Abductively Robust Inference

Finnur Dellsén

This is a preprint of a paper published in *Analysis*; please cite published version.

Abstract: Inference to the Best Explanation (IBE) is widely criticized for being an unreliable form of ampliative inference – partly because the explanatory hypotheses we have considered at a given time may all be false, and partly because there is an asymmetry between the comparative judgment on which an IBE is based and the absolute verdict that IBE is meant to license. In this paper, I present a further reason to doubt the epistemic merits of IBE and argue that it motivates moving to an inferential pattern in which IBE emerges as a degenerate limiting case. Since this inferential pattern is structurally similar to an argumentative strategy known as Inferential Robustness Analysis (IRA), it effectively combines the most attractive features of IBE and IRA into a unified approach to non-deductive inference.

1. Inference to the Best Explanation

On standard formulations, *Inference to the Best Explanation* (IBE) is a form of non-deductive inference in which one infers a hypothesis because it would, if true, provide a better explanation of one's evidence than any other available competing explanatory hypothesis.[1] An explanation is considered 'better' than another to the extent that it exhibits various explanatory virtues – e.g. parsimony and explanatory scope – which jointly constitute the explanation's 'loveliness' (Lipton 2004: 59-62). Although early discussions of IBE saw it as a fundamental and free-standing form of inference warranting full belief in its conclusions, it has now become more-or-less standard for proponents of IBE to view it as an approximation to, or heuristic for, some form of probabilistic reasoning in which rational agents assign subjective probabilities to hypotheses.[2] In Lipton's influential turn of phrase, 'Inference to

[1] See, e.g., Harman 1965, Thagard 1978, Lycan 1988, and Lipton 2004.

[2] See, e.g., Okasha 2000, McGrew 2003, Lipton 2004, and Henderson 2014.

the Best Explanation proposes that loveliness is a guide to likeliness (a.k.a. posterior probability)' (2004: 115).

It is worth noting that many inferences that are commonly characterized as instances of IBE are only indirectly inferences in which one compares explanatory hypotheses with respect to their loveliness. While paradigmatic instances of IBE involve inferring a hypothesis H from the fact that it best explains some evidence E, many proponents of IBE also classify it as IBE when one infers H in virtue of H being entailed by some other hypothesis H^* that best explains E. For example, Lipton suggests that in Newton's days IBE licensed an inference from various terrestrial experiments, E_T, to laws governing planetary motion, L_P, since E_T is best explained by Newton's laws of motion, L_N, which in turn entail L_P. Thus, while L_P certainly doesn't explain E_T, L_P is still inferable from E_T via IBE (Lipton 2004: 63-64). In fact, the possibility of inferring *indirectly* via IBE in this way was already exploited by Harman (1965: 91) when he argued that all enumerative inductions could be construed as instances of IBE.[3]

One familiar criticism of IBE attacks the idea that the various explanatory virtues that jointly constitute explanatory loveliness are correlated with rational probability assignments (see, e.g., van Fraassen 1985; Barnes 1995; Bartelborth 2005). In this paper, however, I will set such worries aside and instead assume, if only for the sake of the argument, that some such connection holds (at least *ceteris paribus*) given a suitably chosen set of 'explanatory virtues' for IBE to operate with. What I will be concerned with is the peculiar structure of IBE, which involves *comparing* a set of *available* competing hypotheses before inferring that the loveliest such hypothesis at least somewhat likely to be true. These structural features of IBE correspond closely to many actual cases of theory-choice in science and philosophy, where making an inference depends crucially on comparisons between extant alternatives as opposed to evaluations of individual theories *in vacuo*. Indeed, it is at least partly because the structure of IBE seems to correspond to actual scientific and philosophical practice that realists of various stripes often endorse and defend IBE.

[3] According to Harman, the best explanation for an evidential proposition of the form (H_O) 'All observed As are Bs' is the proposition (H_A) 'All As are Bs', which in turn entails the proposition (H_N) 'The next observed A will be B'. Thus, for Harman, IBE warrants inferring H_N from H_O, via H_A, even though H_N certainly does not explain H_O.

