Experimental Philosophy, Noisy Intuitions, and Messy Inferences

Jonathan M. Weinberg

University of Arizona

to appear in *Advances in Experimental Philosophy & Philosophical Methodology*,

J. Nado, ed., Bloomsbury, 2016

(pre-print version)

Much discussion about experimental philosophy and philosophical methodology has been framed in terms of the reliability of intuitions, and even when it has not been about reliability per se, it has been focused on whether intuitions meet whatever conditions they need to meet to be trustworthy as evidence. But really that question cannot be answered independently from the questions, evidence *for what theories arrived at by what sorts of inferences*? I will contend here that not just philosophy's sources of evidence, but also its inferential resources, are in great need of closer examination.

I. The methodological inadequacy of "reliability"

Consider the fundamental methodological question of when one should or should not trust a source of evidence that delivers the output P: to what extent should we go ahead and accept P on the basis of a source's say-so? In trusting a source of evidence, we allow ourselves to accept P itself on the basis of its deliverance that P – always in a manner with the potential to be overturned by the vagaries of further research, and perhaps only with some barrage of qualifiers in here, such as *pro tempore*, or for the purposes of inquiry, or in accord with some conventional standard.

Now, one might have reasonably hoped that we could just import the notion of reliability wholesale from epistemology, into a methodological principle of when and how much to trust a source of evidence, namely, one ought trust a source about P to the extent that it is reliable about matters of P. A source that delivers a higher proportion of truths is one that can be more responsibly relied upon in inquiry than one that delivers a lower proportion. This would of course be subject to our best information about subdomains or circumstances in which the source is of greater or lesser reliability, but the core idea would still be: one can responsibly make use of a source only, and at most to the extent that, it is reliable. Something like this idea -- that reliability is the determinative consideration when deciding when to make use of a source -- is in the background in much recent discussion about methodology and intuitions. For example, it is at the heart of Boyd and Nagel's (2014) defense of the appeal to intuitions in epistemology, and plays the sort of central role one would expect in Goldman's (2010) discussion of philosophical methodology. And it seems like it would be nice, in at least two ways, if reliability really were the main determination of methodological trustworthiness. First, it can seem like a straightforward consideration to deploy (at least if we set aside theoretical concerns like those in the vicinity of the Generality Problem). Second, it seems highly plausible that intuitions on the whole are in fact reliable, and indeed to try to argue otherwise is to put oneself in danger of collapsing into a very broad skepticism, should intuitions turn out to undergird a very broad swath of our cognition, as many philosophers have claimed (e.g., BonJour (1998); Bealer (1998).)

But things turn out to be not so simple. It turns out that any degree of reliability noticeably less than perfect is consistent with a high degree of untrustworthiness, and any degree of reliability noticeably greater than nil, in which at least some detectable information about the target is present, can be consistent with trustworthiness. I am arguing here for the methodological inadequacy of the notion of reliability. My claim here is that if *all* we know about a source is its baseline reliability, that by itself cannot settle the question of whether and when to trust it. Low baseline reliability is consistent with trustworthiness, and high baseline reliability is consistent with untrustworthiness. Further sorts of information beyond baseline reliability are crucial in making determinations of trustworthiness, most particularly about the instrument itself, and about our general theoretical situation.

Of the two directions, it is easier to see how a low reliability source might nonetheless be trustworthy, in the right setting and given the right additional information. Suppose that we know that the reliability of a given device is 51%, say, in terms of telling whether a given sample is an acid or a base. (For ease of exposition I will just consider cases where there each datum is a simple binary, but the argument easily expands, *mutatis mutandis*, to more complicated sorts of outputs.) But allow that we know that the device's reports are both repeatable, and so, for example, it does not totally use up the sample in any trial. Moreover, the reports are independent even when applied to the same sample, and so, it will not necessarily just give the same answer every time, when re-applied to the same sample. Moreover, suppose we know it is not subject to any biases in its readings, in that its mistakes are a matter of chance and not, say, on average more likely to mistake an acid for a base than vice-versa. Under such assumptions, one can apply the device over and over again to the same sample, and while each individual report may have a 51% chance of being correct, we know the *average* response of the device to any given sample will be increasingly likely to be true, as the number of readings increases. This is a straightforward application of the Condorcet Jury Theorem: specify any degree of reliability you want to achieve, and you can determine easily how many trials you must perform on any given sample in order to be able to infer to its overall pH status to that degree.[1] But note that we had to know more than just the baseline reliability of the source here -- in particular, the independence of its deliverances is an essential part of the situation that allows this kind of aggregation to work. Any one reading may not be sufficiently trustworthy on its own, but we can build a trustworthy practice based on its readings, when further founded on our knowledge of various aspects of the device, most particularly what

[1] See Huemer (2008) and Talbot (2014) for applications of this principle in the context of our armchair intuitional practices.

kinds of errors it is or is not likely to make.[2]

So far, so good -- it should not be surprising that further information of the right sort can allow us to leverage better inferences out of a highly noisy instrument. But what about the other direction? How can a highly reliable source nonetheless not be trustworthy? One easy way to generate such a situation is when the source's output is wildly in conflict with other evidence we may already have, for example, if an apparently well-conducted experiment seems to indicate the reality of ESP, despite massive amounts of already existing evidence to the contrary. (See, e.g., Wagenmakers et al. (2011) response to Bem (2011).) But suppose that one has no specific evidence against some hypothesis H, and in fact H starts off at least as likely as every other rival to H. And then a highly reliable source provides stark evidence for H. Should we not then accept H, at least for the purposes of inquiry, to the extent of that source's reliability?

