**Note to Reader: This paper, recently published in *The European Journal for the Philosophy of Science* 15**, **21 (2025)*,* is a more recent version of a paper with the same title previously deposited on the archive. This version differs considerably from the earlier version-- material has been dropped and an additional section has been added. Please quote from the published version.**

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

**James Woodward**

**Abstract**

This paper investigates the status of inference to the best explanation (IBE), in contrast to inference to the only explanation (IOE) against the background of Woodward's what-if-things- had-been-different (w) account of explanation. It argues that IBE is not a defensible form of inference. By contrast IOE is defensible and objections to its use (e.g., on the basis of claims about underdetermination) are exaggerated. Although some accounts of explanation in conjunction support IBE, the w-account does not. It is also argued that we should think of explanation as an independent goal of scientific investigation that is valuable in its own right and not because it is a means to discovering truths via IBE. The correct picture of the connection between explanation and truth is simply that successful explanation requires a true or effectively correct explanans. However, we cannot establish that an explanans has this feature by appealing to its potential explanatory power-- that it would if true explain well. Instead, evidence that is independent of potential explanatory power is required. This is what IOE provides.

Key words: Inference to the best explanation; inference to the only explanation; what-if-things-had been-different account of explanation; explanatory virtues; explanatory unification, simplicity

**1. Introduction**

This is a paper about inference to best explanation (IBE), inference to the only explanation (IOE) and, more generally, about the place of explanation in scientific thinking. It attempts to provide a "big picture story" about these and related matters, showing how a variety of claims about explanation and its role in inference fit together in a coherent way. Here is a brief overview: By IBE I understand the claim that the fact that a hypothesis *h* would if true provide a better explanation (call this *h*'s potential explanatoriness) than alternative hypotheses would provide if true for some set of phenomena provides strong support for acceptance of *h*, independently of other evidence for *h,* as long as *h* is a good enough explanation. I argue that this is not a legitimate inference principle. Although many other writers have rejected IBE (e.g., van Fraassen, 1980, Achinstein,1993, , 2021, 2024, Roche and Sober, 2013, Salmon, 2001), my discussion goes beyond these in a number of ways. It investigates the implications of a specific account of explanation-- Woodward's-what-if things-had-been-different account (the w-account described in section 4)-- for IBE, an issue that has not been hitherto explored. It is argued that, unlike a number of other accounts of explanation, the w-account does not support the legitimacy of IBE. Typically, defenses of IBE invoke various explanatory virtues such as unification, simplicity, and the provision of mechanistic information (e.g. Lipton,1991/2004, Carbrera, 2020). I argue that these either are not explanatory virtues at all or that their use in IBE leads to inferences that are (for good reason) regarded as unwarranted in scientific practice. Next, assuming that considerations of potential explanatoriness do not function to provide grounds for acceptance via an IBE, the question naturally arises of what it any alternative role such considerations play. My answer is that these considerations function to guide pursuit, rather than serving as grounds for acceptance. In particular, to the extent that inquiry aims at explanation we may think of it as proceeding via two step-process: First (i) researchers generate alternative potential explanations for a set of target phenomena and then (ii), in a separate step, they use empirical evidence to discriminate among these. Ideally, this involves establishing that one potential explanations as the true of correct explanation by ruling out all of the alternative potential explanations on some combination of empirical and theoretical grounds. Assuming that there is some explanation for the phenomena in question, this justifies an inference to the only remaining alternative. This process is what is sometimes called inference to the only explanation (IOE) (see Bird, 2022, Woodward, 1990).

 As a number of writers observe, one of the primary motivations for belief in IBE is the claim that often, perhaps always, the choice between alternative hypotheses is underdetermined by ordinary empirical evidence, so that appeal to transempirical virtues like potential explanatoriness is required to resolve this impasse. I argue against this claim, providing reasons for thinking that ruling out alternative possible explanations by ordinary empirical evidence is in many cases an achievable ideal and that appeal to non- empirical grounds for hypothesis choice is not required. (This part of my argument also goes beyond many defenses of IOE, since I emphasize such considerations as the fact that our choices are typically among "effective" theories and hypotheses.) I also argue that if explanatory considerations function to guide pursuit rather than serving as grounds for acceptance, this provides a sense in which truth per se is not the only aim of science. Scientific theorizing does (often) aim at discovering truths but it is also important that it does not aim at the discovery of just any old truths, but rather truths that are valuable or relevant in various ways, including truths that can figure in explanations. In this sense, the discovery of explanations can be regarded as a distinctive, independent aim of science, a theme which I develop in Sections 2-3. This aim is not replaceable by the goal of finding hypotheses that have such features as simplicity, unifying power, or other supposed theoretical virtues.

` This is, as I have said, an attempt at a big picture. I do not claim that the various elements I have described stand in relations of logical entailment or that some are strictly obligatory, given others. Rather I claim that the picture I present makes sense in that the individual elements in the picture are plausible in their own right and that they hang together in a coherent way. Thus, my picture is presented in a spirit of "think about matters this way -- you'll see that they form an attractive picture. "

 **2**. **Background and Motivation: Truth and Explanation as Aims of Science.**

Philosophical discussion of explanation has long recognized the notion of a potential explanation-- a set of claims[[1]](#footnote-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 (that is, their potential explantoriness) without knowing whether these claims *are* in fact true. I will adopt this assumption, which is shared by advocates of IBE. I will also proceed on the assumption that any discussion of IBE needs to be situated within some definite model of explanation. As we will see, different models of explanation differ in the support they provide for IBE. The model of explanation on which I will rely is the what-if- things- had- been -different account (hereafter w-account) described in Woodward, 2003, Hitchcock and Woodward, 2003.This model is discussed in Section 4 but for the most part I will rely on the aforementioned discussions for motivation. I will add, however, the requirements of the w-account are relatively undemanding and widely accepted -- they amount at bottom to the idea that explaining an outcome is a matter of finding factors on which that outcome depends.

 Next, my focus will be entirely on the use of IBE in science and to some small extent in common sense reasoning contexts. I will have nothing to say about the use of IBE in metaphysics or analytical epistemology.

