

Explaining Away Intuitions

Jonathan Ichikawa

Arché Research Centre, University of St. Andrews

What is it to explain away an intuition? Philosophers regularly attempt to explain intuitions away, but it is often unclear what the success conditions for their project consist in. I attempt to articulate some of these conditions, taking philosophical case studies as guides, and arguing that many attempts to explain away intuitions underestimate the challenge the project of explaining away involves. I will conclude, therefore, that explaining away intuitions is a more difficult task than has sometimes been appreciated; I also suggest, however, that the importance of explaining away intuitions has often been exaggerated.

Keywords: intuitions

When a philosopher's theory has counterintuitive consequences, she often feels pressure to *explain away* the offending intuition. What is it to explain away an intuition? How can we tell whether an attempt to explain away an intuition has been successful? I'll start with some case studies, presenting and evaluating philosophers' attempts to explain away intuitions; this should help us to identify the central features of successful and unsuccessful attempts. Then I'll turn to the question, under what circumstances is a philosopher obliged to explain intuitions away? My answer is: fewer than is sometimes supposed. I'll conclude with some remarks about the value of explaining away intuitions.

1. Three Elements to Explaining Away

In his "How to defeat opposition to Moore", Ernest Sosa (1999) suggests that safety is a necessary condition for knowledge.¹

Safety S knows that p only if, S would believe that p only if p were the case

Corresponding author's address: Jonathan Ichikawa, Arché Philosophical Research Centre, The University of St Andrews, 17-19 College Street, St Andrews, Fife, KY16 9AL, United Kingdom. Email: ichikawa@gmail.com.

¹ In his more recent work, Sosa demures from Safety. See, e.g., (Sosa 2007a, 26).

And he argues that, the “undeniable intuitive attractiveness” (Sosa 1999, 141) of such a necessary condition notwithstanding, sensitivity is not necessary for knowledge:

Sensitivity S knows that p only if, were p not the case, S would not believe that p .

Sosa apparently feels some pressure towards explaining away the intuitive appeal of Sensitivity; he attempts to do so by suggesting that safety and sensitivity, being counterfactual contrapositives of one another, are not equivalent, but might easily be thought to be. When one embraces Sensitivity, one is reacting to the truth of Safety, and, confusing the one for the other, one announces that Sensitivity is true.²

The pattern Sosa suggests is structurally like that of someone who thinks that Sonic collected gold coins, because her memory of 1990-era video games is hazy, and she is confusing Sonic (who collected gold *rings*) with Mario, who did collect gold coins. Gold coins are not the same as gold rings, but it is easy to confuse one for the other. She is reacting to an implicitly-known truth, but that truth becomes somewhat garbled, and issues into a mistaken intuition. This is, of course, a perfectly respectable form of an explaining-away. Plausibly, this story is a correct explaining-away, for some people, of the judgment that Sonic collected gold coins.

The Sonic and Mario case includes, as a central element, a psychological claim about the etiology of a certain belief: the explaining-away relies essentially on the claim the judgment that Sonic collected gold coins derives in part from a conflation between gold coins and gold rings (or between Sonic and Mario)—the explaining-away predicts that the individual who makes the mistaken judgment is making it for this particular reason. If this psychological claim is false, then the explaining-away fails. As I suggested above, I expect that this explaining-away is likely to succeed, for at least some individuals. Some people do have Sonic and Mario muddled in their memories. (Most of those people spent less time playing video games in the early 1990s than did the present author.)

Sosa’s own attempt relies on a similar empirical psychological claim, namely that those who accept sensitivity as a necessary condition on knowledge (or perhaps: some significant class of those who accept it) do so because they are confusing sensitivity with its contrapositive, safety.

This psychological claim by itself, of course, does not establish that safety is a necessary condition on knowledge, or that sensitivity is not. Sosa’s argu-

² I am using the capitalized ‘Safety’ and ‘Sensitivity’ to refer to the principles stated above; ‘safety’ and ‘sensitivity’ refer to the counterfactual properties that, according to Safety and Sensitivity, are required for knowledge.

ment for these conclusions is independent of the psychological explaining-away. (The argument, broadly speaking, is that Sensitivity is subject to counterexample, and that Safety best explains various cases.) Sosa motivates and defends a particular theory—Safety is true, but Sensitivity is not—then recognizes that it faces certain counterintuitive consequences; he is therefore at this stage committed to the claim that some of our intuitions are incorrect. He gilds this pill with an attempt to explain those intuitions away. This explanation, like all explainings-away, relies centrally on a psychological claim: that safety and sensitivity, being contrapositives, are easily confused. For the attempt to be successful, at least these two conditions must be met: the psychological claims to which the explaining-away appeals must (a) be true, and (b) predict the offending intuition.

