Since that account may be less well-known than some of the models discussed above and since to my knowledge there has been no discussion of its implications for IBE, I want to spend some time drawing out some of those consequences.

The key idea of the w- account is that explanation works by correctly answering what - if-things- had-been-different questions (hereafter w-questions)-- that is by describing how if the factors cited in explanans were to be different in various ways, the target explanandum would change. Put differently, the idea is that a successful explanation correctly describes patterns of dependence between the factors cited in the explanans and variations in the explanandum-- it tells us (at least in some respects) what factors the explanandum depends on and describes the dependence relation obtaining between explanans and explanandum . The conditionals associated with these answers to w-questions are understood as "interventionist" or some other form of non-backtracking counterfactual. In keeping my remarks above, "correct" description of patterns of dependence should be understood in a way that includes "effectively" correct descriptions or those that are good approximations[22]. When a dependence relation holds under some range of interventions or changes in background conditions it is said to be *invariant* under these. Generally speaking when two generalizations G1 and G2 are related in such a way that the conditions under which G2 is invariant is a proper subset of the conditions under which G1 is invariant, G1 can be used to answer more w-questions than others

As outlined in Hitchcock and Woodward, 2003, this account licenses certain kinds of comparisons among potential explanations but not others. Here are some examples -- the list is not meant to be exhaustive. A potential explanation EX1 might if true answer some set of w-

[22] Also, although I lack the space to argue for this claim here, what matters for explanatory import within the w-account is getting the dependency relations effectively correct and nothing more. (It is the dependency relations that do the explanatory work) In particular, as suggested previously, it is possible for a theory to get the ontology "wrong" in some deep way (at least when judged from the perspective of a successor theory) but nonetheless get dependency relations largely right and count as explanatory for this reason. For example, one might think that Newtonian gravitational theory is fundamentally mistaken in its ontological claims (there is no such thing as a gravitational "force" and gravitational phenomena reflect spacetime structure). Nonetheless the w-account will count Newtonian theory as explanatory insofar as it correctly describes how the motions of bodies depends on the masses of and distances to other bodies. As a number of writers have observed, what tends to be preserved across theory changes are approximate dependency relations but not the ontology that goes with these. This is a reason for taking the dependency relations to be central to successful explanation. For more in defense of this idea see Woodward, 2023.

questions *w1* and a second explanation EX2 may answer all of the questions *w1* and more besides-- so that *w1* is a proper subset of the w- questions *w2* answered by Ex2. (As noted above this wil typically be because the generalizations in EX2 have a greater range of invariance than those in EX1. For example, understood in the way described above which countenances "effective" theories as explanatory, GR answers the w-questions answered by Newtonian gravitational theory and answers more questions besides. Another possibility is that potential explanation EX1 would if true explain some rather qualitative or coarse-grained characterized features *E1* of a phenomenon while if true EX2 would explain more fined grained or quantitative features *E2* of that phenomenon, where *E2* implies *E1* but not vice-versa. For example, EX1 might be a potential explanation of the qualitative fact that subjects tend to recall more recent items in a memory test more accurately than less recent items and EX2 might be a potential explanation of quantitative features of patterns of recall. Yet another possibility is that a potential explanation EX1 purports to identify some of the factors on which an outcome *E* depends but not all of these and an alternative potential explanation EX2 purports to identify more of these: EX1 claims the occurrence D of a disease depends on whether one has been exposed to a pathogen but says nothing more about other factors on which D depends. Ex2 if true provides this information about the pathogen and additional information about what D depends on as well--for example, relevant information about the state of the patient's immune system. Note that in this case too, the generalization figuring in EX2 will have a greater range of invariance than that figuring in EX1.

In each of these cases, there is an obvious sense in which. EX2 is potentially more informative about dependency relations than EX 1 and in this respect might be judged more valuable qua explanation. Or at least it might be judged that it is worth knowing whether EX2 holds in addition to knowing whether EX1holds. This is so whether or not we want to conclude that EX2 would if true provide a "better" explanation of what EX1 explains than EX1 would provide if true. Note however that, as argued above, in these cases EX1 and EX2 do not compete in the sense that if we judge that one of these explanations is correct, we must judge the other as incorrect or reject it. Although GR correctly answers more w-questions than N, as argued above N correctly answers some w-questions and in a way that is consistent (up to some very good level of approximation) with the answers provided by GR. Similarly an explanation that cites infection with a particular pathogen as the explanation of a patients disease is not wrong or mistaken merely because there is another correct explanation that cites both the pathogen and the state of the patient's immune system.

As argued above, in cases of this sort, the defender of IBE should not endorse a version of that doctrine that implies we should accept the EX2 explanations and reject the EX1 explanations on the grounds that the former would if true be best in comparison with the latter if true. The EX1s should be accepted as explanatory even if they are not best explanations . In addition IBE should not be understood in such a way that only the EX2s can be inductively inferred to since only these provide best explanations. Again, not thinking of explanations like the EX2s and the EX1s

as competing candidate explanations to which IBE is to be applied is a natural way of implementing these restrictions.[23] .

The upshot of our discussion so far is that is that the kinds of comparisons of explanations described above (and the kinds that are supported by the w-account) don't license inferring to one potential explanation instead of another in the sense that they provide grounds for accepting one and rejecting the other. Similarly they don't license concluding that one potential explanation is better inductively supported than another on the grounds that it would if true provide a better explanation. Within the w-account for different potential explanations to compete in the sense that acceptance of one requires rejection of the other, they must make inconsistent claims about dependency relations or imply inconsistent answers to the same w-questions or at least have inconsistent implications about possible evidence. Moreover, when potential explanations do compete, the w-account (at least as developed so far) provides no basis for ranking them as to their potential explanatory goodness.

These conclusions should not be surprising. As suggested above, IBE is a method which is most naturally applied to competing hypotheses and the kinds of explanatory comparisons of the EX2s and EX1s considered above don't involve hypotheses that compete in the relevant sense. What we need to apply IBE are explanatory considerations or virtues that can guide choice among inconsistent hypotheses that agree regarding the available evidence but disagree elsewhere. Put differently, considerations relevant to explanatory assessment divide into at least two classes--those that do not support choices among competing hypotheses and those that do. IBE requires the latter.