 I will also assume that science sometimes aims at discovering or establishing truths[[2]](#footnote-2) or at least claims that are good approximations or "effectively" valid within some domain of interest. (See below-- Section 4.) Hereafter I will often use "truths" to encompass all these possibilities. Moreover, I will assume that successful explanation requires such truth or effective validity of what does the explaining. Again, these assumptions are common ground with defenders of IBE since IBE is an inference method that is supposed to connect potential explanatoriness with evidence for truth. If successful explanation does not require the truth (or something similar) of what does the explaining, it is hard to understand how IBE is supposed to function. Moreover, as we shall see, for IBE to be even prima-facie plausible, it requires something like a notion of approximate truth or effective validity of explanantia. Similar remarks apply to IOE.

 However, even when science 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 seem uncontroversial but notice that there is nothing in this picture that requires us to suppose that discovering 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 acceptance. Instead, these observations are consistent with the view that the discovery of correct explanations is valuable in its own right. On this alternative view, we want our explanations to satisfy a truth requirement, but it gets things the wrong way around to suppose that the role of explanatory considerations is to guide us to truth. Instead, a focus on explanation leads us to attempt to discover particular kinds of truths-- those that figure in explanations.

 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 it makes reasonably accurate predictions much of the time -- 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. 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.

 But although candidates for predictive hypotheses with the above characteristics would, if accurate, be more valuable or worth knowing than alternatives, 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 a reason to think that *h* is in fact accurate. (Call this Inference to the Best Predictor, in analogy with IBE). That a hypothesis would, if predictively accurate, be highly valuable to discover is no reason to conclude that the 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 analysis of when and why the predictive hypothesis works.)

 Prediction is not explanation, but I suggest that a similar point holds for IBE. Just as we can't infer that a hypothesis is an accurate predictor merely on the basis that if accurate, it would score high on a list of virtues that we think that a good predictors should have, so also we can't infer from potential explanatoriness (or virtues that a potential explanation would have if true) to the conclusion that such a potential explanation is true or has strong inductive support. As the analogy shows, hypotheses can have potential “informational" value[[3]](#footnote-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 or grounds for acceptance. Successful explanation is one such informational goal. I stress here that what I am claiming is not just that the discovery of informative hypotheses is a goal in science. I am claiming that a particular kind of information-- information about explanatory relations, where this understood along the specific lines of the what-if-things-had-been-different account of explanation described in Section 4, is such a goal.

 What do I mean by saying that the discovery of explanation is a goal in its own right? Philosophers often draw a contrast between "epistemic” or "veritistic" goals which are "internal" to science and other sorts of goals(or virtues), which are regarded as in some way "external" or perhaps "merely pragmatic". For example, my research may have the "virtue" of helping me to get rich but this is plausibly understood as an "external" non- epistemic virtue. I contend that successful explanation is an "internal" epistemic virtue that a scientific theory can possess-- that is part of what I mean by claiming that explanation is a goal in its own right. I thus reject the common idea that the epistemic virtues are confined to those whose role is to serve as a means or sign of truth and the accompanying claim that every other virtue or goal of scientific theorizing is "external" in the way that getting rich is or else merely a means to truth.

 One reason for rejecting the claim that truth is the single overarching epistemic goal of science to which everything else "epistemic" is a means is that this runs immediately into the 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 ability to figure in explanations) 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 valuable in their own right and internal to science. However, one might still wonder why the focus should be on explanation rather than other goals that might serve to pick out valuable kinds of truths. I will say more about this below, but as an initial motivation, consider the common 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, with “simple" being understood in some way that is not coextensive with "explanatory". In addition to other problems such as the difficulty of providing a clear characterization of what is meant by this feature, "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 possible candidate for such a justification but one that has not found a generally accepted defense, particularly if one is operating with the generic notion of simplicity commonly assumed in philosophical discussion[[4]](#footnote-4) . Alternatively, one might hold that simpler hypotheses are (usually?) easier to reason or calculate with or to test and are valuable for that reason[[5]](#footnote-5). This is also an instrumental justification but one that raises the worry that it does not give simplicity judgments the "objective" status one would like them to have if they are to play a central organizing role in science.

 I suggest, by contrast, that 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 -- the w-account described below-- we can avoid the 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, it does not seem that we are naturally led to ask for some further justification of why the discovery of such explanations is valuable.

 If we think of explanation as an independent aim but not a means for discovering truths via an IBE, the most natural alternative is that potential explanatoriness belongs to what is sometimes called the context of pursuit[[6]](#footnote-6) rather than the context of acceptance-- a claim that has been recently advocated by other writers (e.g., Nyrup, 2015, Carbrera, 2021, Wolf 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,* so that this is the sense in which discovering explanations is an aim that guides inquiry. However, on this picture, the discovery that *h* if true would explain *e* is not by itself a reason for accepting *h*. Instead, accepting *h* requires what Achinstein, 1991 calls "independent warrant"-- that is, evidence that is independent of *h'*s potential explanatoriness and that differentially supports *h* against alternative potential explanations. This is the idea that establishing an explanation is a matter of inference to the only explanation rather than IBE. Again, I take this to be a very natural way of realizing the independent aim conception of explanation.

**3. Inference to the Only Explanation and Explanation as an Independent Goal**

More specifically, by Inference to the Only Explanation (IOE) I mean an inference having the structure laid out in Bird (2022) -- see also Bird, 2018, 2010, 2007. Following Bird, 2022, with slight modifications, I take IOE to use of the following premises:

1) Target explanandum *E* has some explanation

2) *h1..hn* are the only potential explanations of *E* (where as explained below this means the only hypotheses that if effectively valid would explain).

3) *h1... h n-1* are ruled out by some combination of empirical evidence and more theoretical assumptions, leaving only *hn*.

4) *hn* explains *E* and should be accepted.