How does Sosa's explaining-away here fare? His psychological claim is something like this:

Confusion We're inclined, when we recognize that some counterfactual is necessary for some condition, to intuit also that the contrapositive of that counterfactual is necessary for that condition, even when it is not.

Sosa says little in particular defense of Confusion; a few examples of similar patterns of error would be helpful. Someone more ambitious might cite more explicit psychological data that directly established such patterns. But I am inclined to think that Sosa is on firm ground in suggesting that many people do not ordinarily distinguish counterfactual conditionals from their contrapositives. Indeed, many people might, after brief reflection, endorse the equivalence of contrapositive counterfactuals. It takes a bit of cleverness to come up with a counterexample to the equivalence: $(A \Box \rightarrow C) \equiv (\sim C \Box \rightarrow \sim A)$.

(Sosa's own counterexample is: if this faucet were to leak, it would not both leak and be tightly closed; but if it were to both leak and be tightly closed, it would leak (Sosa 1999, 150).)

So Sosa seems to be on reasonably firm ground in establishing (a), the truth of Confusion. And (b) is, in this case, straightforward; it is obvious that safety and sensitivity are contrapositives, so it is clear that Confusion predicts the offending intuition.

These are necessary conditions for explaining away, but they're not sufficient. For it is consistent with (a) and (b), for instance, that one "explain away" the intuition that p with the psychological claim that one's intuitions are sensitive to the truth, along with the philosophical claim that p is true. This, of course, is no explaining away. There is something to explaining away over and above mere explaining. A third condition, therefore, is necessary:

(c) the explanation must not rely on the truth of the target intuition.³

Stronger and weaker readings of this condition are available. At least, the explanation must not *entail* the truth of the target intuition. But candidate explainings-away could fail to be sufficiently independent from the truth of the target intuitions without entailing them; for instance, an alleged explaining-away that dramatically probabilified the target intuition would, for that reason, fail.

A closer attempt to characterize what is demanded by the third condition is this: the explaining-away must have it that the target intuition be insensitive to its truth, in that, were its content not true, the subject would have the intuition anyway. For example, if the reason you think that Sonic collected gold coins is that you're failing to distinguish, in memory, gold coins from gold rings, then your belief is not sensitive to whether he did in fact collect gold coins; you would think he did, even if he did not—even if he collected gold rings instead.⁴

To challenge Sosa's explaining-away with respect to meeting this condition (c) would be to admit that counterfactuals and their contrapositives are not often distinguished in unreflective judgments, but to argue that this is as it should be, because, apparent counterexamples notwithstanding, counterfactuals *do* contrapose. We've been given an explanation of the intuition, but one on which it is veridical. Someone taking this line ought to have something to say about the apparent counterexamples, like the faucet-leak case above; a counter-explaining-away would probably be in order.⁵ My present project is not definitively to judge Sosa's explaining away in one direction or the other; it is to use the case study in order to help articulate the rules of the game. What is necessary for a good explaining away, and how does one evaluate an attempt to explain away?

Plausibly, (c) will not quite result, in conjunction with (a) and (b), in a *sufficient* condition for explaining away. It may be, for instance, that all three conditions will be met in an appropriate case of abductive reasoning. For

³ Such phrases as "the truth of the intuition" refer to the truth of the *content* of an intuition; as I use the term, an intuition is true just in case it has a true content. (Similarly, a belief is true just in case its content is true.) Thanks to an anonymous referee for pressing this clarification.

⁴ Applying the rule to the philosophical example at hand is somewhat trickier. Whether sensitivity is necessary for knowledge is, plausibly, a necessary fact, and it is not clear how it is to be evaluated when negated in the antecedent of a counterfactual. "You think that sensitivity is necessary for knowledge because you're confusing sensitivity for safety, which is inequivalent; so you would think that sensitivity is necessary for knowledge whether or not it really is." This difficulty is common to knowledge of necessary truths generally. There is, I trust, an intuitive sense in which we can make sense of "counterpossibles" such as these. See, e.g., (Nolan 1997), (Kment 2006).

⁵ As, for example, in (Ichikawa forthcoming).

example, suppose that I observe green emeralds and inductively conclude that all emeralds are green. Perhaps, on pain of widespread skepticism, we should not be willing to conclude that we can explain away my conclusion by pointing out that I formed my belief inductively, and that I would think that all emeralds were green, even if some unobserved ones were not.⁶

Even if (a)–(c) do not comprise sufficient conditions for explaining away, I hope that I have established each to be necessary. In §2, I'll apply these conditions to three further case studies.