The best-known criticism of IBE on structural grounds is no doubt van Fraassen's objection that an explanatory hypothesis may provide the best explanation of the currently available competitors only because we haven't (yet) considered a hypothesis that would explain even better. Thus van Fraassen contends that IBE might well lead us to choosing 'the best of a bad lot' (1989: 142-143). Indeed, for this reason, van Fraassen claims that we should treat any conclusion of IBE as a random member of a set of explanatory hypothesis most of which are false (1989: 146). While that might be something of an exaggeration, the fact remains that the bad lot objection points to a significant epistemic risk inherent in the structure of IBE. For the purposes of this paper, however, I will be setting this problem aside. Accordingly, I shall assume that the set of available hypotheses from which one is choosing in IBE is *ideal* in the sense of including a correct explanation.

Another structural problem – what Douven (forthcoming) refers to as *the asymmetry problem* – arises even in situations in which the true explanation has been considered. While the fact that a hypothesis H explains one's evidence better than some other hypotheses might indicate that H is likelier to be true than those other hypotheses, this *comparative* claim is compatible with H being very improbable indeed.[4] However, nearly all proponents of IBE are committed to IBE licensing an *absolute* judgment to the effect that H is least somewhat likely to be true.[5] This problem is not solved by adding the caveat that the explanation provided by the inferred hypothesis H must also be 'satisfactory' (Musgrave 1988: 238-239) or 'good enough' (Lipton 2004: 63, 154), since that is meant to impose only a minimal constraint on the explanatory loveliness of H.[6] My discussion below has implications for how

[4] Note that this holds even when the truth is among the hypotheses that are being compared, since the truth could be very improbable given the evidence, e.g. when one's evidence is meager or misleading.

[5] A possible exception is Kuipers (2000).

[6] Of course, one could impose a much stronger constraint to the effect that the best explanation must be so explanatorily lovely as to guarantee that the best explanation is more probable than some threshold, e.g. 0.7 or 0.9. (Although this idea cannot be attributed to either Musgrave or Lipton, Climenhaga (forthcoming) considers a suggestion along these lines.) However, in contrast to the standard formulation of IBE this would require that there is some way of specifying what it is for any given hypothesis to be explanatorily lovely in absolute as opposed to merely comparative terms. This is doubtful for at least two distinct reasons. First, it is unclear how to set any absolute threshold for explanatory loveliness such that exceeding the loveliness-threshold correlates with exceeding the associated probability-threshold in all (or even just most) cases. Seemingly unanswerable questions of the form 'Exactly how simple/consilient/etc. must a given hypothesis be for it its probability to exceed

to address this problem in that I argue for an inferential pattern that significantly ameliorates the epistemic risks of IBE in this regard. However, my main focus will be on a somewhat more specific structural problem with IBE – one that has hitherto not been given due attention.

2. The Problem of Multiple Plausible Rivals

There are several live scientific hypotheses that purport to explain the origin of life on Earth, i.e. why living organisms arose from non-living matter on this planet a few billion years ago. Chief among these is a hypothesis known as *RNA world*, which roughly states that life began with the formation of RNA molecules that were capable of self-replication and which would later evolve into the DNA and protein molecules which are the building blocks of today's living organisms. The explanation provided by the RNA world hypothesis is arguably quite lovely indeed, e.g. in that it posits no new kinds of entities beyond the already familiar RNA molecules and yet elegantly explains, if true, how genetic information would have been stored, replicated, and transmitted in the way required for living organisms to evolve. To be sure, biologists have also proposed several other explanations, the most plausible of which arguably involve positing some other self-replicating medium, e.g. various other nucleic acids such as PNA (peptide nucleic acid), TNA (threose nucleic acid), or GNA (glycol nucleic acid). Furthermore, some biologists also take seriously hypotheses according to which life began with the formation of metabolizing cells rather than any kind of genetic material.