9. Simplicity and Unification as Explanatory Virtues

---

[23] These judgments about evidence are supported by standard theories of confirmation, including Bayesianism. Suppose that we use P(H/E)/ P (H) as our measure of the inductive support provided by E for H, and let E be the evidence associated with EX1 in the above examples (evidence for NG, the pathogen as a cause etc). Suppose for simplicity that both (EX2) and (EX1) entail E Then

$$\frac{P(EX2/E)/P(EX2)}{P(EX1/E)/P(EX1)} = \frac{P(E/EX2)/P(E)}{P(E/EX1)/P(E)} = \frac{P(E/EX2)}{P(E/EX1)} = 1$$

This just to say that the evidence E does not discriminate between Ex2 and EX1. This is not at all surprising. For E to discriminate there needs to be a difference in the likelihoods P(E/ EX2) and P(E/ EX1) and *ex hypothesi* there is no difference.

There certainly are proposed explanatory virtues fall into the latter class-- that involve comparisons between different explanations that are inconsistent-- telling us to prefer the explanation that most completely satisfies the virtue in question and which in turn might be linked to evidential support via IBE. In addition to the virtue associated with IS explanations which licenses such comparisons and is discussed above, two of the most prominent possibilities are simplicity and unification. For example, if simplicity (or unification) is taken to be an explanatory virtue, then given two competing explanations of the same body of evidence, if one is simpler (more unified than) the other and simple (unified) enough this presumably implies that, ceteris paribus, the more simpler (more unified one) would, if true , be the better explanation. And, assuming IBE, it follows that the simpler (more unified) explanation is the one that is better inductively supported[24].

The idea that simplicity and unification are explanatory virtues is so often repeated (by some) in the philosophy of science literature discussions that it may seem a kind of heresy to question it. Focusing first on simplicity, consider the following example: Suppose our evidence *E* is that *X* and *Y* are correlated (and this is at present all of the relevant evidence). Suppose we can eliminate the possibility that *Y* causes *X*, perhaps on the basis of time order considerations. Consider the following two candidate explanations for *E*: h1: *X* causes *Y*, *h2*: there is a third variable *Z* that is a common cause of *X* and *Y*. (Here in accord with my previous discussion I assume *h1-h2* are mutually exclusive and exhaustive[25])

A very plausible interpretation of "simplicity" in this context yields the judgment that *h2* is less simple than *h1*: *h2* postulates an additional variable Z and two causal connections, one from *Z* to *X* and one from *Z* to *Y*, while *h1* postulates just two variables and one causal connection. Thus insofar as simplicity is an explanatory virtue and simpler explanations are better explanations (and assuming, as seems plausible, that both *h1* and *h2* are "good enough" explanations), it seems we should conclude that *h1* is the best explanations of *E* and to the extent we are willing to follow IBE, that *E* provides better support for *h1* than *E* provides for *h2*.

---

[24] Some may think it possible for there to be two different candidate explanations, one more unified (or more simple) than the other, but with exactly the same implications for all possible evidence-- with the more unfiled one hence providing a better explanation and having stronger inductive support. Per my earlier discussion, I think that there are no clear cases of this. But even if there are such cases, there are many other clear cases of explanations (like the one considered below) , one more unified or simpler than the other, that compete in the sense that they have implications for other possible evidence that might be used to discriminate between them. If unification or simplicity are explanatory virtues, they ought to be applicable to this second class of cases.

[25] Suppose you think that as formulated the hypotheses are not mutually exclusive; instead the first hypothesis should be formulated as X causes Y and Y has no other causes and similarly for the other hypothesis (Z is the only cause of X and Y) . It is not clear that this will make any difference to the comparative judgments in the example-- it still seems that Z causes X and Z causes Y and X and Y have no other causes comes out as less simple than the other alternative.

This conclusion may seem plausible to some metaphysically inclined readers but in my view it is very strongly at odds with scientific practice. Even if our present evidence is just *E*, good scientific practice is to take seriously the possibility that there may be a third confounding variable--a *Z*-- (or many of these) which is responsible for the correlation between *X* and *Y* and to take steps to explore that possibility and if possible to rule it out. There are many devices for doing this: as noted earlier, one is a randomized experiment in which (assuming *X* and *Y* are binary) the values of *X*s and *Y*s are randomly allocated to a treatment and a control group, where the randomization makes it unlikely that there is such a third confounding variable which is correlated with *X* and *Y*. Alternatively, in a non-experimental context, if a *Z* is discovered such that X and Y are independent conditional on *Z*, this can be suggestive although not conclusive evidence that *h2* is correct. Additional observational evidence in conjunction with background assumptions like the Causal Markov condition and faithfulness (in the sense of Spirtes et al., 2000) can further support (or undermine) *h2*. The important point for our purposes is that scientific practice does not regard it as legitimate to exclude or downplay the possibility of a confounding common cause (or to regard this assumption as less well evidentially supported than *h1*) just on the grounds that *h2* is less simple than *h1*.

This example illustrates the general idea defended above: when there are alternative possible explanations *hi* all of which would if true explain some evidence/explanandum *E* (where this all the relevant evidence we have at present), at least in many cases it is *not* good scientific practice to proceed by assessing which of these alternative would if true provide the best explanation of *E* and then concluding on that basis that this is the hypothesis best supported by *E*. Rather, good scientific practice is to one looks for additional evidence besides *E* that allows one to discriminate among these alternative hypotheses. Often it will be possible to discover such additional discriminating evidence but if it is not we are not entitled to take one of the hypotheses to be true or correct just on the basis of potential explanatory considerations.