2. Case Studies

Consideration of some particular attempts to explain away intuitions will, I hope, both illustrate the application of the conditions articulated in §1, and motivate the thought that explaining away intuitions is more difficult than is often appreciated; many prominent attempts to explain away intuitions fail these necessary conditions in relatively obvious ways. Philosophers too often underappreciate the difficulty of positing psychological theories about the origins of particular intuitions that meet all three of (a)–(c); to do so, one must posit a judgment pattern that is specific enough to predict the target intuitions, but one general enough to be clearly insensitive. Still, it must be sufficiently constrained to be plausibly attributed to those with the target intuitions. Three examples will illustrate the tensions that these criteria generate.

2.1 Hawthorne

In his book *Knowledge and Lotteries*, John Hawthorne considers a puzzle for invariantist approaches to knowledge. The puzzle is this: I read in *The Times* that Manchester United won; I trust *The Times* on good grounds; in fact, it is both generally reliable and correct in this instance. So, on pain of general skepticism, we should say that I know that w —that Manchester United won—and also that I know that *The Times* said that w .

I also know that *The Guardian* includes results of football games, and, like *The Times*, is very reliable about it. (I'm inclined to trust both papers equally.) I have not read *The Guardian* today, but I know (perhaps on inductive grounds, perhaps through testimony) that it says something about yesterday's Manchester United game.

Since I know that *The Times* reported that w , and I know that w , and that *The Guardian* reported one way or the other about w , it might seem that I should be able knowledgeably to infer from these facts that either both *The Times* and *The Guardian* reported correctly that w , or else *The Guardian* mis-

⁶ Thanks to Cameron Buckner for helpful discussion here.

takenly reported that $\sim w$. But, intuitively, that disjunction does not comprise knowledge. Hawthorne's commitments—including closure, invariantism, and anti-skepticism—have the implication that in the case given above, I know the disjunction. But this is counterintuitive. Hawthorne faces this recalcitrant intuition:

Ignorance In the case considered, I do not know that either both *The Times* and *The Guardian* reported correctly that w , or else *The Guardian* mistakenly reported that $\sim w$.

Ignorance is intuitive; it is, perhaps, just the sort of intuition that might motivate contextualism about 'knows'. But Hawthorne, the invariantist, wants to reject Ignorance, explaining its intuitiveness away. His explanation proceeds as follows.

Members of a certain class of disjunctions, Hawthorne says, have a property that inclines us to intuit that they are unknown, even when they are known. Hawthorne calls such propositions 'junk disjunctive knowledge'. (He attributes the name to Roy Sorensen.) Roughly speaking, a junk knowledge disjunction is a disjunction that is known only by virtue of knowledge of one of the disjuncts; were you to learn that disjunct to be false, you would reject the disjunction. This contrasts with the usual case, in which when you know that A or B, you can infer that B when you acquire evidence that not-A. Hawthorne emphasizes, reasonably enough, that useful disjunctive knowledge is not junk:

[W]hen information is usefully encoded by a disjunction, one's knowledge of that disjunction is not grounded simply in knowledge of one of the disjuncts. If you tell me that you will go either to Paris or to Rome this summer, my knowledge of that disjunction is not grounded in my knowledge of one or the other of the alternatives. Correlatively, I am primed to do disjunctive syllogism were I to acquire the belief that one of the disjuncts is false. So if I later learn that you have decided not to go to Paris, I will conclude that you will be going to Rome. (Hawthorne 2004, 71–72)

So the intuition that I do not know the disjunction above is a case where we have the intuition that, for some piece of junk disjunctive knowledge, I do not know it. This, Hawthorne suggests, provides the resources to explain away the offending intuition. What he needs is for skeptical intuitions about junk disjunctive knowledge to be unreliable. He claims this result in the suggestion that we intuit junk disjunctive propositions to be unknown, even when they are in fact known.

When one balks at the idea that I know that either *The Times* and *The Guardian* correctly reported a Manchester United victory or else

The Guardian made a mistake, one imagines that I am so situated that, were I to read *The Guardian* and notice it said that Manchester United had lost (or drawn), I would be able to use my disjunctive knowledge to infer that it was *The Guardian* that had made a mistake. It would indeed be absurd to suppose that I am so situated, but it does not follow that I do not know the disjunction. We are apt to confuse useful disjunctive knowledge with junk disjunctive knowledge. Having recognized that disjunctive knowledge is not useful, we are prone to think that it is not knowledge at all. (Hawthorne 2004, 72–73)

The claim appears to be that generally, when we encounter disjunctions that are not, in the relevant sense, useful, we illicitly intuit that they are not known. This is the psychological claim against which we test my (a)–(c) above.