IOE thus embodies two key ideas: first that we are choosing among competing potentially explanatory hypotheses under the assumption that one and only one of these (among the competitors) is a correct or effectively valid explanation and, second, that this involves a form of eliminative inference according to which all of these alternatives except one is ruled out or eliminated by empirical evidence, often in combination with other assumptions of a more theoretical nature. (See below for examples.) The idea that providing evidential support in science often involves systematic search through alternative hypotheses and the ruling out of alternatives is a common idea-- a very partial list of defenders includes, in addition to Bird, 2022, includes Platt, 1964, Woodward, 1990, Earman, 1993, Achinstein, 1991, Kitcher, 1993 among many others. A number of writers also explicitly contrast this procedure with IBE, which they reject: Woodward, 1990[[7]](#footnote-7), Achinstein, 1991, Norton, 2021/2024. I think of what follows as drawing on (and as supported by these previous discussions, although I do have some moderately new things to say about the details of how the elimination procedure works (Section 6).

 The idea that connection between explanation and evidential support is a matter of inference to the only explanation has a number of attractive features. First, it accounts for many of the features of scientific practice to which defenders of IBE appeal. Like IBE-based views, IOE recognizes that the discovery of explanations is central to much of science, that explanations need to appeal to assumptions that are truth-like and that the assessment of competing potential explanations is an important scientific activity. (However, as argued below, IOE and the independent goal picture need not assume that it is possible to make the kinds of comparisons of explanatory goodness to which IBE appeals.) Because of these similarities, when presented with cases of apparently legitimate inferences which IBE-based views interpret as comparisons of explanatory goodness, it can be plausibly argued that these are instead cases of IOE. 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 is 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. In fact in the same passage, Darwin goes on to argue that the main alternative hypothesis he considers -- 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 than some alternative.

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 with IOE-- that the rays are negatively charged particles is the only explanation consistent with his evidence[[8]](#footnote-8).

 Of course, I agree that citing particular examples does not (and cannot) show that there are not other examples that conform to IBE but are not cases of IOE. My point is rather that, as the above 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 cases of IOE. In fact, as Norton, 2021, 2024 shows, many other cases in the literature that that have been cited as instances of IBE are very plausibly construed as cases of IOE.

 A second appealing feature of the IOE 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 or best supported? A similar problem arises for more specific candidates for explanatory virtues such as unification and simplicity. I think it is fair to say that there is no generally accepted answer to this question. By contrast, IOE 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.

 A third appealing feature of the idea that explanation is a goal of science, realized by the IOE picture is this: scientific theories contain many claims that go beyond what is required to describe "observables" or to achieve "empirical adequacy". This is often thought to involve additional inductive risk and raises the question of why theorizing willingly incurs such risks. The idea that explanation is an independent goal of science provides a straightforward answer: we value explanation, and this often requires reference to unobservables because we have very strong evidence that explananda of interest are explained by these rather by observable factors. For example, an extended series of experiments ruled out alternative explanations of Brownian motion that appealed to observable factors such as convection currents and other environmental influences; thus, in accord with IOE, leaving the conclusion that this explanation requires reference to unobservable molecular collisions as the only remaining possibility. We are willing to incur whatever inductive risk is involved in these additional commitments because we care about explanation in addition to inductive support and because we are often in a position in which we can be confident that we have eliminated alternatives to the explanatory claim in question, thus achieving an acceptable level of inductive security.

 A related feature of scientific theorizing that follows from this picture is this: scientific theories seem to contain *modal* or counterfactual claims that go beyond claims about what actually happens-- theories make claims about what would happen if various conditions were to be realized even if those conditions are not in fact realized or if we are unable to determine whether they are ever realized. This makes sense if explanation is a scientific goal in the manner described because (at least on the w-condition account of explanation defended here) successful explanation requires such modal or counterfactual information

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

I turn now to a more detailed look at the w- account, as well as some of the implications of that account for IBE. 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. (Think of these dependence relations as what are described by laws and causal generalizations or by implications of these conjoined with information about initial and boundary conditions.) The conditionals associated with answers to w-questions are understood as "interventionist" or some other form of non-backtracking counterfactuals. In keeping with 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[[9]](#footnote-9). 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 *G2.*

 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-questions *w1* and a second explanation *EX2* may answer all 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 will 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 w-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 this recall pattern. 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 (e.g, about the state of the patient's immune system) about what *D* depends on as well. 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 *EX1* and in this respect might be judged more valuable qua explanation. Nonetheless it is also true, according to the w-account, that the *EX1s* are also explanatory (in virtue of answering w-questions) even if they explain less than the corresponding *EX2s*. Thus, on the w-account, a 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 or has greater scope-- the existence of the successor theory does not render the earlier theory unexplanatory or incapable of inductive support. This is a consequence 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 effective theories, a requirement on explanation of exact truth and applicability across all domains and scales would render IBE inapplicable to almost all cases.

 In this connection, it is worth remarking that there is related internal problem with IBE that needs to be addressed. On the one hand, this doctrine is framed in terms of inferring to the best explanation among a set of alternatives --it is this best explanation that is most strongly supported or accepted. On the other hand, the advocate presumably wants to avoid construing IBE in such a way that only the *EX2s* and not the *EX1s* above can be inferred on the basis of IBE since it is the *EX2s* that provide best explanations. In addition to being at odds with scientific practice this would greatly restrict the application of IBE-- roughly, when a "fundamental" theory exists, IBE can be used only to infer to this since this provides the best explanation. One strategy for avoiding this would be to restrict the application of IBE to "competing" alternative explanations. Intuitively, *N* and *GR* don't compete in the way in which, say, *N* and *MOND* (Modified Newtonian Dynamics) do[[10]](#footnote-10). *N* and *GR* don't compete qua explanatory hypotheses in the sense that within the domain of application of *N* their predictions are close to one another. Again, a restriction of this sort seems required if IBE is to produce even prima-facie sensible results. In what follows I will assume it applies to both IBE and IOE.