So far I've focused on the issue of whether the supposed greater simplicity of *h1* provides grounds for regarding it as better supported than *h2* by *E*. However, we can also use the example to raise questions about whether simplicity is an explanatory virtue , at least in the way advocates of IBE claim. Suppose, as obviously possible, that when we get additional evidence, *h2* turns out to be the true or correct explanation of *E*. Do we then conclude that although *h2* is the correct explanation, it nonetheless provides a "worse" explanation than *h1* would have provided, had it been correct? It isn't just that this sounds odd-- although it does. It is hard to see what might be a non-question-begging basis for this judgment. After all, in the case we are envisioning *h2* completely accounts for the correlation *E* between *X* and *Y* and, by hypothesis, also explains or at least is supported by whatever additional evidence we have obtained.

A more plausible assessment is that *h1*, and and *h2* are both equally good potential explanations of *E*. This is the judgment supported by the w-account of explanation-- under the above scenario, each if correct would provide a full account of the factors on which *E* depends. On this view, explanations can differ in the extent that they describe what an explanandum or set of these depend on, and they can also differ in which explananda they cite dependency relations for, but when a potential explanation of *e* would if true, fully describe the factors on which *e* depends (as we are assuming *h2* does) , there are no legitimate further ground for claiming that the

explanation is more or less "good" depending on how simple it is. Thus simplicity when understood as above is not an explanatory virtue at all.

This claim may evoke the following response: Although simplicity when understood as it is in the example above may not be an explanatory virtue, this is not the appropriate understanding of the notion for purposes of IBE. Instead some other notion is the appropriate one. I am not unsympathetic to this response (see my remarks on tuning in Section 11), but of course it puts the onus on the advocate of IBE to distinguish the "appropriate" notion of simplicity from others and to explain why it is linked to inductive support in the way claimed. If not all varieties of simplicity support a judgments of explanatory goodness, the advocate of IBE owes us an account of what distinguishes the varieties that do from those that do not. This reinforces the point, made earlier, that the failure of advocates of IBE to engage with the extensive literature on simplicity is misguided[26].

I think that similar points apply to another alleged explanatory virtue-- unification. Suppose, as before, that *X* and *Y* are correlated (*e*) but now the only two hypotheses that are consistent with background knowledge are *h2*, understood as above, which postulates a single common cause *Z* that accounts for the correlation and *h3* that says instead that there are two common causes, *U* and *W*, both distinct from *Z*, that contribute to, and together, fully account for the correlation. (Fully account in the sense that conditional on *U* and *W*, *X* and *Y* are independent.) We noted above that the notion of unification is unclear in important respects (or at least it has resisted any kind of content general clarification) and that there seem to be different varieties of unification, some of which may be more relevant to explanation than others. However, if explanatory unification means anything definite at all, *h2* is surely more unified that *h3*. Nonetheless it seems highly problematic to infer, on the basis of IBE, that *e* provides stronger evidential support for *h2* than *h3*. It also seems dubious that *h2* if true would provide a better explanation than h3 would if it was true. Instead, since by hypothesis both if true would equally capture the dependency relations relevant to *e*, it is more reasonable to conclude that each would be an equally good explanation if true. Again, this is the judgment reached by the w-account.

As with simplicity, advocates of IBE might respond that the sort of unification that is an explanatory virtue is different from the notion at work in the above example. Again, the onus is on the advocate to explain how to distinguish the IBE-friendly notions of unification.

As I remarked above, the suggestion that unification (at least without further qualification) is not the kind of explanatory virtue that licenses IBE may be met with incredulity. Isn't, e.g., Newtonian mechanics spectacularly successful as a unifying theory and isn't this why it is accepted? (See Lipton, 2004) In fact, things are more complicated than this simple gloss

---

[26] Sober (1994, 2015) has argued persuasively that when simplicity is legitimately invoked in hypothesis assessment, such assessments typically rest on background empirical claims about particular domains of investigation-- what he calls "subject matter specific (and *a posteriori*) considerations" (1994). Such empirical considerations make the use of simplicity in hypothesis assessment unmysterious but this doesn't support the claim that simplicity is a virtue bearing on potential explanatoriness-- rather the empirical considerations bear directly on whether the hypothesis in question is true or false.

suggests. It is of course true that Newtonian theory is a highly successful explanatory theory and that we value it in part for this reason. There is also an obvious sense in which it unifies celestial and terrestrial phenomena. It does not follow, however, that we should regard it as inductively well supported *because* if true, it would provide an explanatory unification. In fact as shown by several detailed recent studies (Smith, 2014, Harper, 2011) the evidential support for Newtonian theory -- both the support to which Newton appealed and subsequent evidence-- involves highly complex reasoning strategies that go beyond simple appeals to unifying explanatory power.

Let me add that it is true that the example above (involving one vs two common causes) is a very simple one and in a number of ways not representative of serious explanatory unifications in science which involve examples like Maxwell's unification of electricity and magnetism, the electroweak theory, the standard model in particle physics and so on. However, consideration of these examples reinforces my argument in the following way. In the richer examples, unlike the case considered above, one consequence of a successful unification is that it generates new and correct answers to w-questions that are not answered by previous theories. Thus, we can account for the appeal of the unified theories within the framework of the w-account just by invoking this feature. There is no need to invoke unification as an explanatory virtue over and above successful answering of additional w-questions.

My discussion of simplicity and unification as candidate explanatory virtues illustrates several more general points. First, unificationist accounts of explanation of the sort developed by Friedman and Kitcher will presumably judge the single common cause explanation a better explanation than the two common cause explanation while the w-theory account (along with a number of other causal theories of explanation) rejects this judgment. Thus in assessing IBE, we cannot bypass these accounts, but rather need to directly assess their merits.

A second point is this: once one decides to focus just on the alleged explanatory virtues, an obvious question arises: why invoke any notion of explanation at all (at least at a fundamental level) for the purposes to which IBE is to be put? After all, if what matters for the purposes of IBE is whether the hypothesis of interest possesses features like simplicity, unifying ability and so on, why not drop any reference to explanation and instead argue directly that the hypothesis that most exemplifies these features is the one that is best supported, regardless of whether the features are distinctively explanatory virtues? This idea is explicitly advocated in Elliott, 2021 and it seems to me that, dialectically speaking, it makes a great deal of sense, given the commitments of defenders of IBE. Why get embroiled in a discussion of how explanation is liked to unification and simplicity, if it is really just unification and simplicity that matter for IBE and inductive support?