Interestingly, even the majority of biologists who consider the RNA world hypotheses to be by far the best explanation of the available evidence are quite hesitant to infer that it is

---

the threshold?' would have to be addressed. Second, probability appears to behave very differently from absolute levels of explanatory loveliness in that the probability of one explanatory hypothesis inevitably takes away from the probability of competing explanatory hypotheses, which is not true of explanatory loveliness. To see why, note that a group of incompatible explanatory hypotheses $H_1,...,H_n$ may all be very lovely in an absolute sense, but they cannot all be very probable (on pain of violating the axioms of the probability calculus).

true (and that its alternatives are false).[7] It is therefore doubtful that IBE is descriptively correct in this case. More importantly, biologists' reluctance to infer the loveliest explanatory hypothesis seems, contrary to IBE, normatively appropriate. After all, the sheer multitude of available competing explanations – each one of which has a non-negligible likelihood of being correct – suggests that the truth is quite likely to lie with one of these many alternatives. To be sure, it may still be perfectly reasonable for biologists to use the RNA world hypothesis as a working hypothesis to guide further scientific inquiry, since one can do no better than to operate with the best theory available. However, the point remains that in this case the epistemic merits of inferring in accordance with IBE is undermined by the availability of multiple plausible competing explanatory hypotheses – a factor that IBE simply ignores.

In case proponents of IBE take issue with my description of this particular case, let us note that the problem here is perfectly general. Suppose $\{H_1,...,H_n\}$ is a set of available competing explanatory hypotheses (assumed to be *ideal* in order to set aside the bad lot problem). According to IBE, whether a hypothesis H_i in this set is inferable will depend exclusively on whether H_i is lovelier than other hypotheses in the set (and perhaps also on whether H_i is 'sufficiently' lovely). Intuitively, however, the extent to which it is reasonable to infer H_i also depends on whether there are many plausible alternatives to H_i that are lovely enough to be taken seriously as alternative explanations for E (even if they are not nearly as lovely as H_i itself). The general problem here for IBE is that it has no resources for taking into account the possibility that an inference to H_i may be undermined by the availability of *multiple plausible rivals* to the loveliest explanatory hypothesis. Of course, IBE could still be a good rule of thumb in many cases; nevertheless, the problem demonstrates that IBE ignores a factor that is relevant to how reasonable it is to infer the hypothesis that best explains the available evidence.

---

[7] This common attitude is summed up nicely in Bernhardt's sympathetic discussion of the hypothesis, entitled 'The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)' (Bernhardt 2013: 1).

It is worth noting that those who favor other explanations, such as the PNA, TNA and GNA world hypotheses, are also hesitant to infer that these other theories are true. The point here is that biologists are generally hesitant to infer that the theories they think provides the best explanation of the evidence is in fact true, contrary to what IBE recommends.

One might think that this problem can be solved with minor modifications to IBE by either requiring that the inferred hypothesis *H* provide a *far better* explanation than competing available hypotheses, or by requiring that we set some fairly high *absolute threshold* for *H*'s explanatory loveliness.[8] However, it is easy to see that no solution of this kind will get to the heart of the problem. Consider a variation of the previous case in which there are only two plausible competing explanations, e.g. a case in which all available explanations for the origin of life have been empirically ruled out except for the RNA world hypothesis and the corresponding PNA world hypothesis. In that case, it would surely be *more reasonable* to infer that the RNA hypothesis is true than it was in the original case.[9] And yet altering the case in this way keeps fixed both the absolute explanatory loveliness of the loveliest explanation and the relative explanatory loveliness of the loveliest and second-loveliest explanation. Accordingly, the modified versions of IBE we are now considering wrongly predicts that this altered case is on a par with the original case with regard to whether it is reasonable to infer the RNA world hypothesis. It should be clear, then, that we need to look elsewhere for a solution to the problem.