But it turns out again that features of the theoretical situation can still make it inappropriate to extend our trust so far. I will use a extreme case to make the point stark. Let us start with a device much more reliable than the one in the previous case, in the sense of having 95% specificity and sensitivity for both acids and bases: out of every 100 samples of acids, it will correctly give the reading that it is an acid 95 times, and likewise for bases. But it is also unlike that first device, in that its results are not independent, should we re-apply it to the same sample: if it first outputs that a sample is an acid, it will always say that, no matter how many times we check. Thus we cannot ramp up its reliability via repeated observations, like we could with the first device.

Now, suppose we have arrived at a point in inquiry where we have a set of 15 samples, and for each permutation of acid/base predictions for the members of that set, we have an equally probable hypothesis that makes exactly that prediction. Let us assume that we are certain that the truth must lie among those hypotheses, though, so their probabilities will sum to 1. So, one hypothesis predicts that they are all acids; another, that they are all bases; another, that the first four are acids and the latter nine are bases; and so on for all 2^15 permutations here.

Now, we perform the measurements, and our device reads that all 15 are acids. Let H be the hypothesis that makes exactly that prediction. So, its predictions appear to be *exactly* confirmed, and it is the only hypothesis whose predictions are so confirmed, since every other hypothesis will diverge from the readings for at least one sample, and many of them for more than that. It can thus seem that *of course* we should go ahead and endorse H, on such evidence from such a highly reliable source.

But a quick calculation reveals that, although we should correctly view H as decidedly more likely than any single one of its rivals, its probability on the evidence is still only (0.95)^15 – just a bit above 46%. In fact, if we consider the set of hypotheses that diverge from the observations by either exactly one or exactly two readings, it is somewhat *more* likely that that the correct hypothesis to be found somewhere in that set, than it is that H is the true one. We cannot trust the device here, with regard to H,

[2] This is what I call in my (2015) the "error profile" of the instrument.

certainly not anywhere near the level that the device's high reliability would have antecedently suggested to us.

That is not to say that the device has been epistemically worthless – far from it. Of what had originally been 2^15 hypotheses, we can consider these observations as having pretty much ruled out about 2^10 of them, and to have given us a sense of which of the remaining competitors are more likely or less so, with H having pride of place among the former. Beyond that, though, we must view our inquiry as still very open. The instrument is highly reliable, and while one and only one hypothesis predicts the observations that the instrument provides for us, nonetheless we find ourselves very far from arriving at an answer as to which hypothesis is the right one. Just how far depends on what sort of threshold of likelihood we consider acceptable. I had to use a dizziyingly large hypothesis space in order to get that ultimate probabiliy below 0.5, because I wanted to example to give us that that stark a mathematical result. Yet I expect that philosophers do not generally want to consider anything-above-0.5 as as a sufficient degree of confidence for accepting a theoretical result, however. If we set that threshold instead at 0.8 (which is probably still rather too permissive a standard), then merely *five* rival hypotheses would be enough to keep us below it, for this degree of reliability.[3]

Thus, no degree of poor reliability is sufficient to rule out trusting a source of evidence in inquiry, so long as it is at least modestly above chance. Nor is any high degree of reliability sufficient to establish trustworthiness, so long as it is at least in any practical sense less than utterly certain. One must attend to more than the general reliability of a source of evidence, when deciding whether and how to put it to methodological work. These cases draw our attention to important aspects of methodology that a purely reliabilist account of source selection cannot accommodate.

The key moral of these two cases is simply that when we are deciding whether or not to rely on a source of evidence for guiding our inquiries at some stage, it is just not enough to know how reliable it is (again, outside of knowing that it is perfectly reliable or perfectly unreliable or close enough to either for practical purposes). I take it that we seek reliability not just at the level of inputs to our processes of theory selection, but moreover – and more importantly -- at the level of the *outputs* of such processes. We are trying to arrive at a theory that, at a minimum, is highly likely to be true. But these cases teach, first, that further methodological resources can yield reliable theory selection from not-especially-reliable data; and second, that weaknesses in our inferential resources can make even a highly reliable source nonetheless inadequate to our theoretical demands. I will argue that philosophy is currently in something like the latter situation -- but can take a lesson from the former situation, about using further resources to supplement a methodological weakness.

Here is a very useful observation from Weatherson (2003), on the extent to which many areas of philosophy seem perhaps to rely too much on the power of a small number of data points in guiding theory selection:

[3] See also Gibson et al. (2013), both for greater mathematical rigor and for an application of similar reasoning in the context of methodology in linguistics.

> In epistemology, particularly in the theory of knowledge, and in parts of metaphysics, particularly in the theory of causation, it is almost universally assumed that intuition trumps theory. Shope's *The Analysis of Knowledge* contains literally dozens of cases where an interesting account of knowledge was jettisoned because it clashed with intuition about a particular case. In the literature on knowledge and lotteries it is not as widely assumed that intuitions about cases are inevitably correct, but this still seems to be the working hypothesis. (1)

One might reasonably object that this at least a little bit overstates the degree to which current philosophical practices in these areas are vulnerable to one-off counterexamples (and Weatherson's paper on the whole is sensitive to such concerns). For example, it is reasonably standard to attempt to explain a recalcitrant intuition away, or to say that one's theory is still the best overall fit, even if it misses a few cases (since in general, everyone's theory misses a few cases). But I fear that the standard moves here actually underscore how weak our resources are for handling the possibility of intuitive error. I am not sure that we have ever, in the history of the profession, had a case where there were tough cases for everyone, but someone said that *the other side* had the better overall trade-off. It is just too easy to impeach another's theory without being able to make the charges stick sufficiently to convince that one's own view must be adopted in its stead. We have at present no good resources for debating and evaluating philosophers' appeals to which theory has the best overall fit, and when we make such appeals, they are often with highly cherry-picked evidence sets. We don't have any general consensus as to what rules to use for scoring different failures -- why missing *this* intuition should count worse against a theory than missing *that* one. We give lip service to the strength or centrality of a case, but again he have no means to referee the frequent disagreements about whether a case is really really central, or how strong an intuition really really is.