 The w-account has two additional implications that will be discussed in greater detail below. First, in accord with what I suggested earlier, there does not seem to anything in the w-account which provides a direct connection between explanation and reasons for belief or inductive support. That a candidate explanation *EX3* would if true answer a greater range of w-questions than some alternative candidate explanation *EX4* seems to provide no reason at all to think that first explanation is more strongly inductively supported or worthy of belief. Suppose our evidence *E* is the terrestrial and astronomical phenomena *P* known to early 19th century scientists and our candidate explanatory theory is Newton's (*N*) which purports to account for *P*. Stipulate that *GR* if true would provide better explanations of *P* and other phenomena than *N*. This is not a reason for thinking that *GR* is true or that there are reasons for belief in *GR* that are stronger than those supporting *N*.

 Of course, some may take the features just described as a reason to reject the w-account in favor of some alternative treatment of explanation. In response, I emphasize two points. First, the emphasis on dependency relations in the w-account is very intuitive and, arguably a feature of any account that applies to causal explanation. Second, even if one prefers an alternative account, it is worth exploring the implications of the w (and similar) accounts for IBE. In particular, I take the discussion that follows to illustrate the point that it matters for the assessment of IBE which account of explanation one adopts. One can't be neutral on this issue.

 **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 inductive support in virtue of its potential explanatoriness 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 "good enough", thus avoiding van Fraassen's (1980) "best of a bad lot" objection. I also assume, again in accordance with the usual understanding, that alternative explanations to which IBE is to be applied do not include explanations that have been falsified or strongly undermined by current evidence.

 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[[11]](#footnote-11).

 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/acceptance 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 and grounds for acceptance 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 followed Lipton, 1991/ 2004 in largely avoiding 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”, and 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.

 I don't think this claim about the irrelevance of the traditional models is correct. First, some of these models, fit naturally with or support IBE, given assumptions about confirmation or grounds for acceptance with which they are typically associated, while others provide no such connection. For example, the DN model of explanation, along with a hypothetico-deductive model of confirmation (assumed together by many writers such as Hempel), views explanation and confirmation as closely associated: if *E* is a potential DN explanation of *M*, then *M* will provide H-D confirmation for *E*. Thus, modulo these assumptions, potential explanatoriness tracks evidential support or grounds for acceptance. Similarly, if *E* is a potential IS explanation of *M* (*E* if true would confer high probability on *M*), then, on a likelihoodest conception of evidential support, which is commonly tied to the IS model, *M* is evidence for *E*. (Of course the HD model and arguably the DN model as well as the IS model are defective, but my point is simply to illustrate how certain views about explanation and support are mutually enforcing.) Defenders of unificationist models of explanation like Friedman 1974 and Kitcher 1989 also take the fact that *h* would if true provide a unified explanation of disparate explananda *M1* and *M2* to be, so to speak, dual to an account according to which, when such unification is present, *M1* and *M2* provide increased evidential support for *h*, via "consilience". (Friedman, 1983 is particularly clear about this claimed relationship in the context of spacetime physics.) By contrast other models of explanation suggest no such connection between potential explanatoriness and evidential support or grounds for acceptance. This true, for example, of Salmon's SR and causal mechanical models: in connection with the latter, the fact that a proposed explanation *E* would if true successfully trace the continuous causal processes that led to some explanandum *M* is no reason to think that *E* is true. (It is no coincidence that Salmon was critical of IBE-- see Salmon, 2001.) A similar conclusion (about the absence of a connection between potential explanatoriness and grounds for acceptance) holds, I claim, for the w-account of explanation. It thus matters what accounts of explanation and grounds for acceptance we adopt when we evaluate IBE.

 A related consideration is this: the traditional models, whatever their limitations, were at least relatively precise. The current talk of explanatory virtues is, by contrast, rather underdeveloped and non-specific. As a result, some of the lessons learned from discussion of the traditional models has not been taken up in treatments of IBE. Consider as an illustration, the suggestion that "unification" is one of the explanatory virtues that can figure in an IBE. Discussion (e.g. Morrison, 2000, Woodward, 2003) of unificationist models of the sort proposed by Friedman and Kitcher has shown that it is very difficult to capture the kind of unification, if any, that is relevant to explanatory goodness and that there are many example of unification in science that don't seem explanatory at all. This does not show that there is no connection between explanation and unification, but it does strongly suggest that anyone who wants to claim that potential unifying ability is a virtue that can figure in an IBE needs to be precise about what is meant by unification and how this is connected to explanation and grounds for acceptance.

 In sections 7-9 of this essay, following the discussion of underdetermination in Section 6, I will consider some of the explanatory virtues mentioned above in more detail. I will provide some simple examples showing that simplicity and unification do not connect to comparative judgments of explanatory goodness or reasons for belief in any straightforward way, thus reinforcing similar conclusions reached elsewhere in the literature. By contrast, mechanistic information is certainly relevant to successful explanation but is not linked to grounds for acceptance in a way that allows it to figure in IBE. Insofar as agreement background knowledge is relevant to successful explanation, this is only because it bears on the truth of the proposed explanation-- we can't make sense of the claim that background knowledge is relevant to how well a potential explanation would explain if true. Thus, none of these supposed virtues provide motivation for IBE.

**6. Underdetermination as a motivation for IBE**

Although it does not strictly require this assumption[[12]](#footnote-12), the idea that IBE is a legitimate inference form flourishes against the background of a picture of science that was common in the last half of the past century but is less uniformly accepted by philosophers of science today. 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 or other extra-empirical considerations) does not differentially support one of these over the others. We (allegedly) 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 non-empirical or theoretical virtues that we can identify a hypothesis as possessing, independently of ordinary empirical evidence. Considerations of potential explanatories are obvious candidates for such virtues, especially if it is thought that other candidates, like simplicity and unification can be folded into them[[13]](#footnote-13).

 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. 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. These considerations support IOE.