Were this psychological claim true, Hawthorne's story might be a plausible one. It predicts the intuitiveness of Ignorance, and does so in a way that builds in insensitivity. So (b) and (c) are met.⁷ What of (a)? Is Hawthorne's psychological claim true? Surprisingly, Hawthorne says little in particular defense of it. And it is not hard to see that it does not, in generality, stand. For some junk disjunctive knowledge, though obviously junk, is nevertheless obviously known. For instance, I know that Barack Obama was born in Hawaii; on this basis, I know that Obama was born in Hawaii or Rhode Island. Admittedly, claiming this knowledge generates a false implicature—namely that, were I to come to believe Obama was not born in Hawaii, I would believe him to have been born in Rhode Island—nevertheless, the knowledge claim does sound true, in a way that the target one did not. But each is equally junk. Hawthorne's attempt to explain away, then, seems to fail the (a) condition; the psychological premise on which the story relies is not true. A successful explaining-away must first be a successful explaining.

Perhaps Hawthorne, or someone sympathetic to his approach, will argue in response that I have not correctly formulated the psychological principle on which he is relying. It is not the mere fact of junk disjunctive knowledge that generates skeptical intuitions, but some further feature of the *Times* and *Guardian* case: junk knowledge that meets some additional criterion. Upon identifying this criterion, then, one can reformulate the psychological element of Hawthorne's explaining-away, avoiding my counterexample-based objection. In principle, this is a legitimate move. That it is an open question

⁷ One could try to deny (c), admitting Hawthorne's link between uselessness and intuitions of non-knowledge, but insisting that those intuitions are veridical. This would be in effect to make usefulness of the relevant sort a necessary condition for knowledge of a disjunction. This move strikes me as too extreme to be at all plausible, but perhaps someone could make a case for it. It is the structural parallel to the suggestion that safety and sensitivity are after all equivalent in the Sosa case.

whether a better psychological principle is available shows that it is an open question whether there is a possible explaining-away of the recalcitrant intuition in this case. Of course, the onus is on Hawthorne to articulate such.

But here is a worry for attempts to refine explainings-away along these lines. The more one refines one's psychological principle to avoid obvious counterexample, the more danger one runs of offering an explaining away that is drained of its rhetorical force. In his original case, Hawthorne cited the claim that junk disjunctive knowledge is generally intuited to be unknown. An important feature of the explaining away—the part that ensured the (c) condition—was that this human tendency, if actual, is a mistake. All parties agree that there is some junk disjunctive knowledge; if our intuitions consistently judge junk disjunctions as unknown, this does give us reason to discredit these kinds of intuitions. But if, in response to the observation that the general psychological claim is false, Hawthorne were to limit the psychological claim to one affecting only some limited subset of junk disjunctive knowledge, it is less obvious that a mistake is being made. The contextualist, for example, may well agree with a strengthened version of Hawthorne's psychological theory—he may say, for instance, that we're inclined to intuit knowledge attributions to be false when the content of the alleged piece of knowledge is a disjunction that includes a disjunct that raises certain skeptical possibilities to salience. But whether this tendency represents a mistake is exactly what is at issue between Hawthorne and the contextualist. On pain of begging the question, then, Hawthorne cannot cite a principle like this is his explaining-away of skeptical intuitions.

To defend Hawthorne's explaining away of this intuition in an effective way, one would have to posit a psychological claim that, unlike Hawthorne's original attempt, is plausibly true, *and also*, unlike the attempt just considered, uncontroversially represents an error. Perhaps this could be done, although I confess I cannot see how.

2.2 Stanley

A similar problem besets an attempt to explain away by Jason Stanley in *Knowledge and Practical Interests*. Stanley's approach to knowledge has a counterintuitive consequence with respect to this case:

High Attributor-Low Subject Stakes. Hannah and her wife Sarah are driving home on a Friday afternoon. They plan to stop at the bank on the way home to deposit their paychecks. Since they have an impending bill coming due, and very little in their account, it is very important that they deposit their paychecks by Saturday. Hannah calls up Bill on her cell phone, and asks Bill whether the bank will be open on Saturday. Bill replies by telling Hannah, 'Well, I was there two weeks ago on a Saturday, and it was open.' After reporting the discussion

to Sarah, Hannah concludes that, since banks do occasionally change their hours, ‘Bill does not really know that the bank will be open on Saturday.’ (Stanley 2005, 5)