We philosophers are thus in a situation like my toy example of a reliable-but-not-trustworthy source above. Across a wide range of domains, reliance on intuitions has perhaps narrowed down our set of live hypotheses to a fairly compact number. Even granting a high level of baseline reliability to intuitions, they do not have sufficient power of resolution to enable us to decide well between the resulting competitors, and we currently lack the further methodological resources to do better. Since the mere reliability of our intuitions has not proved adequate to our methodological needs, it is time to see whether -- and how -- we can do better.

II. Thinking about methodology

At a very high level of abstraction, we can think of inquiry as involving a space of theories within which to search for the one that is maximally likely to be true, and the primary -- but not sole -- determinant of our search is the evidence coming in from our sources. At this nosebleedingly-high altitude, we can identify at least four components to the norms governing evidence-driven methodologies.[4] None of this is

[4] I don't mean anything too particular by "evidence-driven" here, but I include the restriction to indicate that there may be very different sorts of philosophical methodologies

meant to be at all specific to philosophy, and indeed I will illustrate briefly with some parallels in linguistics, and my intention is that this abstracted picture speaks to many methodologies in the empirical sciences. Also, just to be clear, in identifying sets of norms that are operative in a methodology in some community is not thereby to endorse those norms as correct.

A. Source selection – What sources of evidence are to be taken as creditable, and with what sorts of initial weightings? For example, in classical syntax, certain sorts of linguistic judgments by native speakers are given a high degree of evidential weight, such as acceptability or unacceptability of particular sentences, and not so much on those speaker's judgments about why any given sentence strikes them as acceptable or not. In philosophy, of course, there is a parallel question regarding intuitions about cases, which then play out further in such debates as whether folk intuitions are to be trusted over those of trained philosophers, or the relevance of how esoteric or far-fetched a case might be, and so on.

B. Theory structure – What does the space of competitor theories look like? What forms of theories are to be valorized, and which discouraged? Theories of grammar for a natural language in classical syntax are computationally structured and discrete. One may contrast this with functionalist grammars, which are both highly attuned to frequency of usage and admit a great deal of gradation. Linguists disagreeing across this theoretical divide may still agree as to what the relevant data are to which their theories must answer. In philosophy, we can see debates as to whether and when our theories should take the form of necessary and sufficient conditions, or have modal force, or also just how complicated and baroque a theory of a philosophical domain can be without counting as somehow beyond the pale.

C. Modes of inference – What rules of inference, and criteria of theory/model-selection, are taken as canonical? How does one proceed from the deliverances of the sources in A. towards winnowing down the set of competitor theories in B.? For classical syntax, theories must cleave fairly close to the observed data of linguistic judgments, but there are specific resources for accommodating apparently errant cases, as the product of performance errors, computational resource limitations, and the like. Grammars must also be learnable by child speakers on positive data, and should avoid too much language-specific machinery. There seem to be fairly close parallels here in philosophy, with perhaps further work to be done by constraints about logical self-consistency.

D. Vindicating metatheory – What reasons does the community offer to itself, and to other communities of inquiry, as to why it makes sense for them to use the methodology as described by A., B., and C.? Continuing with our example, here one would appeal to such theoretical commitments as the competence/performance distinction, or the nativist apparatus for explaining language acquisition, or the modularity of syntactic processing. For philosophers, it can matter here a great deal whether we take ourselves to be providing an account of our own concepts, or of something robustly extra-mental.

for which this analysis may just not fit at all, e.g., Foucauldian genealogy, without meaning to be at all derogatory towards those methodologies.

With this rough and abstract framework of theory-selection-as-search in place, let me offer the following analysis of the methodology I will term the *classical philosophical appeal to intuitions* (CPAI): the norms for data, inference, and theory of classical analysis all fit each other very well. Our main data are verdicts about particular cases, as to whether such cases do or do not fall under particular categories, as typically provided by armchair consideration of the cases in a manner distinguishable from recalling those verdicts from previous investigation, or accepting them from another's testimony, or arriving at them by means of explicit inference from applying one's theory of the category in question. Whatever our psychological capacity may be that gives us access to such verdicts in such a way, many philosophers seem happy to call it "intuition" even while disagreeing starkly as to just what underwrites that capacity. (One can contrast George Bealer's (1998) rationalist account in terms of concept mastery; Jennifer Nagel's (2012) appeal to the notion of epistemological intuitions as implemented in our tacit domain-specific capacity for "folk psychology"; and Williamson's[5] (2008) approach in terms of our general capacity for judgment.) For the purposes of analyzing this practice here, that term will suit us fine.

Although we are certainly not limited to intuition as a source of evidence in philosophy, and it is generally legitimate to supplement with other sorts of evidence, it is still the most basic sort of evidence that our theories must be responsible to, in this methodology. And when we do advert to other evidence while using this methodology, we often do so in the service of evaluating some reported intuition as legitimate or not. For example, when we test an intuition for perhaps being driven by pragmatics, we are not so much looking to bring a second line of evidence to bear on the matter, so much as to place more scrutiny on our main line of intuitional evidence in the first place. We can thus see that this methodology is committed to a high-but-not-practically-infallible level of reliability for these intuitions.