 My defense of IOE and the inductive strategies that involve exclusion of alternatives has several components. First, the idea that undermining alternatives is crucial to inductive support is common to many of the most plausible accounts of confirmation and gains normative support from this fact. In some cases this undermining may take the form of evidence that is deductively inconsistent with alternatives, given auxiliaries we accept, so it is obvious why acceptance of the remaining alternative is warranted. However, it is natural to extend the notion of undermining to the provision of evidence and other information that makes all of alternatives except one highly improbable. As noted in Earman, 1992, Bayesianism implements just such a probabilistic version of the eliminative ideal-- we look for evidence *e* such that, for each of the alternative hypotheses, *hi*, *i≠j*, *Pr (e/ hi)* is low and *Pr(e/hj)* is high, thus supporting *hj.* Assuming that different researchers have somewhat different priors, this assigns a central role to evidence *e* in discriminating among alternative explanations. One finds a similar idea within statistical learning theory, which is widely used in statistics and machine learning: "we start with a class of hypotheses and use the empirical data to select one hypothesis from the class" (von Luxburg and Scholkopf, 2011).

 Second, 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--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, motivated by the w-account of explanation, is that what matters for the explanatory success of these theories is the effective correctness of the dependency relations that they postulate, within the domains in which these are applied. These are what is captured by, e.g., the Newtonian gravitational force law and the field equations of GR.

 As I see it, this focus on the role of dependency relations in explanation contrasts with the role played by "ontological" commitments associated with various theories-- e.g., Newtonian theory's commitment to an ontology of gravitational "forces" as well as a notion of absolute spatial position in contrast to GR. If one focuses on such ontological commitments, as many philosophers do, it is natural to infer that, given GR’s status as a successor theory, Newtonian theory must be fundamentally false because its ontological commitments are at variance with those of GR. By contrast, 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. It is these claims about dependency relations (and not the accompanying ontology) that need to be effectively correct for Newtonian theory to explain. Again, this fits well with the w-condition account of explanation which focuses on such dependency relations and not ontological correctness. Although I lack the space for detailed discussion, readers may think of this as a version of structural realism (as described in e.g., Lorenzetti and Ladyman, 2023) applied to explanation: as far as explanation goes, the part of a theory that does explanatory work is the "structural" part where this is identified with what I have called dependency relations.

 When the parts of these theories that do the explaining are understood in this restricted way (rather than, say, as claims about exact and literal truth at all possible energy scales or about the correctness of fundamental ontology), it is easier to see how those parts can be strongly confirmed by available evidence -- they claim less than some philosophers suppose. Moreover, it is simply an empirical fact that there are no known alternatives to the above theories, again when understood as effective claims within restricted domains. 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. 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. An illustration is provided by Weinberg's (1996) argument that the combination of non-relativistic quantum mechanics (in fact, 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. Another example is provided by the demonstration by Ehrenfest and Poincare (discussed in Norton, 1993) that, given additional minimal assumptions, the only explanation of the observed blackbody spectrum is one which introduces a quantum discontinuity. 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 the generic case to be one in which there are (formulated) alternatives to the above hypotheses and theories, consistent with the available evidence, but which are taken to be less good as potential explanations and thus are taken to be less well supported for that reason. This is not what one sees in many areas of science.

 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 (in conjunction with other assumptions) exclude whole subsets of these at once[[14]](#footnote-14). Such generate and search strategies can minimize the inductive risk posed by unconceived alternatives, which are sometimes thought to pose a fatal objection to IOE. 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) proceeds 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[[15]](#footnote-15).

 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 force of some kind centered on the sun. Newton (and his successors) also used additional complicated iterative and confirmatory procedures, as described in Smith (2014) and Harper (2011)[[16]](#footnote-16). 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[[17]](#footnote-17).

 I conclude that when several alternative hypotheses are consistent with the known evidence it is often possible to find additional non-IBE evidence that supports one of these hypotheses and excludes the others. IOE is often a realistic goal. This does not show that IBE is a mistaken strategy, but it does show that there is an alternative to it.

**7. Simplicity and Unification as Explanatory Virtues**

I noted above that there are some conceptions of explanation (or at least of explanatory virtues) which may seem to have a connection with grounds for belief that is lacking for other conceptions. Simplicity and unification are two obvious candidates for this role and in what follows I consider both.

 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 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*.

 Assume *h1*- *h2* are mutually exclusive and exhaustive.

 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 (if true) 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 explanation of *E* and, to the extent we are willing to follow IBE, that we have reasons to believe *h1* rather than *h2.*

 This conclusion may seem plausible to some readers but, in my view, it is very strongly at odds with scientific practice. 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-- in other words, to follow the eliminative strategy discussed earlier. 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 both correlated with Z and are independent conditional on *Z* (and we can exclude the possibility that *X* causes *Z* which causes *Y*) this suggests 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 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 alternatives would if true provide the best explanation of *E* and then concluding on that basis that this is the best supported hypothesis. Rather, good scientific practice is to look for additional evidence besides *E* that allows one to discriminate among these alternative hypotheses. It may or may not 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.

 The issue raised by this example also generalizes: Supposing, as the w-account claims, explanation is a matter of correctly identifying dependency relations, this is independent of the number of such relations that happen to be present, the number of variables they involve and so on-- all of which presumably matter for simplicity.

 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 all, at least in the way advocates of IBE claim. Suppose, as is 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 a non-question-begging basis for this judgment might be. 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 *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. As argued above, on this view explanations can differ in the extent to which they describe what an explanandum depends 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 further grounds for claiming that qua potential explanation, it is more or less "good" depending on how simple it is. As this example illustrates, explanations that differ in "simplicity" can nonetheless do equally well in satisfying the requirements of the w-account[[18]](#footnote-18). Similar points apply to unification, construed as an explanatory virtue. 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.) Although the notion of explanatory unification is unclear in important respects[[19]](#footnote-19), if this notion means anything definite at all, *h2* is surely more unified that *h3*[[20]](#footnote-20). Nonetheless it seems highly problematic to infer, on the basis of IBE, that there are stronger grounds for accepting *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 and can figure in IBE 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. At the very least, this requires a more detailed engagement with the literature on explanatory unification than is found in the IBE literature.

 The suggestion that unification (at least without further prescification) 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, 1991/2004.) In fact, things are far more complicated than this simple gloss suggests. Of course, Newtonian theory is a highly successful explanatory theory and 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 worthy of belief *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 well 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. These involve examples like Maxwell's unification of electricity and magnetism, the electroweak theory (unifying the electromagnetic and weak forces), 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.