Intuitively, Hannah’s last utterance is both sensible and true. Stanley, however, gives an approach to knowledge on which it is false. According to Stanley, whether S knows that p depends on whether p is important to S, but not on whether p is important to the attributor of knowledge or non-knowledge. So on Stanley’s view, Hannah is wrong when she says that Bill does not know: Bill does know, because the stakes are not high for Bill. That the stakes are high for Hannah is irrelevant with respect to Bill’s knowledge. And indeed, since Hannah knows that Bill is disinterested, she has no reason to think that he does not know. So Stanley’s view faces this recalcitrant intuition:

Correct In the case considered, Hannah is correct to deny knowledge of Bill.

His task, therefore, is to explain Correct away. Stanley writes:

Here is an intuitively plausible account of what is occurring in High Attributor-Low Subject Stakes. . . . When High Stakes wants to know whether another person knows that p , it is presumably because High Stakes has an important decision to make, one that hinges upon whether or not p What High Stakes is interested in finding out, then, is whether someone else’s information state is sufficient for High Stakes to know that p . In short, the *purpose* High Stakes has in asking someone else whether or not p is true lies in finding out whether, *if that person had the interests and concerns High Stakes does*, that person would know that p

We are now in a possession of a perfectly intuitive explanation of the intuitions in High Attributor-Low Subject Stakes. Hannah and Sarah are worried about their impending bill, and so they want to know whether the bank will be open on Saturday. It is to resolve this question that they phone Bill. What they want to know from Bill is whether he has evidence such that, *were he in their practical situation*, it would suffice as knowledge. . . . Of course, were Bill to share Hannah and Sarah’s practical situation, he *would* be in a High Stakes situation, and so would not know, on the basis of the evidence that he actually has, that the bank will be open on Saturday. So Hannah and Sarah are perfectly *correct* to conclude that the answer to their actual concern—whether Bill would know that the bank will be open if he were in Hannah and Sarah’s practical situation—is negative.

. . . So, we are strongly inclined to go along with Hannah and Sarah’s judgments, since we recognize that they are perfectly correct about the information in which they are really interested. (Stanley 2005, 101–103)

Stanley's strategy against Correct seems to comprise these claims:

- (1) When Hannah asks Bill whether he knows that the bank is open, *what she really wants to find out* is whether his evidence would be sufficient for her to know whether the bank is open.
- (2) Bill's evidence would not be sufficient for Hannah to know whether the bank is open.
- (3) In general, when someone asks whether X because she really wants to find out whether Y, and the answer tells her that not-Y, she will have the intuition that not-X.
- (4) In general, when someone has the intuition that not-X because she asked whether X, intending to find out whether Y, and learned that not-Y, we will have the intuition that her intuition that not-X is correct, even when X is true.

These four claims seem to be sufficient, if true, for a successful explaining of Correct away. The four claims together do seem to explain Hannah's reaction to Bill, and our intuition that it is correct, and to do so in a way that diminishes the probative force of that intuition. Let us grant (1) and (2), at least for the purpose of argument. Stanley's (3) and (4) are psychological claims; if they are true, then, combined with (1) and (2), they would explain Correct away. As before, the (b) and (c) conditions are met. But what of (a)? Are Stanley's psychological claims true?

It is surprising, I think, that Stanley does not go to more effort to defend (3) and (4). There is, in the text, a sort of "just so" story about the purpose of communication, meant to show how truths like (3) could arise. But there is nothing like an extended defense of the actual truth of either (3) or (4). And indeed, neither appears to be particularly plausible.

Suppose I am getting dressed and hoping to impress my friend Katherine who is a fashion maven. I ask you this question: 'Do I look hip and awesome?' (In fact, let us stipulate, I do look hip and awesome.) I ask this question because I want to know whether Katherine will be impressed. You divine my true purpose and do not even bother looking at me. 'Katherine is in a bad mood, and we not be impressed no matter what you wear.' Stanley's (3) has it that I should now have the intuition that I do not look hip and awesome; Stanley's (4) has it that third-party observers should intuit that I speak truly if I say, 'I do not look hip and awesome.' But neither of these predictions are met. I remain agnostic about the question, and it is clear to third-party observers that I do look hip and awesome. So I think that Stanley's explaining-away, like Hawthorne's, relies on too careless an empirical psychological claim.

Of course, as in the previous case, there is room for some considerable back-and-forth here. Perhaps (3) and (4) could be modified to avoid my objection; perhaps there is an alternative possible successful explaining-away of Correct. The challenges for Stanley will be similar to the ones I discussed for Hawthorne above.