The moral that I draw from the above examples is quite different. I take them to show that explanation has, so to speak, a life of its own: a concern with finding explanatory hypotheses is not the same thing as finding hypotheses that are simple, unifying or possess many of the other features on the standard lists of explanatory virtues. This is the path I have been following in this essay. Discovering hypotheses and theories that explain is a distinct goal of science and not merely a merely a means to other goals having to do with unification, simplicity and so on.

10. The Role of Background Knowledge

The proposed virtues considered so far (simplicity, unification etc.) at least fit coherently with the guiding question that underlies IBE (how well would this hypothesis explain if true?). However, other candidate virtues do not. Consider, for example, the suggestion (made by XX and YY, among others) that fit with background knowledge is among the explanatory virtues that can guide IBE. Suppose a potential explanatory hypothesis *h* is inconsistent with the totality of our background knowledge B. Since IBE is supposed to be applied to choose among candidate hypotheses which are not known to be false, if this background knowledge is genuinely knowledge, it seems that *h* should be excluded based on this consideration alone. In other words, h should be excluded because it is *false* and not on the basis of an assessment of how well it would explain if true.

Suppose instead the criterion is fit with background *beliefs* B where this allows for the possibility that some of those beliefs may be false. What situation or scenario are we then to envision when we ask, given background beliefs *B*, whether *h* if true would explain *e*? Are we supposed to consider a scenario in which both *h* and *B* are true, but *B* remains inconsistent with *h*? That's incoherent. A scenario in which *h* is not strictly inconsistent with B but disconfirmed or undermined by it? The IBE test involves assuming that *H* is *true.* If so, why should it matter whether our background beliefs seem to make *h* less likely? If this is the case, in the envisioned scenario, with *h* assumed true, our background knowledge is misleading regarding the status of *h* and so it is unclear why *B* should be taken to undermine the explanatory status of *h*. Alternatively we might imagine our background beliefs altered in such a way that they are consistent with *h* being true. But then consistency with our actual background beliefs (in unaltered form) seems to do no work and presumably it is consistency with our actual background beliefs that should matter for the assessment of *h*.

Suppose instead that *h is* supported by our background knowledge or beliefs. What additional boost in potential explanatoriness is provided by this support, given that in applying the counterfactual test associated with IBE, we are already assuming that *h* is true?

Consider, in this light, the following example: (**Note to reader: I'm trying to find the source of this example-- if anyone knows, would you please help me out?).** The scenario is that X returns home in the evening to find that a beer is missing from the six pack he purchased earlier. He considers two hypotheses-- h1: the beer was drunk by his wife who was home all day, h2: the beer was taken by aliens from outer space who broke into his refrigerator. Obviously we are much more inclined to accept h1 and the claim is that we infer to h1 via IBE on the grounds that h1 is a better potential explanation because it has a better fit with our background knowledge[27].

The alternative assessment that I favor is that the inference to *h1* in this example does not require IBE. Instead our background beliefs imply that it is overwhelmingly likely that *h2* is

[27] Kevin McCain has drawn my attention to a similar example in his 2019. In McCain's example, noodles are missing from the refrigerator and the alternative hypotheses are that they were eaten by one's roommate or that noodle thieves broke into the refrigerator and took only those. I think the same analysis applies to this example as to the missing beer.

false and we reject it for that reason, thus inferring *h1* as the only remaining possibility. In particular, we have no evidence that aliens exist, and apparently strong evidence that they do not visit earth (in the form of the absence of evidence where it would otherwise be expected) More specifically, if they had entered Xs house there would be particularly strong evidence for that -- the US has elaborate machinery for tracking intrusions into its airspace, including intrusions by alien space craft. For these reasons in evaluating the alien hypothesis we don't need to consider how well it would explain if true-- we reject it because we think we have extremely strong evidence against it.

Suppose, however, we do decide to evaluate *h2* by applying the IBE test. Now matters become more complicated. In applying that test, we need to take seriously the requirement that we are to ask is how well the alien hypothesis would explain if true. In other words, we are to suppose that is true that the aliens took a beer from Xs fridge and then ask, assuming that is the case, how well this would explain the absence of one of the beers. It seems to me that in this case the explanation would be a good one-- or at least that it might be made good with suitable elaboration. Indeed it seems very strange to both suppose that it is true that the aliens took the beer (which implies that no one else did, the beer did not spontaneously vanish etc.) and to suppose that in these circumstances this is some how an inferior defective or non-best explanation of the absence of the beer. I think that the temptation to think otherwise comes from a failure to take the counterfactual test (would this be a good explanation if true?) seriously[28].

11. Tuning

I said above that although I do not think that there is a connection between potential explanatory considerations and inductive support of the sort envisioned in IBE, there is one important exception. Interestingly, this is a consideration that does not usually appear on the list of explanatory virtues discussed in the literature on IBE. It is discussed in various forms in portions of the philosophical literature on explanation and elsewhere but it is also not, for the most part, explicitly incorporated into the standard models of explanation. (The issue of how it might be incorporated into w- condition account is explored briefly below.) The consideration in question, perhaps best thought of as a family of considerations, has to do with *tuning*. Ceteris paribus, a hypothesis which is in some way too finely tuned, is thought of as explanatorily defective. Moreover, finely tuned hypotheses are often thought to lack (or to have only weak) inductive support. So it is arguable that tuning provides an IBE-like link: the fact that a potential explanation does not involve objectionable tuning can be a reason for thinking that it is inductively well-supported. Conversely, the fact that a candidate explanation requires elaborate

[28] To reiterate a point made earlier, nothing in my discussion is meant to imply that ft with background knowledge is unimportant in explanatory assessment. Rather my claim is that it is important because it bears directly on whether a potential explanation is true or correct. I'll add that it is arguable that it may also be relevant to something like the fertility of a potential explanation-- that is the extent to which it can be successfully elaborated, as discussed in section 12. When a potential explanation fails to fit with background knowledge that knowledge cannot provide guidance in how to elaborate or develop the explanation. It is a valuable feature of an explanation that it can be elaborated but the possibility of elaboration is not relevant to how well an explanation in its present, unelaborate state explains.

tuning may be taken to a reason for not regarding it as well-supported even if (given the tuning) it seems empirically adequate.