 The decision to focus on the explanatory virtues and bypass any general account of explanation, raises the following question: 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 linked to unification and simplicity, if it is really just unification and simplicity that matter for grounds for belief.

 I draw the opposite moral from the above examples. 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 or unifying. 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.

**8. Mechanistic Information**

Another explanatory virtue invoked in discussions of IBE is the provision of information about mechanisms (e.g. Lipton, 1991/2004)In contrast to the supposed virtues of simplicity and unification, I take it to be uncontroversial that mechanistic information is often explanatorily valuable. The problem is rather that it is hard so see how such information could contribute to grounds for belief in the way IBE requires. Suppose that (*E*) a correlation between aspirin ingestion *A* and headache relief *R* is observed. Consider the following two explanations of *E*. *h1*: *A* causes *R*, *h2*: *A* causes *R* and this happens via mechanism *M.* Assuming that the provision of mechanistic information is an explanatory virtue, it seems unavoidable that *h2* would *if true* provide a better explanation of *E* (and presumably other explananda as well).. This follows on the w-account since mechanistic information if correct tells about a wider range of dependency relations, including those involving mediating causes (See, e.g., Woodward, 2002). But it seems obvious that this fact about potential explanatoriness lends no support at all for that part of *h2* which has to do with *M*. *E* tells us nothing about whether the operative mechanism is *M* or something else. To establish that *M* is the mechanism we need additional evidence besides *E* that favors *M* over alternative claims about the mechanism by which by *A* causes *R*. Note also how this case illustrates the clear difference between the claim that in virtue of possessing feature *F*, a hypothesis *h* if true would be highly informative and the claim that the presence of *F* provides support for the truth of *h*. Claims about mechanisms are valuable because of the explanatory information they provide *when they are correct*, and not because their merely potential explanatoriness provides them with inductive warrant.

**9. 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 McCain, 2019 and Lipton, 2004, among others) that fit with background knowledge is among the explanatory virtues that can guide IBE. It is hard to understand how this is supposed to work. 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.

 Consider the following example from McCain (2019) which he intendeds to illustrate the (legitimate) role that background information can play in IBE. The evidence is that noodles are missing from the refrigerator, and the alternative hypotheses are (*h1*) that they were eaten by one's roommate or (*h2*) that, instead, noodle thieves broke into the refrigerator and took only those. Suppose our background knowledge *B* consists of the information that thieves who break into houses and steal only noodles are rare to non-existent. McCain claims that *B* can be regarded as an ingredient in an IBE that supports *h1*. However, *B* strongly supports the assessment that *h2* is false and hence should not be regarded as a potential explanation for the purposes of IBE. We thus don't need IBE to infer that *h1* is more strongly supported than *h2.* On the other hand, suppose we *are* willing to consider *h2* as a potential explanation. When we ask how well *h2* would explain *if true,* we thus must be considering a case to which *B* somehow does not apply. (Supposing that *h2*is true, it must be that the outside thieves *are* operative, contrary to what is suggested by *B*.) In such a case, it seems that *h2* would explain the absence of noodles just as well as *h1* and consequently, via an IBE, has an equal claim to acceptance, which is not the correct conclusion.

 In connection with the examples in this and previous sections (7-9) a referee has suggested that what they establish is a dilemma: one can accept either IBE or the w-account of explanation but not both. Of course I agree that the w-account is inconsistent with IBE but I think a stronger conclusion is warranted. The examples concerning simplicity, unification, mechanism, and background knowledge are plausible counterexamples to claims that these are explanatory virtues that can be used in IBE, independently of whether one accepts the w-account of explanation. The w-account provides an overarching rationale for why the counterexamples work but, even apart from this, it remains true that the use of simplicity and unification described in Section 7 are at odds with scientific practice, that merely adding a possible mechanistic detail to a hypothesis does not make it more belief worthy, and that fit with background knowledge is not a distinctively explanatory virtue.

 **10. Conclusion and a Comparison**.

Bird (2022) has recently discussed both IBE and IOE (see also Bird, 2018, 2010, 2007). The editors have asked me to compare my view with his. In what follows, I do this with an eye to also summarizing my views and what is distinctive about them. As remarked earlier, Bird's understanding of IOE is essentially the same as mine--it proceeds by assuming that there is some explanation of the phenomena of interest, then considering the alternative hypotheses that are potential explanations of this and finally getting evidence that rules out all but one of these via a process of something like eliminative induction. Bird claims that only when this process is carried out that we have *knowledge*. I haven't described my claims in this essay in terms of "knowledge", preferring talk of "hypotheses for which there are grounds for acceptance", but for present purposes have no objection to Bird's terminology.

 Since Bird agrees that successful IOEs can be carried out, he also agrees that standard claims about the inevitable underdetermination of theory by evidence are mistaken. With respect to such underdetermination claims, I see my discussion as going beyond Bird's in several respects (which is not to say that what I say is inconsistent with Bird's claims.) First, I emphasize the importance of systematic search to generate alternative potentially explanatory hypotheses as a way of addressing Stanford's problem of unconceived alternatives. Second, I emphasize the idea that the competing potential explanations among which we choose involve "effective" hypotheses and theories and that these concern claims about dependency relations (rather than, e.g., ontology). As I argue, this makes the underdetermination problem much more manageable. I also note that without restrictions of the sort described, the problem of unconceived alternatives seems intractable for IOE. As noted above, despite the very strong evidence for Newtonian gravitational theory circa 1850 or so, there existed at that point an alternative not yet conceived potential explanation provided by General Relativity which of course was not excluded by the available evidence. If GR is treated as an alternative, competing hypothesis to Newtonian theory, this threatens to undermine the use of IOE to support Newton's theory. As I argue, we can avoid this conclusion by taking the relevant alternative hypotheses to which we apply IOE to be those that make different claims about dependency relations within the domain of applicability of Newtonian theory. Some restriction of this sort seems necessary to many applications of IOE, particularly those in which the competing alternatives are whole theories that may stand in hierarchical relationships rather than competing hypotheses about particular causes.