Perhaps it will be objected that I am demanding too much of Hawthorne and Stanley's attempts to explain away; that, read in their broader context—which includes independent arguments for their respective views, and independent criticism of their contextualist rivals, who do capture the intuitive judgments—they have no great obligation to tell a thorough psychological story about the source of the mistaken intuitions. As will emerge below, there is much about this suggestion with which I agree. I think that the emphasis that some philosophers have placed on explaining away intuitions is excessive, and that sometimes, it is legitimate for philosophers simply and barely to accept counterintuitive consequences, without providing explanations about from where the faulty intuitions derive. But it must be admitted that to give up on explaining the origins of an intuition is to give up on explaining the intuition away. There is a difference between explaining away an intuition and biting a bullet. Insofar as Hawthorne and Stanley are attempting to explain away intuitions—which is what they both say they're doing—I have argued that they fail. I have said nothing against the alternate strategy of simply biting the relevant bullets; to evaluate that strategy is to evaluate the merits and drawbacks of the views on the whole, along with those of their competitors.

The shared moral of the Hawthorne and Stanley cases, I think, is that philosophers who wish to explain away intuitions must take care to explain them with plausible psychological principles. There is room for a considerable variation on how one goes about establishing such psychological plausibility; in many cases, consideration of a number of examples is sufficient. I think that was so in the Sosa case discussed above. My final case study, from Tamara Horowitz, represents a much more empirically grounded attempt to explain away intuitions.

2.3 Horowitz

Tamara Horowitz (1998) provides a good example of a more thorough, careful, and empirically-informed attempt to explain away intuitions than do the examples given by Hawthorne and Stanley above. Horowitz is considering a particular sort of moral intuition that is sometimes used to support the existence of a morally relevant distinction between killing and letting die—or,

more generally, between doing and allowing.⁸ (Horowitz's particular target in her paper is Warren Quinn, but his view, I think, represents a widespread approach.)

Consider a consequentialist who thinks that the moral value of an action is determined entirely by its results. In particular, that someone dies as a result of an action counts equally against that action, regardless of whether the person is killed, or whether the person is merely permitted to die. Here is a familiar anti-consequentialist argument:

Given the choice between saving one drowning person in one place, or saving five drowning people in another place, it would intuitively be a good thing to do to save the five people, letting the one person die. But, given the choice between letting five drowning people die versus killing one person (perhaps by driving over him, where he is trapped on the road) in order to save the five people, it would be intuitively bad to save the five people by killing the one person. Since both cases are alike with respect to how many people live and die, there are more morally significant factors at work—in particular, the fact that in the second case, you would be *killing* someone makes that decision worse than in the first case, where you would be merely *letting* someone die.

As in the cases above, we have a theory and a recalcitrant intuition. What is the consequentialist to do? He is committed to the falsehood of this intuitive claim:

Better In the case considered, it is better to let five people die than to kill one person.

How can he gild the pill? Horowitz provides an insightful suggestion. There is a general psychological tendency, she says, citing the influential

⁸ One caveat: Horowitz may be thinking of her own project in a way that does not fit exactly into the mold I have cast for explaining away intuitions. In particular, she focuses more on a philosophical project that seeks to explicate the norms that are encoded in human beings, than on facts about extra-mental ethical reality. "Some philosophers," she writes, not apparently intending to distance herself from them, "particularly ethicists and epistemologists, see as one of their tasks the discovery of norms, ethical or epistemological, that we more or less live by" (Horowitz 1998, 367). She declines ultimately to take a stand on the questions of objective morality, offering (Horowitz 1998, 381) official agnosticism about a morally significant difference between the relevant cases, but expressing—rightly, in my view—some skepticism about the alleged distinction, in light of her psychological proposal. However, it seems clear that an ethicist intending to engage in question of moral reality could put Horowitz's proposal to use in an attempt to respond to intuitions about killing and letting die. In what follows, I will speak loosely and attribute that project to Horowitz, with the understanding that her own official view may be somewhat weaker, limiting itself to psychological claims.

work of Kahneman and Tversky, to evaluate gains and losses asymmetrically. And what is counted as a “gain” or a “loss” depends on the relatively arbitrary setting of a “neutral” point. It is the same phenomenon that underwrites other sorts of *prima facie* surprising decision asymmetries. Here is an example (described by Horowitz 1998, 369–370):

Case 1: First, I give you \$300. Then, I offer you a choice: you can either (a) take another \$100 and walk away, or (b) flip a coin: if it’s heads, I’ll give you \$200 more; if it’s tails, you get nothing more.