I am thus taking it that certain sorts of anti-intuition metaphilosophers are simply wrong as a matter of description as to what philosophers do.[6] But the analysis *is* consistent with those who want to deny any central evidential role for the intuitions themselves, that is, facts of the form that philosophers might commonly report by saying something like, "I have an intuition that…".[7] The account here is consistent either with taking such psychological facts to constitute our evidence itself, or instead to be the source of the case evidence but is not itself the weight-bearing part of our evidence. (Such facts will still be among our evidence for these philosophers, since they endorse E=K, and these psychological facts about we do or don't report as intuited are among what we know.) On the former sort of anti-intuitionism, it doesn't really make sense that philosophers should bother with debating whether a given reported intuition does or does not come from the right kind of psychological mechanism. On the latter sort, however, it is highly analogous to scientists worrying about whether a given observation is or is not an artefact of their experimental

[5] Though one should note that he is an influential member of the minority that abjures the use of the term "intuitions", he is still defending the methodological practices that are under discussion here.

[6] E.g., Deutsch (2010, 2015); Cappelen (2012); Earlenbaugh and Molyneaux (2009).

[7] Such as Williamson, as noted above in n5. Ichikawa (2014) might be another example.

apparatus.

As for our theories, there is a strong norm for them to be logically sharp-edged – that is, to be *exception-intolerant*,[8] in a manner often secured via universal quantifiers and necessity operators. That *almost all* instances of knowledge are also instances of justified, true belief is taken to be a fairly worthless epistemological generalization, if it can only be defended in that hedged and partial form. There is also a norm preferring more compact theories to more expansive ones, in which the latter might get tagged negatively as, say, "baroque", and especially to run the risk of suffering accusations of being ad hoc. Each of these norms makes sense in the light of the other: point-data are putatively armchair-accessible, and can put severe pressure on logically sharp-edged theories. And by allowing only a fairly modest degree of complication, with strong norms against baroqueness or being ad hoc, and moreover with strong norms in place encouraging rigorous exploration of possible counterexamples, we can take it that a theory that survives such a process of counterexample testing should thereby legitimately have its credibility raised. In our terms here, our norms of what good theories can look like are serving in part to provide a theoretical context in which a small number of cases can properly be determinative for selecting the best theory.

Moreover, a vindicating metatheory in terms of something like our mastery of the relevant concepts is further backed both the sharp edges of the theories, and the plausibility that this methodology can legitimately be practiced from the armchair.[9] In turn it legitimates the use of off-beat and esoteric cases. While verdicts in such cases can go little distance if any in determining, say, what is typically true of some philosophical category, or true in normal conditions, they are sufficient to threaten the right sorts of claims pitched in a more demanding modality. And all of this dovetails nicely with our desire to be able to use our well-developed logical tools to extract further consequences, perhaps even surprising ones, once we have amassed a trove of such theoretical results.

I am trying to draw our attention to some of the ways in which the reliability of intuition is only one part of what is needed to make the method of cases make sense -- an integral part, surely, but not enough by itself without a number of other closely integrated pieces of methodological machinery. But when all the pieces of a methodology fit together so tightly, we should not be surprised, then, if a need to modify one piece of the methodological puzzle will require alterations in the others as well. My contention is that our growing awareness of just how much noise there is with our data, will require serious reconsideration of these other aspects of the method. In particular, we find ourselves with a mismatch between, on the one hand,

[8] Thanks to Wesley Buckwalter for this useful piece of terminology. Also, see Nado (2015) for further examination of the methodological consequences of the exacting demands made by our preferred forms of philosophical theories.

[9] Although there has been some important dissent from any sort of conceptualism, no other vindicating metatheory has emerged with anything near the popularity of that view. I believe this is in part because it has proved hard to justify these other elements of the practice without this further commitment.

just how subtle the differences are between our competitor theories, and just how formally stringent they are, and on the other hand, just how noisy our primary data source is here. There's probably not much we can do purely from the armchair to raise our baseline reliability much higher than it is now – though I would emphasize that that is in part because we have been able to discover at least some potential sources of error, and to modify our armchair practices accordingly. (To take just one example, it is customary to deploy bits of formal notation in philosophical writing to help avoid scope ambiguities, even outside the context of explicit logical proofs.) We don't have any reason to think that even more armchair reflection on intuition will root out very much more of the noise that remains. So how else can we try to address this revealed disconnect within the methodology of cases?

There is some promise within experimental philosophy of using experimental methods and a greater understanding of the psychology of intuitions to amplify our powers of error-detection here. This is the most methodologically conservative approach, in that all of the x-phi work would be aimed at rendering a cleaned up data set to which our traditional modes of inference could apply. The objective is still to provide sharp verdicts as to what is or isn't a case of knowledge, or free action, and so on, and the other aspects of the methodology as sketched above could remain pretty much as they are. Other than its abandonment of the armchair, the resulting methodology otherwise retains all the core features of CPAI. I pursue the prospects for such an approach elsewhere.[10] But here, let me consider what work might be done by making adjustments elsewhere in the methodological matrix.

III. Can we make room for messy theories?

Suppose we decide that we need to start acknowledging that our intuitive case-verdict data really is noisy, and seek to accommodate that situation in our theorizing. We would need to revise CPAI in way that no longer relied on a fairly small number of well-targeted counterexamples to put lethal pressure on rival theories. We could thus consider a broader set of structures for our theories -- in particular, ones that *exception-tolerant* -- and see what additional resources they allow us to bring to bear.