 Next, when it comes to the mechanics of how eliminative induction works, I fully agree with Bird that this often draws on knowledge that is "theoretical” (following Williamson, Bird describes this as "evidence" but I don't think this matters for present purposes) rather than purely "observational". However, I offer specific suggestions involving physics examples, about some of the forms that use of theoretical information in inference takes. For example, it is sometimes possible to prove, given uncontroversial assumptions, that the only hypothesis capable of explaining the available evidence must take a certain specific form-- the result of Weinberg's discussed in Section 7 as well as the proof that any hypothesis capable of explaining the observed black body spectrum must rely on energy quantization are illustrations. Results of this sort illustrate specific ways in which "theoretical" knowledge can be used to address underdetermination problems.

 In his 2022 Bird also claims that although IBE does not yield knowledge it can nonetheless it can nonetheless provide " a rational basis for preferring one hypothesis to another" and for judging the "relative plausibility of hypotheses”. Moreover, he seems to accept that the standard list of explanatory virtues (simplicity etc.) are what should go into the use of IBE for this purpose. Here our views differ. My discussion in Section 7 claims that some standard explanatory virtues like simplicity and unifying ability are, at least when applied to the cases I discuss, not virtues at all and that appeal to them does *not* support judgments of relative plausibility of the sort Bird contemplates. My discussion also differs from Birds in that I criticize other appeals to explanatory virtues in IBE such as mechanistic information and consistency with background knowledge.

 More generally, my discussion goes beyond Bird's in arguing that discussions of IBE should take place against the background of an explicit account of explanation, rather than just a list of supposed explanatory virtues. Some accounts of explanation fit naturally with and support IBE and others do not and any supporting account must have the right structure to support the kinds of comparative judgments of explanatory goodness on which IBE relies. In particular I have argued that one prominent account of explanation-- the w-account-- does not support IBE. I also emphasize the idea, which I don't think is present in Bird's discussion, that the truthlike condition that is appropriate to impose on explanation is a notion of effective validity within a domain and that that this should be understood as a requirement on the dependency relations postulated by an explanation rather its other components.

 Finally, I have emphasized the idea that explanation is an important aim or goal in science in its own right (although far from the only aim) and that we need the idea that there are other scientific aims in addition to the discovery of truths. As noted above, Bird talks instead of knowledge as an aim of science and argues that the kind of knowledge that provides understanding of explanatory relations is important in science. I lack the space to comment on the knowledge/truth contrast but, putting that aside, we seem to agree on the role of understanding/explanation as a goal. I have tried, however, to say a bit more about the role of this goal and how it gives scientific inquiry some of its distinctive features such as its willingness to postulate unobservables and modal structure. I have also sought to emphasize that discovering hypotheses that explain is a different goal than that of finding hypotheses that are simple or unifying.

# References

# Achinstein, P. (1991) *Particles and Waves*. New York: Oxford University Press.

# Achinstein, P. (2001) *The Book of Evidence*. Oxford: Oxford University Press.

Akaike, H. (1974) "A New Look at Statistical Model Identification" *IEEE Transactions on Automatic Control* 19: 716-23.

Bird, A. (2022) *Knowing Science* Oxford: Oxford University Press.

Bird, A. (2010) "Eliminative Abduction: Examples from the Medicine" *Studies in History and Philosophy of Science* 41: 345-52.

Bird, A. (2018) "Inference to the Best Explanation, Bayesianism and Knowledge" in McCain, K. and Poston, T. (eds) Best Explanations. Oxford: Oxford University Press, 97-120.

 \_

Bird, A. (2007) “Inference to the Only Explanation” *Philosophy and Phenomenological Research* 74: 424–32.

# Cabrera, F. (2021) "String Theory, Non-Empirical 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*

Dorling. J. (1973) "Demonstrative Induction: Its Significant Role in the History of Physics" *Philosophy of Science* 40: 360-372.

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.

Friedman, M. (1974) “Explanation and Scientific Understanding”, The Journal of Philosophy, 71(1): 5–19.

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

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

Hicks, M., Jaag, S. and Loew, C. (2023) *Humean Laws for Humean Agents*. New York: Oxford University Press.

 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. (1991/2004) *Inference to the Best Explanation* Routledge: London and New York

Lorenzetti, L. and Ladyman, J. (2023) "Effective Ontic Structural Realism" Forthcoming, *British Journal for the Philosophy of Science*.

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.

Norton, J. (2024) *The Large-Scale Structure of Inductive Inference*.BSPS*open*/University of Calgary Press,

 Norton, J. (2021)  [*The Material Theory of Induction*](https://sites.pitt.edu/~jdnorton/homepage/cv.html#material_theory). BSPS*open*/University of Calgary Press,

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

# Platt, J. (1964) "Strong Inference" *Science* 146: 347-53.

Prasetya, Y. (2021) "Which Models of Scientific Explanation are (In)Compatible with IBE?"*British Journal for the Philosophy of Science* 75

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. (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

Sanford, P. (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 Rowe, D., Sauer, T. and Walter, S. 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.

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.

# von Luxburg, U. and Scholkopf, B. (2011) "Statistical Learning Theory: Models, Concepts, and Results" in Gabbay, D. and Woods, J. (eds) *Handbook of the History of Logic*, volume 10: 651-706. Amsterdam: Elsevier.

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 Philosophy of Science.*

Woodward, J. (2002) "What is a Mechanism? A Counterfactual Account.”  *Philosophy of Science* 69: S366-77.

 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.