3. Abductively Robust Inference

As a first step towards solving the problem, notice that when it comes to explaining the origin of life of Earth there is an important claim that holds true on a number of the most plausible available competing explanatory hypotheses, viz. that life on Earth began with the formation of a genetic replicator of some sort or other. To be sure, this claim is false according to the metabolism-first hypothesis mentioned above, but that hypothesis is arguably less plausible than each of the RNA, PNA, TNA, and GNA world hypotheses – all of which do entail that claim. Unsurprisingly, then, the hypothesis that life arose from a genetic replicator – a

[8] See the previous footnote for further problems with the second of these two suggestions.

[9] This is evidenced by the fact that the process of incrementally confirming scientific theories often involves eliminating alternative explanations for a given phenomenon even when those alternative explanations are not considered to be the most plausible. A case in point is are numerous alternatives to Einstein's general theory of relativity that were proposed and systematically refuted in latter half of the 20th century – for discussion, see Earman 1993: 173-181.

hypothesis entailed by, but not equivalent to, the RNA world hypothesis – can be found in many biology textbooks and has even entered some influential biological definitions of 'life'.

The key insight here is one that is familiar from a relatively underexplored inferential strategy discussed in a different context by Woodward (2006). Suppose we have some data *D* from which we hope to infer a conclusion *S*. Suppose also that *D* does not by itself imply *S*, so that some additional assumption(s) are required for *S* to be inferred from *D*. Furthermore, suppose that a number of competing possible assumptions $A_1,...,A_m$ are available, each one with some plausibility. If *D* implies *S* given any one of these assumptions (or if the probability of *S* given *D* and each assumption exceeds some particular threshold),[10] then the inference from *D* to *S* is said to be *inferentially robust* with respect to $A_1,...,A_m$. Now, the idea behind what Woodward calls *Inferential Robustness Analysis* (IRA) is that if it is known that one of the competing assumptions $A_1,...,A_m$ is true (though it isn't known which one is true) then the fact that the inference from *D* to *S* is inferentially robust with regard to $A_1,...,A_m$ provides a strong reason for us to infer *S* from *D*.

Woodward rightly criticizes IRA on the grounds that the conditions that would need to be satisfied to infer via IRA are 'very strong', so that 'its range of application looks rather limited' (Woodward 2006: 222).[11] As Lloyd (2015: 58) observes, this is a huge understatement. After all, it is hard enough to imagine any interesting cases in which we know for certain that one (but not which one) of some competing assumptions $A_1,...,A_m$ is true; it is even harder to think of cases in which each one of these assumptions (together with some data *D*) entails any remotely interesting conclusion *S*. So, while IRA would certainly provide

[10] In what follows, I ignore the caveat in the parenthesis since I will be concerned exclusively with cases in which there is a logical entailment between *D* and *S* given any of the assumptions $A_1,...,A_m$.

[11] Woodward (2006: 223) briefly mentions (but does not endorse) the suggestion that the problem may be alleviated by weakening these conditions, e.g. by using a weaker notion of robustness on which the inference from *D* to *S* need only hold under 'most' of the competing assumptions $A_1,...,A_m$. However, this suggestion is clearly problematic in that some of the competing assumptions $A_1,...,A_m$ may be far more plausible than others, in which case an inference from *D* to *S* may be very weak even if it holds on a majority of the assumptions $A_1,...,A_m$.

a very strong reason to infer claims that are robust in the above sense, we would hardly ever be in a position to identify inferentially robust claims.[12]

However, I now want to suggest that the core insight of IRA – viz., that the robustness of an inference under a variety of competing assumptions is an indicator of truth – can still be put to good use within a broadly-speaking 'explanationist' framework in which available hypotheses are compared with respect to their explanatory loveliness. The basic idea is that a claim may be inferred if it is entailed by all of the available competing explanatory hypotheses that do best on whatever explanatory virtues are taken to constitute explanatory loveliness. To make this more precise, we define a family of notions of 'abductive robustness' as follows (where k is an arbitrary natural number):

> For a given set of available competing hypotheses $\{H_1,...,H_n\}$ that potentially explain some evidence E, a claim C is said to be *abductively robust*$_k$ iff C is entailed by all of the k loveliest such hypotheses in $\{H_1,...,H_n\}$.