Before we do so, though, we should ask whether we have good reasons to think that philosophical truths must generally have an exception-intolerant form. Beyond being true, what other work do we want a philosophical theory to do for us? We want them to be "illuminating", or perhaps "perspicuous"? Surely, but then is there a reason only logically sharp-edged theories can do that? Of course soft-edged theories won't generate surprising theorems nearly as copiously as sharp-edged ones – a small set of propositions with just a few interlocking univerally-quantified pieces can produce a truly astounding amount of structure, as the Peano axioms nicely illustrate. Replace "All numbers have a successor" with "Most numbers have a successor, typically", and you can't really get too far in your number theory. But it is not clear why propensity to rich formal treatment should be a *requirement* on our theories. Theorems are wonderful, both in their own right and where they can legitimately be put to work, but most of human knowledge just does not take that form. And for that matter, even the formal sciences have their own devices of exception-tolerance to

[10] Weinberg (forthcoming-a).

deploy where they may be useful, such as the highly useful mathematical notion of conditions that hold “almost everywhere”.

Moreover, the same aspects of logically sharp-edge claims that makes them amenable to proofs, also seems to make them susceptible to paradox! Weaken the premises of most paradoxes to less than strict universality, and the paradox disappears. Moreover, we can still use formal tools where appropriate, including formal models; we will just recognize that such models, in virtue of their particular formal structure, may be making claims that are sharpened past the point of actual philosophical accuracy. This should not count against such models even a little bit, since we already often expect such models to be simplifications of and abstractions from their target philosophical truths, and there is nothing wrong with that.

One might think that the *modal force* of philosophical truths is an important feature that they have. But propositions that are themselves exception-tolerant can still hold with various sorts of modal force. Depending on one’s views about ceteris paribus laws, one might think that most natural laws that our sciences actually traffic in, are exception-tolerant. Exception-tolerant propositions can also support counterfactuals: if we were now to be set upon by tigers, then it is highly likely they would have four legs (and so, say, we would not expect to be able to outrun them).

Perhaps philosophers are especially interested in very compelling subset of modally strong claims: claims of identity. If our goal is to establish a claim that starts “knowledge is…”, where that “is” an “is” of identity, then we would indeed seem to fail that goal and fail it badly should the right-hand side pick out a class that was not at least extensionally equivalent to knowledge. But should we really want to be only, or even primarily, interested in theories that express identities? It’s been a long time since Socrates, after all, and philosophers as otherwise dissimilar as Jerry Fodor and Timothy Williamson have raised very good doubts as to whether many such philosophical truths will be forthcoming in any meaningful way. Even if we decide that such identities are worthy of philosophical pursuit, it does not follow that they are the only kind of theoretical result worth achieving in philosophy. For I wonder whether we are leaving important philosophical facts unused on the floor, by refusing to make heavier use of soft-edged claims. There are counterexamples to KK, sure, but is it not an important fact about knowledge that, say, KK is true of the vast majority of knowledge claims that we self-consciously deploy, such as in argumentation? Whether or not all justified true beliefs are knowledge, surely there is a mammoth overlap between JTB and K -- is that really not a result worth taking epistemological notice of? These questions are not rhetorical; I don't think it is obvious what the answer to them are. But what I want to draw our attention to here is that, in particular, it is not obvious that the answer to such questions must always be *no*.

One kind of such structure that already has some currency in current philosophical practice is that of 'family resemblances' or 'cluster concepts', where we do not look for more from an analysis than a rather motley set of disjuncts that may have little to no deeper unity. But this approach has not generally proved an attractive framework, outside some recent reemergence in philosophical aesthetics.[11] This perhaps makes

[11] E.g., Gaut (2005).

sense: in many parts of aesthetics, one expects the concepts to twist and wind around our contingent and all-too-human practices of art creation and appreciation. We do not necessarily expect a category like *art* or (even more) *horror film* to have too much depth beyond our historically-configured patterns of attribution. For categories like *knowledge* or *causation*, though, philosophers seem in general to expect more organic unity. There are of course exceptions to those general trends[12], and the issues under discussion here may be cause for re-think about the merits of such theories. But let us see what other options can be put upon our methodological table.

One candidate that should be considered here are generics. Even if we grant that K=JTB can't be true as an exceptionless universal, one might nonetheless be struck be the *weirdness* of Gettier cases on the whole, and wonder whether *knowledge is justified true belief* might still be a useful piece of epistemological lore when considered as a generic.

One very attractive feature of generics is that multiple inequivalent generics can all be true. *Tigers have four legs* and *tigers have stripes* are both true generics about tigers, for example. Similarly, *knowledge is justified true belief* and *knowledge is true belief caused by a reliable process* are rival claims when understood as necessary and sufficient condition analyses -- but they are both surely true if understood as generics.[13] Could we perhaps take the whole class of major rival theories of what knowledge is, and turn them instead into a compilation of valuable generic claims about knowledge? With this one pirouette, we would transform the state of our inquiry to one mainly characterizable in terms of our ignorance as to the One True Theory of knowledge, into a copious stockpile of epistemological discoveries. Maybe this would be an unacceptable shrinking of philosophical ambition. But maybe, instead, the feeling that philosophy's progress is painfully slow is an artifact of a slightly hyperbolic set of expectations as to what philosophy should deliver to us, and those expectations themselves a false hope produced by too narrow a focus on one form that our theories can take.