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. [↑](#footnote-ref-1)
2. I take estabilshing truths to be generally accepted shorthand for something more complicated, including providing evidence or grounds for claiomed truths, avodance of acccepting falsehoods and so on. [↑](#footnote-ref-2)
3. The idea that there are informational virtues in addition to confirmational ones and that explanation is an informational virtue is bruited in Salmon, 2001 and is discussed at greater length in Carbrera, 2017. [↑](#footnote-ref-3)
4. To say a bit more about this: The literature on simplicity is enormous, spanning many different disciplines besides philosophy. As remarked above, philosophical discussions of simplicity, particularly in the context of IBE, tend to invoke a rather generic and undifferentiated notion, supposedly applicable across many different domains of inquiry and independent of empirical assumptions about those domains, the population from which one is sampling and so on. Philosophical discussion also sometimes seems to assume that the claim that simplicity is linked to truth and inductive support is so obviously true that there is no need to provide further details such as a precise characterization of the relevant notion of simplicity and an argument-- mathematical or otherwise-- linking this to inductive support or grounds for acceptance. This contrasts with the approach taken in disciplines like statistics and machine learning, where the simplicity notions are sometimes employed but where these are far more specific and mathematically precise (involving, e.g., counts of free parameters, as in Akaike, 1974, or shattering coefficients, as described in von Luxburg and Scholkopf, 2011). Here mathematical arguments are developed connecting these characterizations to precise notions related to acceptance or inductive support, such as expected success in out of sample prediction. Moreover, it is made clear in this literature that the claimed connections between these simplicity concepts and anything related to induction holds only for very specific background assumptions (e.g., that the sample is drawn i.i.d. from a fixed probability distribution.) I mention this because it shows that (i) it is not unreasonable to ask for precise results about the connection between notions like simplicity and inductive support and (ii) there are strong reasons for thinking that such results do not exist at the level of generality at which philosophical discussion is conducted. [↑](#footnote-ref-4)
5. For recent discussions of this sort of justification for appeals to simplicity in the context of the Best System Account of laws of nature, see the papers in Hicks et al., 2023. [↑](#footnote-ref-5)
6. A referee observes -- and I agree-- that it is common in many scientific publications to include a "General Discussion" in which a theory is suggested that best explains the evidence we have so far, but without any claim that this theory should be accepted or is definitely correct and with the recogntion that it may be possible to get additional evidence that discriminates between this theory and rivals. This does not count as a case of IBE as I understand it since it is not claimed that the theory should be accepted and it is recognized that any underdetermination may be transient and thus addressed by additional evidence. I'll add that I understand the notion of "pursuit" very broadly-- it includes, as in the referee's example, theories that are judged worthy of consideration, entertainment, and further exploration and development.

 [↑](#footnote-ref-6)
7. For example, after describing the inference by Alvarez and Alvarez to an asteroid impact as the cause of a massve extinction event as a case of eliminative inference, Woodward, 1990 writes: "The eliminative strategy also reflects the idea that warranted causal inference is *not*, as is sometimes claimed, a matter of inference to the best explanation; it is rather a matter of inference to the *only* acceptable explanation among some range of alternatives". [↑](#footnote-ref-7)
8. That Thomson's remarks are plausibly interpreted in terms of IOE is also noted by Nyrup (2015). [↑](#footnote-ref-8)
9. It has recently been claimed (e.g., Rice, 2021) that scientific hypotheses and models are full of falsehoods and distortions (in the form of "idealizations" etc.). I lack space for detailed discussion but, again, what matters from my point of view is whether the models get the relevant dependency relations approximately right. To the extent that they do, they can be used to explain, even if the models make false claims in other respects. For example, a two-dimensional lattice model such as the Ising model can successfully capture dependency relations involving phase transitions in a continuous three-dimensional system even though the model, interpreted literally, makes false claims about the dimensionality etc. of the system. Interestingly on one interpretation, Rice seems to agree. He emphasizes that true counterfactuals about the target system can be extracted from models containing falsehoods. On my view, it is such counterfactuals and the dependency relations with which they are associated that figure in successful explanations. [↑](#footnote-ref-9)
10. MOND proposes a modification of Newton's second law to explain observed features of galaxy behavior in a domain in which unmodified versions of Newtonian mechanics are usually taken to be applicable. [↑](#footnote-ref-10)
11. Arguably this is sometimes part of what goes on when people adopt conspiracy theories. As a mischievous 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. [↑](#footnote-ref-11)
12. 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. [↑](#footnote-ref-12)
13. See, e.g, McCain and Poston (2023) for such an argument, including the claim that IBE is widely used in science. [↑](#footnote-ref-13)
14. 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. One does not have to be passive in the face of the possibility that there are unconsidered alternatives. [↑](#footnote-ref-14)
15. A similar point holds for other designed based strategies in causal inference such as the use of instrumental variables-- when sucessful these also work by ruling out alternative causal hypotheses except for one. Of course one may be mistaken in thinking that the conditions for the successful application of these strategies hold, but it is extremely implausible that they can never be successfully used to rule out alternatives in the way required by an IOE. [↑](#footnote-ref-15)
16. Dorling, 1973 describes a number of other examples of such deductions from the phenomena. [↑](#footnote-ref-16)
17. Another relevant consideration is this: the fact that a hypothesis *h* makes extremely precise empirical predictions that are well confirmed can itself help to make it implausible that there are viable alternatives to *h.* For example, the standard model successfully predicts the value of the magnetic moment of the electron to a value of one part in a trillion. If there is a viable alternative to the standard model "out there" in theory space, this would need to make the same prediction while nonetheless differing from the standard model in other respects (and predictions). It seems plausible that physicist's understanding of the possible alternatives to the standard model is secure enough that it strongly supports the judgment that there is no alternative with the right properties. [↑](#footnote-ref-17)
18. This may evoke the following response: Although simplicity when understood as above may not be an explanatory virtue, there is some other notion of simplicity (suitably restricted and refined) that is such a virtue and that can be legitimately employed in an IBE. All that I can say is that this puts the onus on the advocate of IBE to distinguish the explanation-linked notion of simplicity from others and to explain why it is linked to reasons for belief in the way advocates of IBE claim. [↑](#footnote-ref-18)
19. Again see Morrison, 2000 and Woodward, 2003. [↑](#footnote-ref-19)
20. For example, the best known unificationist accounts of explanation -- those developed by Friedman and Kitcher will judge the single common cause explanation a better explanation than the two common cause explanation. [↑](#footnote-ref-20)