In case 1, most people take the sure \$100, rather than gambling.

Case 2: First, I give you \$500. Then, I offer you a choice: you can either (a) give me back \$100 and walk away, or (b) flip a coin: if it’s tails, you have to give me back \$200; if it’s heads, you lose nothing.

In case 2, most people gamble for the chance to keep the full \$500. But of course, the two cases are exactly equivalent: both are a choice between gaining a sure \$400, or a 50/50 chance of gaining either \$300 or \$500. That subjects tend to respond asymmetrically shows that they are responding to features other than those that actually define their options. Kahneman and Tversky’s suggestion is that people respond asymmetrically to gains and losses, defined relative to whatever “neutral” point is salient to them. In the case above, the neutral point is set by the amount of cash that starts in their hands. Gains are treated with less significance than are losses.

Horowitz’s interesting suggestion is that the same psychological phenomenon underwrites *Better*. When all six people are already in mortal peril, the neutral point is set at six deaths; to save five of them is to gain five, which is straightforwardly better than gaining one. But if one’s neutral point has Roady living, and the five swimmers dying, then one weighs Roady as a loss, which is counted more significantly than are the swimmers, who would represent mere potential gains.

Insofar, then, as the asymmetry in intuitions relies on a quite general asymmetry between perceived gains and perceived losses, we have the resources to explain the appeal of *Better* away: We have the intuition that saving the five in case 1 is better than saving the five in case 2 because we weight the loss of the individual in case 1 as less important than the loss of the individual in case 2. In general, we consider potential gains as less important than potential losses, even when their true values are equal.

Here, as before, a psychological generalization predicts the offending judgment, and does so in a way that does not depend on its truth. So (b) and (c) are met. In this instance, unlike the previous case studies, I am inclined to accept (a) as well-established, too. I think this explaining-away succeeds. This is not to claim, of course, that there is not room to challenge it; one

might do so by presenting counterexamples to the generalized psychological claim, or by denying that Better is an example of the relevant general kind. Or, of course, one may accept this explaining-away while continuing to think that there are serious problems—even intuition-based problems—for consequentialism, perhaps comprising alternative intuitions that escape this explaining away. Whether an intuition against a view is explained away is nothing close to the be-all-and-end-all of the view's merits.

3. Must we Explain Away?

The case studies suggest that to explain away an intuition, one offers a psychological thesis to explain why people would have the offending intuition, even if it were not true. One's explaining away is successful insofar as the psychological thesis (a) is true, (b) predicts the offending intuition, and (c) does not depend upon the truth of the offending intuition. If this sounds tautological, perhaps it is, but it is worth a clear reminder. As the case studies indicate, too often, attempts to explain intuitions away fail in obvious ways.

As suggested in my discussion of Hawthorne and Stanley, however, to fail to explain away an intuition need not always be to lose the game. What if an explaining-away does fail? How bad is it to have a view with counterintuitive consequences, while unable to explain the offending intuitions away?

We can start by asking why we bother explaining intuitions away at all. Widespread practice notwithstanding, it is not *prima facie* obvious why philosophers should, in general, be concerned with explaining intuitions, or with explaining them away. Intuitions are psychological entities; philosophical theories are not, in general, psychological theories. Ontologists theorize about what there is; it is quite another matter, one might think, what people *think* there is. Epistemologists concern themselves with knowledge, not with folk intuitions about knowledge.⁹

Some philosophers, even some who recognize that philosophical ques-

⁹ This characterization of philosophical subject matter is, as the previous footnote demonstrates, not entirely uncontroversial. Much of the twentieth century was dominated by the “linguistic turn”, which had it that philosophical questions were ultimately questions about the meanings of natural language terms. A descendant of this tradition has it that philosophical subject matter is conceptual-explication of our concepts is all that philosophy can hope to do. These restrictionist views about the proper subject matter of philosophy are now widely, if not quite universally, thought to be mistaken. A philosopher may study the meaning of the word ‘knowledge’, or facts about the structure or cognitive significance of the concept knowledge, but he may also choose to study knowledge itself; if he does so, he may or may not find study of these psychological elements to be useful intermediary steps. See (Williamson 2007, chs. 1–2).

tions are often questions about extra-mental reality, think of intuitions as playing a special foundationalist role in the epistemology of philosophy. Such philosophers include both defenders of intuition-based philosophy and critics.¹⁰ Can such an approach explain the importance placed on explaining away intuitions? In other work, I have criticized the shared assumption that intuition does play this central evidential role in standard philosophical practice; however, even granting for the sake of argument such a special role to intuitions, it is not clear to me what onus there is on a philosophical theorist to explain away misleading intuitions. After all, we (almost) all agree that some intuitions are mistaken.¹¹ A philosopher with good reason to accept some theory, where that theory is inconsistent with some intuition, thereby has some good reason to describe that intuition as mistaken. If he does not know why we have the intuition—he cannot explain it, much less explain it away—why should this psychological ignorance be an obstacle to his philosophical theory?