Objectionable tuning can take a number of different forms-- for example, a theory with lots of free parameters which are required to take very specific values to account for the available evidence in circumstances in which there are no independent reasons apart from accounting for the evidence for the parameters to take those values may be regarded as tuned to an objectionable degree and as providing less than fully satisfying explanations for this reason. A related form of tuning arises when a theory employs various free parameters and then must assume some precise relation between them (e.g., various combinations of values that cancel) in order to be empirically accurate, and where there is no further explanation for why that relation obtains. Violations of the faithfulness condition in the causal modeling literature which involve combinations of coefficient values that just happen to sum to zero are one example of this[29] and violations of "naturalness" conditions in high energy physics are perhaps another. Yet another form of tuning involves a theory that needs to assume highly specific, non-generic initial conditions (rather than arbitrary assumptions about parameters in the theory's generalizations) for which there is no independent evidence or further explanation in order to reproduce observed results. An illustration is provided by early models of the big bang prior to the introduction of inflationary models. The pre-inflationary models required highly specific assumptions about initial conditions to reproduce various features of the observed universe, with no independent evidence or theoretical rationale for these assumptions. Inflationary models show how some of the observed features follow from any one of a wide range of assumptions about initial conditions. It is often supposed that inflationary models provide better explanations of the observed features in question because of they do not require highly tuned initial conditions and to the extent that the required conditions are generic, the assumption that they hold may also seem less inductively risky than the assumption of more specific initial conditions.

Exactly why (if at all) it is justifiable to regard theories/ hypotheses/ assumptions about initial conditions that are finely tuned as less than fully explanatorily satisfactory is far from obvious and certainly not adequately addressed by the best known philosophical theories of explanation[30]. One natural thought is that highly tuned theories seem to generate a need for further explanations and hence are not regarded as natural explanatory stopping points in comparison with less tuned theories. A theory with lots of free parameters that need to take

[29] Zooming out from details, one might associate the Causal Markov (CM) condition used in causal modeling with the w-condition framework as described above. When CM is violated there will be (unconditional or conditional) statistical dependencies that are not explained-- w questions that are not answered, failures to discover what some pattern in the data depends on. CM encodes the idea that there will always be *some* explanation of those dependencies -- they can't be just brute. By contrast, independencies *are* accepted as brute-- they are natural stopping places in explanation and (special circumstances aside which actually produce tuning, as in certain biological examples) and don't seem to require further explanation. The faithfulness condition is, as argued above, an anti-tuning condition-- it excludes certain overly tuned kinds of explanations of independencies.

[30] Glymour (1980) is an important exception.

highly specific values naturally raises the question of why those parameters take those exact values rather than others -- that is, it generates a further demand for explanation of these values. A theory that postulates specific connections between parameter values raises the explanation-seeking question of why those specific connections hold. A theory that requires highly specific initial conditions raises the question of why just those initial conditions rather than others hold-- a question that may seem less pressing when a result is shown to follow from any one of a large range of initial conditions.

A related thought is that one role for explanation is to reduce arbitrariness or contingency or to show that an outcome is "necessary", given certain assumptions. Hempel's DN model attempted to capture some aspects of this idea via the claim that contingency is reduced when an explanandum is derived from a law. But quite apart from other problems with the DN model, this leaves out a lot since a derivation via a law from a highly tuned intial condition may meet DN requirements but do little to reduce arbitrariness. If, on the other hand, one can show that an explanandum would follow from any one of a very large range of possible initial conditions, this may make the explanandum look less contingent. Similarly for demonstrations that an outcome follows automatically from very generic assumptions, without the need for special assumptions about parameter values-- given these assumptions, one sees that the outcome could not have been otherwise. A striking example, discussed by Glymour (1980) , is provided by the contrast between Ptolemy's and Copernicus' explanations of a numerical regularity[31] relating features of the observed behavior of the superior planets. Ptolemy's theory can reproduce this regularity by making special assumptions about parameter values and epicycles. By contrast the regularity follows automatically within the Copernican system given a few generic assumptions that are central to that theory-- such as heliocentrism and the claim that the superior planets have longer periods than the earth. Copernicus' theory also explains apparent retrograde motion via similar assumptions while Ptolemy's requires special tuning.

I suggested above that when a theory seems explanatorily deficient (or at least incomplete) because it relies too heavily on fine tuning, it seems, at least some of the time, not so well inductively supported. Sometimes this may be because the presence of tuning suggests that there is likely to be an alternative less tuned theory that is better supported[32]. This is illustrated by the Ptolemy/Copernicus example. In some specific cases there are formal considerations that can support this assessment. For example, as discussed by a number of writers (MacKay, 2003, Henderson, 2014), within a hierarchical Bayesian framework, through the operation of the

---

[31] The observed regularity is this: if a superior planet goes through a number of cycles of anomaly while going through a number of revolutions in longitude in a number of solar years, then the number of solar years is equal to the number of oppositions plus the number of revolutions of longitude.

[32] Wolf and Duerr, forthcoming, quote Smeenk, 2018, p.218 as suggesting that in physics it is a common strategy to use fine-tuning as a guide to developing new (less tuned) theories. This is relevant to the role of elaboration in explanation discussed in Section 12.

Bayesian Ockham's razor, theories with lots of free parameters can be less well-supported than theories with fewer such parameters, given a common body of evidence. The HBM framework thus generates a penalty in terms of support for theories with more free parameters[33]. A similar rationale underlies the use of the Aikake Information Criterion. (In both these cases, as in the Ptolemy/ Copernicus example, it is assumed there is a non-tuned hypothesis which is an alternative to the hypothesis that overfits the data.) In a causal modeling context if one assumes a uniform distribution over the possible parameter values characterizing the direct causal links between two variables, then cases in which there is a faithfulness violation will have Lebesque measure zero, thus establishing a connection between this kind of tuning and what might be interpreted as a kind of "unlikeliness" . If one employs a causal discovery program that assumes faithfulness, one is in effect deciding not to infer to models which are tuned in the sense of having a faithfulness violation. Thus, in these cases, to the extent that a good explanation is one that avoids fine tuning, we have formal stories connecting this kind of explanatory goodness and inductive support.