We use this to construct a family of corresponding inference rules:

> *Abductively Robust Inference*$_k$ (ARI$_k$): If C is *abductively robust*$_k$ relative to an ideal set of available competing hypotheses $\{H_1,...,H_n\}$ that potentially explain E, then infer C from E.

It would be easy to construct related, and perhaps more sophisticated, inferential patterns using the same basic idea. For example, one could define a gradable notion of explanatory robustness that varies with both the number of available explanatory hypotheses that entail C and the levels of explanatory loveliness exhibited by those hypotheses.[13] It is also worth emphasizing that ARI$_k$ should, much like currently standard conceptions of IBE, be viewed as a heuristic inferential pattern appropriate for cognitively limited beings rather than as a prescription for ideal epistemic agents.

---

[12] Indeed, notice that IRA seems to collapse into a deductive form of argument given the condition that we know that one of the assumptions is true since the data D and the disjunction of $A_1,...,A_m$ (a disjunction that would be known to be true) would logically entail the conclusion S.

[13] I leave the precise articulation of these inferential patterns to future work. It is worth noting that although patterns of this kind would arguably be more sophisticated from an epistemological point of view, they would also require its users to evaluate the absolute explanatory loveliness of all the hypotheses that entail C. This is problematic for the reasons given in footnote 6.

I refer to ARI_k as an *inferential pattern* because we obtain distinct inference rules depending on what value we give to the variable k. In choosing a value for k, one is confronted with a familiar trade-off between minimizing epistemic risk and maximizing applicability. For example, a higher value for k decreases the epistemic risk of inferring in accordance with ARI_k, but it also decreases the number of situations in which the rule would be applicable (since less is generally implied by every member of a set than by every member of its proper subsets). It would be unwise to try to fix once and for all a specific value for k since different specifications might be appropriate for different purposes. For example, if it is important that some conclusion C be inferred only if C is almost certainly true, then it makes sense to employ an instance of ARI_k in which k is quite high; in other circumstances, k can be considerably lower. That said, the more interesting versions of ARI_k will arguably assign fairly moderate values to k. To see this clearly, let us consider what inference rules are obtained in the two extreme cases in which k equals n and 1 respectively.

On the one hand, setting $k = n$ gives us a very safe inference rule – ARI_n – in which the entire epistemic risk consists in the possibility that all of the available hypotheses are false. It is worth noting that since ARI_n requires the conclusion C to hold in *all* of the available competing hypotheses $\{H_1,...,H_n\}$ that potentially explain the evidence E, ARI_n does not require its users to compare any of the hypotheses $\{H_1,...,H_n\}$ with regard to their explanatory loveliness. Put differently, evaluating the comparative loveliness of some group of hypotheses $\{H_1,...,H_n\}$ is irrelevant when the question is whether a conclusion C follows from E given each of the n hypotheses. In this respect ARI_n, resembles IRA, which also does not require any estimation of comparative explanatory loveliness. Unfortunately, however, ARI_n also inherits the aforementioned applicability problem for IRA, in that we are rarely if ever able to infer any substantial consequences from *all* available competing explanatory hypotheses. Thus, as with IRA, the range of applications of ARI_n would be extremely limited. As soon as we have even a single hypothesis H_i in our set of available competing explanatory hypotheses $\{H_1,...,H_n\}$ such that E and H_i do not entail C, ARI_n becomes inapplicable.

On the other hand, by setting $k = 1$ we obtain an inference rule – ARI_1 – according to which one may infer a claim C from E if C is entailed by the single loveliest of the available competing hypotheses that potentially explain E. Now, recall (from section 1) that IBE also licenses inferences to the deductive entailments of the loveliest available explanatory

hypothesis. Interestingly, then, we get the result that ARI_1 is *identical* to IBE: Both license an inference from E to C just in case C is entailed by the loveliest of the available competing hypotheses that would, if true, explain E.[14] In other words, IBE is a special limiting case of the more general inferential pattern ARI_k – a case obtained when k is set to an extreme value in order to maximize applicability at the expense of epistemic caution. Given this, it should hardly have been surprising that IBE has the systematic epistemic defect discussed in section 2. Indeed, note that IBE/ARI_1 is the *only* instance of ARI_k that licenses the dubious inference to the RNA world hypothesis since even just the two loveliest explanatory hypotheses in that case do not both entail that life began with the formation of RNA molecules.