Now, philosophical generics would perhaps be impossible to test well with CPAI, since they are counterexample-tolerant. But the question does immediately arise: how well can we test them at all, even with a suitably modified methodology? Would we need to bring in a notion of proper function, or statistical normalcy, or what? One methodological upshot would be that esoteric sorts of cases would be seen as of little or no evidential value, akin to trying to look at the amputee ward of the zoo when evaluating that generic about quadruped tigers. All in all, it would not be a trivial methodological change, to aim at generics instead of strict generalizations; indeed, the main moral of this paper is that methodological changes in general ramify more broadly than we have heretofore considered.

Let me put one more theoretical option on the table, drawn more from empirical psychology than from philosophy of language: weighted feature sets, such as prototypes. One interesting fact about the growing set of x-phi studies about Gettier

[12] See, e.g., Williams (1996).

[13] See Lerner and Leslie (2013).

cases, is that for many of them, the Gettier cases come out generally on the "not knowledge" side of the scale, yet often not nearly so far as other sorts of cases, like falsehoods or lucky explicit guesses. Indeed the overall pattern can be a bit embarrassing for defenders of the classical analyses of knowledge, since at least sometimes the folk's epistemometer needle leans equally far towards "not knowledge" for some paradigm Gettier cases as it does for cases which are supposed to count as knowledge, such as skeptical pressure cases.[14] If we are theorizing knowledge in terms of a weighted feature set instead of strict generalizations, we could allow that *Gettierized* could be a feature that applies some negative impetus against counting as knowledge, while perhaps accommodating how that might be a notably weaker anti-knowledge factor than, say, falsehood of the belief. Such a treatment might also prove a valuable way to include in our theory of knowledge factors that have been proposed as partly determinative of knowledge, but which have proved elusive to detect experimentally, such as stakes effects.

But this line of thought brings us to the greatest danger of trying to theorize in terms of weighted feature sets, especially if we will allow in features that are not weighted especially heavily. Namely, we may find ourselves unable to determine what psychologically subtle factors to include as part of the theory of the category, and which should be excluded as mere noise. Perhaps the major lesson from "negative program" x-phi is that there are lots of funny little effects on offer out there, lots of ways in which our attributions of philosophically important concepts can be unexpectedly manipulated. We would need a principled way to include what we take to be legitimate factors in our theories (such as, perhaps, Gettierization and stakes), and still exclude oddball ones (like order or font choice). I have argued elsewhere for the possibility of using effect size for help in separating signal from noise in our intuitions, but on this specific proposal of using weighted feature sets, especially if we are looking to use fairly modestly weighted features, that likely won't do the trick.

IV. What kinds of inferences can select between messy theories?

Now, if *all* we do is broaden the set of competitor theories to include various sorts of exception-tolerant ones, then we will only have made matters worse, with more theories to choose between and without any relevant tools to help us with that choosing. We must consider alterations not just to the forms of our theories but to our modes of inference. In particular, we require modes of inference that can better deploy noisy data.

Perhaps our best model of making inferences with noisy data is modeling itself. If we adopt a view of philosophical inquiry as a kind of model-building, then inference should be a matter of model-selection, and we can look to use the tools of model-selection in our philosophizing. One might think we are already kinda sorta doing that, but then we're doing it kinda sorta not very well. What we lack are decent metrics of goodness of model here. We mostly use binary data, so all we have is whether the model hits or misses at any given point, and we don't really have any way of making any comparison between several different models that all have different sets of hits with some misses. Philosophical fans of the model-building model often

[14] Nagel et al. (2013).

suggest that we can and do appeal to broad, holistic criteria of a good model a la Quine.[15] I am friendly to that general idea, and am not looking to raise difficulties about it here in terms of whether, e.g., parsimony is properly understood as a truth-conducive property of a theory in philosophy.[16] Even setting such concerns aside, however, I want to present two problems with the idea that holistic theoretical virtues can do the necessary work of theory selection.

First, in the absence of any sort of further resources to discipline our evaluations of overall holistic quality, I am not sure that we have any right to expect our appeals to such evaluations to bring in as much psychological noise as those same noisy intuitions that the holistic judgments are supposed to help us overcome. Confirmation bias in particular will be a huge threat here, and it is a kind of bias that we know is not overcome well just by being smarter and thinking harder.[17] And just ask yourself: when was the last time you saw a participant in a philosophical debate take a look at the holistic character of the state of play, and plump for an *opponent's* theory as a result? Not that it *never* happens, but it is perhaps notable just how notable it is when someone does change their flag from one ism to another. We should expect also that various path-dependencies and order effects to play an outsized role.

I need to offer two caveats in understanding this objection. First, I am not saying that we can never make good use of holistic judgments of theory quality. But my concern is that the noise to signal ratio will likely be poor when we attempt to use such judgments to choose between a small set of elite contender theories. We are in a situation highly similar to my example of reliable-but-untrustworthy sources above. Second, this issue is not quite the same problem as that of the incommensurability of different weightings of the theoretical virtues, as it would arise even between rival theorists who agree about those weightings. However, the existence of those different weightings does exacerbate the problem. In general, the more slack we allow in the rope of cognition, the more tightly our minds will bind themselves to whatever theory they happen to prefer.