In other cases such as inflationary versus non-inflationary scenarios, the warrant for an interpretation in terms of inductive support is admittedly less clear because it is hard to see how make any connection with probabilities -- that is, no obvious basis for regarding a highly special assignment of initial conditions as "improbable". Still is perhaps something defensible in the thought that a highly generic assumption about initial conditions that is a superset of highly specific assumptions is less inductively less risky than the specific assumption .

It is an understatement to say that the connection between tuning in its various forms, explanation and inductive warrant, is currently not well-understood. Apart from the Appendix immediately below, I will not try to say more about it beyond what I have said above except to underscore its interest and importance[34].

Appendix to Section 11.

---

[33] Unfortunately this argument (as well as arguments that support AIC) apply only in special contexts: one needs the assumption that the data are drawn from a single stable probability distribution, with a noise term acting as a source of possible overfitting. The evidence base for many theories with free parameters (e.g the standard model in high energy physics) does not satisfy these conditions. For this reason my view is that the argument for penalizing free parameters coming from the Bayesian Ockham's razor or AIC does not generalize to other contexts in which these background assumptions are not satisfied. However, I will not try to argue for this assessment here.

[34] It is also worth emphasizing that highly tuned theories are often (plausibly) regarded as non-simple. So this is one route by which one kind of simplicity consideration can be relevant both to potential explanatoriness and inductive support. However, it should be clear from some of the examples considered previously that there are other varieties of simplicity or simplicity-based considerations that are not straightforwardly linked to potential explanatory goodness. Again, this suggests the need for a more discriminating assessment of the connections between simplicity, explanation and inductive support. My view, as intimated above, is that only some aspects of simplicity (rather than some generic notion) matter for explanatory assessment.

The w-condition account as described in Woodward, 2003 and Hitchcock and Woodward, 2003 does not say anything explicit about tuning. One might wonder how, if at all, that account might be extended, to capture tuning-based considerations. This issue has in effect been explored in Wolf and Duerr, as well as Ylikoski & Jaakko Kuorikoski, 2010 and I draw on them in what follows.

First, although the w-account valorizes correct answers to w-questions, it is arguable that a natural extension is to add that if an explanation, in answering a w-question, generates lots of other w-questions which it fails to answer (e.g. questions about why some highly specific initial condition or a particular parameter value is present when there is no particular reason to expect this) that is a less than ideal feature of the explanation, particularly when an alternative explanation that does not have this feature is available. As indicated above, this captures one set of considerations that leads us to disprefer highly tuned explanations.

Another consideration extends the notion of invariance in Woodward, 2003, Hitchcock and Woodward, 2003 to encompass what Wolf and Duerr call modal robustness. As suggested earlier we prefer explanations that appeal to relatively invariant generalizations because they can be used to answer a wider range of w-questions. Woodward, 2003 talked mainly about invariance under interventions and changes in background conditions. But one can also consider the extent to which an explans- explanandum relation remains stable (in the sense of correctly telling us how target explananda depend upon the factors cited in the explanans) under changes in parameters figuring in the explanans. In general if an explanation only works (in the sense of correctly capturing dependency relations) if we assume that the parameters in the generalizations figuring in the explanation take very specific values and would not work otherwise and there is alternative potential explanation that does not have this feature, this may (sometimes-- see below) provide grounds foe preferring the latter. As Wolf and Duerr emphasize this is also an anti-tuning consideration. Arguably this consideration is particularly compelling when there is uncertainty about whether the values of the parameters in the tuned explanation are empirically correct or when we don't have any independent evidence for those values or when there is reason to think that these may change under changing background conditions or for some other reason[35]. Obviously when an explanation assumes parameters that only hold under very specific background conditions, the explanation cannot be used to tell us what would happen if those conditions were to change. An explanans that is not sensitive in this way can be used to answer a wider range of w-questions.

[35] For example, a demonstration that the entropy of a system will increase given certain highly specific (correct) assumptions about its initial conditions seems unsatisfying in comparison with a demonstration that such an increase will occur for almost all initial conditions. The latter tells us what would have happened if those initial conditions had been different and also makes it clear the behavior in question does not depend on those specific initial conditions. A demonstration that the tosses of a suitably symmetrical coin will result in fair outcomes that depends on the exact dynamics governing the coin and exact the initial conditions on successive tosses is a less satisfying explanation than a demonstration that this outcome follows for almost all initial conditions that the tosser can impose and for almost any plausible dynamics.

That said, it also seems clear that the considerations just described need to be circumscribed in some way-- they don't always seem compelling. An explanation of the tides does not seem problematic because it requires particular parameter values for the masses of the earth and moon and the distance between them or because the gravitational inverse square law assumes a particular value for the gravitational constant G. I have no account to offer about this.

12. The Role of Elaboration

I've previously argued that explanation as a goal is different from the goal of finding hypotheses that are simple or unified or have certain other explanatory virtues. In this section I draw attention to another distinctive feature of explanation as a goal-- the way in which a concern with explanation leads to *explanatory elaboration*. What I mean by this is that when a potential explanation is proposed, it will often suggest or will be accompanied by background assumptions that suggest ways in which the potential explanation can be made more detailed or elaborated. In fact in many cases, the suggested elaboration will be hard to avoid if the potential explanation is to be taken seriously at all-- see the smoking example below. This elaboration, in turn, will typically have additional testable consequences and when these are observed or not this provides additional grounds to distinguish between the elaborated explanation and competing alternatives. By contrast, the aim of finding simple or unified hypotheses typically (often?) does not suggest such elaboration or at least does not suggest fruitful elaborations.