4. Conclusion

We have seen that a traditional conception of abductive reasoning in terms of Inference to the Best Explanation (IBE) faces a heretofore unrecognized problem in cases where there are multiple plausible potential explanations for the available evidence. Inspired by the scientific methodology of Inferential Robustness Analysis (IRA), we approached this problem by constructing a family of abductive inference rules, Abductively Robust Inference$_k$ (ARI_k), which require the desired conclusion to follow from the evidence given each of the k loveliest explanatory hypotheses available. Interestingly, IBE here emerges as a limiting case ($k = 1$) in which one completely sacrifices epistemic caution for the sake of maximizing applicability; similarly, a rule that resembles IRA emerges at the other end of the spectrum ($k = n$), i.e. when one completely sacrifices applicability in order to minimize epistemic risk. This suggests that it would often be wise to steer clear of these two extremes ($1 < k < n$) in order to strike an appropriate balance between epistemic caution and applicability.[15]

[14] Note that since every proposition entails itself, this also covers standard cases of IBE in which one infers C from E in virtue of C being the loveliest available competing hypothesis that would, if true, explain E.

[15] I am grateful to Katrina Elliott, Kevin McCain and two anonymous referees for very helpful comments on drafts of this paper. This work was supported by the Irish Research Council [grant number REPRO/2015/89].

References

Barnes, E. 1995. Inference to the loveliest explanation. *Synthese* 103: 252–77.

Bernhardt, H. S. 2013. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others). *Biology Direct* 7: 23.

Climenhaga, N. Forthcoming. Inference to the best explanation made incoherent. *Journal of Philosophy*.

Douven, I. Forthcoming. Inference to the best explanation: What is it? And why should we care?", in *Best Explanations: New Essays on Inference to the Best Explanation*, ed. K. McCain and T. Poston. Oxford: Oxford University Press.

Harman, G. 1965. The inference to the best explanation. *The Philosophical Review* 74: 88–95.

Henderson, L. 2014. Bayesianism and inference to the best explanation. *British Journal for the Philosophy of Science* 65: 687–715.

Kuipers, T. A. F. 2000. *From Instrumentalism to Constructive Realism*. Dordrecht: Kluwer.

Lipton, P. 2004. *Inference to the Best Explanation*. Second edition. London and New York: Routledge.

Lloyd, E. 2015. Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science* 49: 58–68.

Lycan, W. G. 1988. *Judgment and Justification*. Cambridge: Cambridge University Press.

McGrew, T. 2003. Confirmation, heuristics, and explanatory reasoning. *British Journal for the Philosophy of Science* 54: 553–67.

Musgrave, A. 1988. The ultimate argument for scientific realism. In *Relativism and Realism in Science*, ed. R. Nola, 229–52. Dordrecht and Boston: Kluwer.

Okasha, S. 2000. Van Fraassen's critique of inference to the best explanation. *Studies in History and Philosophy of Science* 31: 691–710.

Thagard, P. 1978. The best explanation: Criteria for theory choice. *Journal of Philosophy* 75: 76–92.

van Fraassen, B. C. 1985. Empiricism in the philosophy of science. In *Images of Science: Essays on Realism and Empiricism, with a Reply from Bas C. van Fraassen,* ed. P. M. Churchland and C. A. Hooker, 245–308. Chicago: University of Chicago Press.

van Fraassen, B. C. 1989. *Laws and Symmetry*. Oxford: Clarendon Press.

Woodward, J. 2006. Some varieties of robustness. *Journal of Economic Methodology* 13: 219–40.