My second objection against the idea that holistic theoretical virtues can do the required work is less psychological and more methodological. I am concerned about a potentially misleading picture favored by that idea's proponents in which we imagine applying those virtues after the evidence has done all of its work already. But considerations of simplicity and parsimony actually need to be brought to bear both more rigorously, and much earlier in inquiry, than this picture allows. We are used to thinking of them as very "big picture" sorts of considerations, yet in fact when we are reasoning from noisy data, they must be used at a stage much closer to the initial analysis of the data itself, in addressing the problem of *overfitting.*

[15] See, e.g., Paul (2012), who is also very explicit in her defense of metaphysical theorizing as a kind of modeling; Nolan (forthcoming).

[16] My discussion here is thus on the whole *pace* Sober (2009). In general, we should be careful here to distinguish *parsimony* considerations from *simplicity* considerations, and here I am only intending to traffic in the latter.

[17] Stanovich and West (2007).

Philosophical theory-selection and empirical model-selection are highly similar problems: in both, we have a data stream in which we expect to find both signal and noise, and we are trying to figure out how best to exploit the former without inadvertently building the latter into our theories or models themselves. Under-utilizing the signal is one kind of danger - but clinging too close to the precise contours of our data stream is yet another. This is just the trade-off we have been discussing. But where there is quantitative data and models, quantitative tools can be brought to bear. Our lack of any substantive means for evaluating the fit of models in philosophy contrasts sharply with the range of tools available in the sciences.

For example, one standard measure for model selection is the Akaike information criterion. Suppose we have some range of models, which vary from each other in two key ways. First, they deploy different numbers of parameters. And second, they vary in their likelihood -- intuitively, how much the parameter choices agree with the observations, measured in terms of how probable the observations are, on the assumption that the model is correct. (Note that this is not the same as how probable the model is, on the assumption that the observations are correct.) The AIC which is defined as

$$2k - 2\ln(L)$$

where k is the number of parameters, and L is a particular measure of likelihood.[18] (Don't even begin to worry about the 2's or the natural log in there, for our purposes here!) Speaking very roughly, the idea is that when considering some range of competitor models, one ought to prefer the one that minimizes AIC, and there are ways of measuring just how much preference to give, for any set of competitor AIC values. In general, models are rewarded for closeness of fit to the data, since a closer fit means a greater maximum on the likelihood function, which in turn means a lower AIC (since the 2ln(L) term is subtracted). And models are penalized by the number of parameters they have, largely to hedge against the risk of overfitting the data.

There's nothing special or magical about AIC, and there are a variety of other tools out there. I am using it as an illustration of the general way such tools work: by having a measure of fittingness for competitor models, and a penalty for additional parameters, and a way of putting those together to yield an evaluation of the bang-for-the-buck of any new proposed way of capturing more observations by increasing the number of parameters. Here's the thing: in order to use any tool of this sort, like AIC, we need some way of generating a quantitative measure of likelihood for philosophical theories. This is not something that we have generally looked to do, certainly not outside of x-phi and not even that widely within x-phi.

Let's return to Gettier cases again.[19] I mentioned above the problem that the effect size of Gettierization is in some places of a comparable size to effects that need to be considered as potential noise effects. There are further reasons to think that the evidence *on the whole* is pointing towards at least some Gettier cases counting as

[18] Burnham and Anderson (2002) is a canonical reference here.

[19] For a more extensive treatment of this example, see my (forthcoming-b).

non-knowledge, but in a way that is rather messier than the received philosophical view of such cases would predict.[20]

For the sake of the illustration, let us simplify matters drastically and stipulate that we have two rival theories: JTB and JTBW, and in which both (i) agree on all of the non-Gettier cases, and (ii) which, as their names suggest, have a clear asymmetric relationship in terms of their parameters, in which JTBW has all three components that JTB has, plus one more, a warrant condition that steers the predicted attribution values in a negative direction for at least some Gettier-type cases. A question we need to be able to answer is: *how much* of a better fit must JTBW be, in order to license our preferring it over JTB? And I do not see how we even know how to begin answering that question in an intellectually respectable way, at this point. We measure these trade-offs in entirely hand-waving, it-seems-to-me kind of ways -- which is to say, we don't really *measure* them at all. Among the methodological innovations that we need in the face of the fact of intuitional noise, we need tools to address such questions. The holistic criteria cannot do this work for us if we can only appeal to them in an unregimented manner. Yet we may legitimately hope that more richly quantitative data would allow for just the sort of rigorous, quantitative measures that are called for here.

(In principle such quantitative data could be possible from the armchair, should we adopt norms of intuition reporting that standardized some sort of scale of strength of intuition or something like that, perhaps even on a model with the fairly simple */?/?? system used in linguistics. But I think we have reason now to think that there would be a lot of disagreement within given communities about a great many cases and just how strongly felt various intuitions are. We should also suspect that these will be susceptible to all sorts of sources of error, such as order effects and confirmation biases, in a way that cannot be managed from the armchair but which can be filtered out using the methods of the social sciences.)

This seems to me an important future priority for experimental philosophy. I think an early hope -- certainly an early hope I had harbored -- was that experimental investigations would, in good time, reveal individual cases to be problematic, or not. Further investigation into the problematic cases would thereupon lead to their being either settled or discarded, or on hopefully rare occasion, relativized. Had things worked out that way, then x-phi could have played an important role in cleaning up philosophy's data set, but then our more traditional modes of inference might have remained untouched. But while that first, problematizing part has clearly happened, by and large very few cases have been either settled by means of x-phi, or shown to be so noisy as to be beyond hope of further research. As I mentioned earlier, some version of the cleaning-up, separating-wheat-from-chaff project is surely still worth pursuing. At the same time, we should pursue other forms of theories and appropriately altered modes of inference and model selection, in case the noise proves (as it has thus far) tricky to factor out of our evidential base.