Suppose a correlation e between cigarette smoking and lung cancer is observed. Consider the following potential explanation of e: (*h1*) cigarette smoking causes lung cancer. There are of course competing potential explanations of E-- for example, R.A. Fisher's suggestion that (*h2*) E is due to a genetic factor G that acts as a common cause of both smoking and lung cancer. If (*h1*) is correct, it suggests a rather specific elaboration: If smoking causes lung cancer, it is very likely that this is because material from cigarette smoke is taken into the lungs where it has a carcinogenic effect. This in turn suggests (even if it does not strictly require) that we should expect various other consequences. For example, it is plausible to expect a dose/response relation -- that the incidence of lung cancer should be higher among those who smoke more or for longer periods of time. One might also expect that the use of filtered cigarettes will reduce the probability of lung cancer. Suppose these additional effects are observed. There is no reason to expect them if (*h2*) is correct and it doesn't look as though there is any natural elaboration of (*h2*) that might explain them. To explain these effects in terms of (*h2*) one would have to assume, not just that G causes both smoking and lung cancer but that there are variants of G that operate in such a way that they cause some people to smoke more and at the same time to have a higher probability of lung cancer, other variants that cause some to smoke less and have a lower probability of lung cancer, variants that both cause some people to prefer filtered cigarettes and to have a lower probability of lung cancer and so on. There is nothing about (*h2*) that suggests or motivates these elaborations. If the above consequences are observed, this supports (*h1*) in preference to (*h2*).

It may be tempting to think that in this scenario *h2* is dispreferred because it is a less good potential explanation, thus vindicating IBE after all. This is mistaken. In the above scenario (*h1*) is preferred over (*h2*) because (*h1*), when elaborated, has additional consequences that are observed. (*h2*) fails to imply those consequences and in fact suggests that they will be absent. It

is the existence of this evidence that tips the balance toward (*h1*). If this evidence did not obtain this would have undermined (*h1*).

Several other points about this example are worth noting. First, the elaboration strategy works by finding additional dependency relations that are suggested or implied by the original hypothesis and background knowledge. This makes use of domain specific information about how the factors cited in the explanation are likely to operate. Considerations having to do with simplicity or unification do not play any obvious role in this process. These seem too domain-independent to play the role of guiding elaboration. In addition, as far as simplicity goes, note that elaboration involves complicating or adding to the original hypothesis, and in that sense making it less simple (and less likely, within a Bayesian framework). The advantage of the elaborated hypothesis is that *if* it turns out to have evidential support, it explains more (and we value explanations and explanations that explain more). Moreover, elaboration tells us where to look for additional evidential support if it exists. If such support exists, this can undermine rival potential explanations, thus justifying acceptance of the remaining hypothesis. That we end up accepting a stronger or more committal hypothesis -- elaborated h-- rather than just h is due to the fact that we care about explanation and we view the choice we face as a choice among competing potential explanations, rather than just a matter of choosing the likeliest hypothesis, which is of course just our original evidence e.

13. IBE Elsewhere in Philosophy

It is no secret that many of the advocates of IBE have backgrounds in epistemology and metaphysics or at least approach philosophy of science with sympathy for metaphysics. By contrast, many philosophers of science with a less metaphysical orientation (Achinstein, Cartwright, Hacking, Morrison, Sober, van Fraassen, among others) reject IBE understood as a normative thesis as well as the claim that IBE is widely employed in science. Relatedly, it is common for metaphysicians to appeal to IBE to support metaphysical conclusions outside of science. Often they argue as follows: Science employs IBE to reach well supported conclusions. This provides reason to think that the use of IBE in metaphysics will also lead to conclusions that are at least somewhat well supported. (See Sider et al. 2007 and Paul, 2012, who writes that "my central claim [is} that most metaphysical claims about the world rely on inference to the best explanation" for explicit endorsement of this line of argument). Thus we find Armstrong XX arguing for the existence of laws of nature and for the further claim that these should be understood as relations of necessitation between universals on the grounds that these claims provide the best explanation for observed patterns of regularity. Other philosophers invoke IBE to argue for realism about mathematical objects, the reliability of testimony, and, ranging further afield, the reality of Jesus' resurrection.

None of this should be surprising. Competing metaphysical theses rarely have different empirically testable consequences. IBE is designed for situations in which we have competing hypotheses that agree on the evidence we have so far, with choice among these hypotheses being made on considerations having to do with potential explanatory virtues, many of which appear to be non-empirical. So IBE seems tailor- made for application to competing metaphysical theses, where choice also has to made on non-empirical grounds and to gain credibility in such applications from the supposed fact that it is also widely employed in science.

Obviously if the argument of this essay is correct, it undermines the use of IBE in metaphysics. If forms of IBE based on such supposed explanatory virtues as simplicity and unification are not normatively defensible in science, they are likely not in metaphysics either. (The one explanatory virtue that we found to be connected to inductive support -- tuning-- does not seem to have any obvious application in metaphysical argument, or at least the arguments made by metaphysicians do not seem to appeal to this.) Indeed, it is tempting to, so to speak, turn the argument around: If a proposed form of inference licenses conclusions about which of various metaphysical theses are true, perhaps we ought to be skeptical the this form of inference is widely used in science, which seems to be an enterprise in which empirical evidence plays a different role than in metaphysics. And those of us who are skeptical of metaphysics may regard the use of IBE in that area of philosophy as an additional reason for skepticism about that inference form.

References

Cabrera, F. (2021) "String Theory, Non-Empircal Assessment and the Context of Pursuit" *Synthese* 198: 3671-3699.

Cabrera, F. (2020) "Does IBE Require a 'Model' of Explanation?" *British Journal for the Philosophy of Science* 71: 727–750.

Cabrera, F. (2017) " Can There Be a Bayesian Explanationism? On the Prospects of a Productive Partnership " *Synthese*

Craver, C and Kaplan, D. (2020) "Are More Details Better? On the Norms of Completeness for Mechanistic Explanations" *British Journal for the Philosophy of Science* 71: 287-319.

Douven, I (2017) "Inference to the Best Explanation: What is it and Why shold we Care?" In McCain, K. and Poston, T. (eds.) *Best Explanations: New Essays on Inference to the Best Explanation.* Oxford: Oxford University Press.

Douven (2024) "Abduction" Stanford Encyclopedia of Philosophy. Accessed 1/9/2024.

Earman, J. (1992) "Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory" Cambridge: MIT Press.

Elliott, K. (2021) "Inference to the best explanation and the new size elitism" *Philosophical Perspectives* 35: 170-188.

Glymour, C. (1980) "Explanations, Tests, Unity and Necessity" *Nous* 14: 31-50.

Hempel, Carl G., 1965a, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press Friedman, M. (1974) "Explanation and Scientific Understanding", *The Journal of Philosophy*, 71(1): 5–19.

Harman, G. (1965) "The Inference to the Best Explanation" *The Philosophical Review* 74: 88-95.

Harper, W. (2011) *Issac Newton's Scientific Method: Turning Data Into Evidence About Gravity and Cosmology.* Oxford: Oxford University Press.

Hempel, Carl G., 1965a, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press.

Henderson, L. (2014) "Bayesianism and Inference to the Best Explanation" *British Journal for the Philosophy of Science* 65: 687-715.

Hitchcock, C. and Woodward, J. (2003) "Explanatory Generalizations, Part II: Plumbing Explanatory Depth" *Nous*. 37: 181-99.

Kitcher, P. (1989, "Explanatory Unification and the Causal Structure of the World", in Kitcher and Salmon 1989: 410–505. *Scientific Explanation* (Minnesota Studies in the Philosophy of Science, Volume 13), Minneapolis, MN: University of Minnesota Press.

Lipton, P. (1991/204) *Inference to the Best Explanation* Routledge: London and New York

MacKay, D. (2003) *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.

McCain, K. (2019) "How Do Explanations Lead to Scientific Knowledge?" In *What is Scientific Knowledge?* ed. K. McCain and K. Kampourakis. Routledge.

McCain, K. and Poston, T. (2023) "Explanation and Evidence" In Maria Lasonen-Aarnio and Clayton Littlejohn (eds.) *Routledge Handbook of Evidence*, Routledge.

Morrison, M. (2000) *Unifying Scientific Theories: Physical Concepts and Mathematical Structures.* Cambridge: Cambridge University Press.

Nyrup, R. 2015: "How Explanatory Reasoning Justifies Pursuit: A Peircean View of IBE", *Philosophy of Science* 82(5): 749-760.

Paul, L. (2012) "Metaphysics as Modeling: the Handmaiden's Tale" *Philosophical Studies* 60 (1):1-29

Prasetya, Y. (2021) "Which Models of Scientific Explanation are (In)Compatible with IBE?" *British Journal for the Philosophy of Science*

Rice, C. (2021) *Leveraging Distortions: Explanation, Idealization and Universality in Science*. Cambridge: MIT Press.

Roche, W., and Sober, E, (2013), "Explanatoriness is Evidentially Irrelevant, or Inference to the Best Explanation Meets Bayesian Confirmation Theory"; *Analysis*, 73. 659-668.

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

Salmon, W. (2001) " Explanation and Confirmation: A Bayesian Critique of *Inference to the Best Explanation"* in G. Hon and S. Rakover (eds.) *Explanation: Theoretical Approaches and Applications* Dordrecht: Kluwer 61-91.

Salmon, W. (1971a) "Statistical Explanation", in Salmon 1971b: 29–87.

Salmon, W. (1971b) (ed.) *Statistical Explanation and Statistical Relevance*, Pittsburgh, PA: University of Pittsburgh Press.

Sanford, P. K. (2006) *Exceeding Our Grasp: Science, History and the Problem of Unconceived alternatives* New York: Oxford University Press.

Smeenk, C. (2018) (2018). "Inflation and the Origins of Structure". In: Beyond Einstein: Perspectives on Geometry, Gravitation, and Cosmology in the Twentieth Century. Ed. by David E. Rowe, Tilman Sauer, and Scott A. Walter. Springer New York, pp. 205–241.

Smith. G. (2014) "Closing the Loop: Testing Newtonian Gravity, Then and Now," in *Newton and Empiricism*, ed. Zvi Beiner and Eric Schliesser, Oxford University Press 262-35.

Schulte, O. (2022) "Formal Learning Theory" Stanford Enccylopedia of Philosophy

Sider, T. Hawthorne, J. and Zimmermann, D. (2007). *Contemporary Debates in Metaphysics*. Blackwell.

Sober, E. (1994) "Let's Razor Ockham's Razor", in *From A Biological Point of View*, Cambridge: Cambridge University Press, 136–57.

Sober, E. (2015) *Ockham's Razors: A User's Manual*, Cambridge: Cambridge University Press.

Spirtes, P., Glymour, C and Scheines, R. (2000) *Causation, Prediction and Search.* Cambridge: MIT Press.

van Fraassen, B. (1980) *The Scientific Image.* Oxford: Clarendon Press.

Ward, B. (2022) "Informational Virtues, Causal Inference, and Inference to the Best Explanation" PSA 2022: The 28th Biennial Meeting of the Philosophy of Science Association (Pittsburgh, PA, November 10-13 2022)

Weinberg, S. (1999) "What is quantum field theory, and what did we think it was? . In *Conceptual Foundations of Quantum Field Theory*, (ed. Cao) Cambridge: Cambridge University Press.

Will, C . (1981, 1993) *Theory and Experiment in Gravitational Physics*, Cambridge University Press.

Wolf, W. and Duerr, P. (Forthcoming) "The Virtues of Pursuit-Worthy Speculation: The Promises of Cosmic Inflation" *British Journal for the Philospohy of Science.*

Woodward J. (2023) "Sketch of Some Themes for a Pragmatic Philosophy of Science" in *The Pragmatists Challenge*.) (ed. Andersen, Mitchell) Oxford: Oxford University Press. 15-66.

Woodward, J. and Hitchcock, C. (2003) "Explanatory Generalizations, Part I: A Counterfactual Account" *Nous* 37: 1-24.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Ylikoski, P. and Kuorikoski, J. (2010) "Dissecting Explanatory Power" *Philosophical Studies* 148: 201-219.

.