V. Conclusion

[20] See also Blouw et al. (forthcoming) for an attempt to impose some order on this messiness.

Experimental philosophy is helping teach us just how *messy* philosophical practice is. I don't think we made the mess -- I think we are revealing a mess that was already there, but which was, as it were, swept under the rug of logically sharp theories. These theories are often elegant, even beautiful, but they turn out to rest only awkwardly atop the lumpy, bumpy floor of our all-too-human philosophical intuitions. One response x-phi can, and should, have to this situation is to clean up that floor -- to tidy up and help take out the garbage that cannot be disposed of from the armchair. But I am urging here that we need also to pursue another line of response: learning how we can stand comfortably amid the mess, by means of a better understanding of it. Where we cannot learn how to get rid of the mess, we must learn how to find a secure footing within it. And to do so (and to leave the metaphor of floor coverings behind us) may require a more radical reimagining of philosophical practice, especially in its inferences, than even experimental philosophers have heretofore imagined.[21]

Works Cited

Bealer, G. (1998). Intuition and the autonomy of philosophy. In DePaul, M. & Ramsey, W., (Eds.), *Rethinking Intuition* (Lanham: Rowman & Littlefield): 201-239.

Bem, D. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of personality and social psychology*, *100*(3), 407.

Blouw, P., Buckwalter, W., & Turri, J. (forthcoming). Gettier cases: a taxonomy. To appear in R. Borges, C. de Almeida, & P. Klein (Eds.), Explaining knowledge: new essays on the Gettier problem. Oxford: Oxford University Press.

BonJour, L. (1998). *In Defense of Pure Reason*. Cambridge: Cambridge University Press.

Boyd, K., & Nagel, J. (2014). The reliability of epistemic intuitions. In Machery, E., & O'Neill, E. (Eds.). *Current Controversies in Experimental Philosophy*. Routledge. 109-27.

Burnham, K. & Anderson, D. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Cappelen, H. (2012). *Philosophy Without Intuitions*. Oxford: Oxford University Press.

[21] I am grateful to Jennifer Nado, Michael Johnson, and Ron Mallon for very useful comments on an earlier draft.

Deutsch, M. (2010). Intuitions, counter-examples, and experimental philosophy. *Review of Philosophy and Psychology*, *1*(3), 447-460.

-----. (2015). *The Myth of the Intuitive: Experimental Philosophy and Philosophical Method*. Cambridge, MA: MIT Press.

Earlenbaugh, J., & Molyneux, B. (2009). Intuitions are inclinations to believe. *Philosophical studies*, *145*(1), 89-109.

Gaut, B. (2005). The cluster account of art defended. *The British Journal of Aesthetics*, *45*(3), 273-288.

Gibson, E., Piantadosi, S., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, *28*(3), 229-240.

Goldman, A. (2010). Philosophical naturalism and intuitional methodology. *Proceedings and addresses of the American Philosophical Association*. American Philosophical Association, 115-150.

Huemer, M. (2008). Revisionary intuitionism. *Social Philosophy and Policy*, *25*(01), 368-392.

Ichikawa, J. (2014). Who needs intuitions? In Anthony Robert Booth & Darrell P. Rowbottom (Eds.), *Intuitions*. Oxford University Press. 232-256.

Lerner, A., & Leslie, S. (2013). Generics, Generalism, and Reflective Equilibrium: Implications for Moral Theorizing from the Study of Language. *Philosophical Perspectives, 27,* 366-403.

Nagel, J. (2012). Intuitions and experiments: A defense of the case method in epistemology. *Philosophy and Phenomenological Research*, *85*(3), 495-527.

-----, San Juan, V., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition*, *129*(3), 652-661.

Nado, J. (2015). “Intuition, Philosophical Theorising, and the Threat of Scepticism.” In E. Fischer and J.Collins, eds., Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method. Routledge.

Nolan, D. (forthcoming). The a priori armchair. To appear in *Australasian Journal of Philosophy*.

Paul, L. (2012). Metaphysics as modeling: The handmaiden's tale. *Philosophical Studies,* 160: 1-29.

Sober, E. (2009). Parsimony arguments in science and philosophy—A test case for naturalism. In *Proceedings and Addresses of the American Philosophical Association*. American Philosophical Association, 117-155.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225-247.

Talbot, B. (2014). Why so negative? Evidence aggregation and armchair philosophy. *Synthese*, *191*(16), 3865-3896.

Wagenmakers, E., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: the case of psi. *Journal of Personality and Social Psychology*, *100*(3), 426-432.

Weatherson, B. (2003). What good are counterexamples? *Philosophical Studies*, 115(1), 1-31.

Weinberg, J. (2015.) Humans as instruments: or, the inevitability of experimental philosophy. In Fischer, E., & Collins, J. (Eds.). (2015). *Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method*. Routledge. 171-187.

----- (forthcoming-a) Going positive by going negative: on keeping x-phi relevant and dangerous. To appear in Buckwalter, W. & Sytsma, J. (Eds.), *The Blackwell Companion to Experimental Philosophy*, Oxford: Wiley-Blackwell.

----- (forthcoming-b). Knowledge, noise, and curve-fitting: A methodological case for K=JTB? To appear in R. Borges, C. de Almeida, & P. Klein (Eds.), Explaining knowledge: new essays on the Gettier problem. Oxford: Oxford University Press.

Williams, M. (1996). *Unnatural Doubts.* Princeton, NJ: Princeton University Press.

Williamson, T. (2008). *The Philosophy of Philosophy*. Oxford: Blackwell.