By way of comparison, consider the attribution of false beliefs.¹² Beliefs, like intuitions, are sometimes mistaken, but, as in the case of intuition, most of our beliefs are true. Sometimes, theorists—be they physicists, phoneticians, philologists, or philosophers—argue for theories that are inconsistent with some of our beliefs. They do not, in general, face any requirement, or even any *prima facie* requirement, to explain why we have those false beliefs; they simply show us a reason to reject them, and ask us to do so. There is no obvious reason why the burden to explain away false intuitions should be any different.

Yet it is clear that at least some philosophers consider there to be such a burden. To take one example, Conee and Feldman (2004) argue against externalist approaches to epistemic justification by claiming that, unlike their preferred internalist approach, these approaches have no resources to explain away skeptical intuitions. Any view that cannot “make sense of” skepticism is thereby objectionable. In what, if anything, does such a burden consist?

¹⁰ Prominent advocates of philosophy based in intuition include Bealer and Strawson (1992) and Sosa (2007a,b). See (Stich 1990, forthcoming), (Hintikka 1999), and (Weinberg et al. 2001) for a few statements of the critical point of view.

¹¹ Ludwig (2007) defends an account according to which intuition is factive. On such an account, of course, inconsistency with an intuition entails that a theory is false. Nevertheless, one can still, on such an account, do something importantly like reject the truth of an intuition: one can reject the truth (and intuitionhood) of an *apparent* intuition. So likewise could one, if one cared to, attempt to explain away the apparent intuition.

¹² According to the account I favor, false beliefs are false intuitions. See (Ichikawa manuscript). See also (Williamson 2004). One need not accept this identity to accept the analogy here offered.

Certainly, a theorist gains *some* advantages by explaining away offending intuitions.

One reason it is a virtue of a theory that it come along with an explaining-away of recalcitrant intuitions is that in general, it is a theoretical virtue of a theory that it explain truths. If it is a fact that many people have the intuition that p , then a theory that predicts and explains this fact thereby gains some limited appeal. With respect to this consideration, there is nothing at all special about recalcitrant intuitions: it is also a theoretical virtue of a theory if it explains why many people have such-and-such *correct* intuitions, or if it explains why objects tend to fall downward, or if it predicts the recent economic downturn. Plainly, if this is the only reason it is valuable to explain away intuitions, it is a relatively edentulous one, and the widespread emphasis on explaining-away is misplaced. If I'm to theorize about, say, the nature of reference, I should not feel at all guilty if I fail to explain why people like chocolate, or why the Detroit Lions are so bad. Why should I feel differently about the fact that some people think that in Kripke's story, the name 'Gödel' refers to Schmidt? This psychological fact is interesting, and is, it seems to me, well worth explaining. But it is not clear why it should be the reference theorist's job to explain it. His job is to explain reference, not to explain intuitions about reference.¹³

4. Must Philosophers Explain Away Survey Data?

I disagree, therefore, with a certain alleged skeptical upshot of some survey-based experimental philosophy. I have in mind in particular data which purports to show that intuitions about central philosophical cases—Gettier cases are the favorite example—vary systematically according to cultural backgrounds. For example, Weinberg, Nichols, and Stich famously produced data suggesting that East Asians are less likely than Westerners are to have the standard skeptical intuition about Gettier cases. Although he means to resist metaphilosophical skepticism, some of Ernest Sosa's writ-

¹³ The philosopher who considers explaining away intuitions to be mandatory is a bit like the farmer in E. B. White's novel, *Charlotte's Web*. (See (Dreier forthcoming). My use of the example is inspired by Dreier's in the obvious ways.) Charlotte, a remarkable spider, sets out to save Wilbur, a rather ordinary pig, from the slaughterhouse. She does this by weaving intricate messages over Wilbur's sty, spelling out such phrases as *SOME PIG* and *HUMBLE*. The farmer is very impressed by Wilbur—what an amazing pig, to be such that a spider will write such messages above it!—but pooh-poohs his wife's suggestion to wonder whether they have an extraordinary spider. He errs in looking to the pig for the explanation of the odd responses the pig evoked, when they should have looked to the spider, who was the one responding. So it is, I think, when a philosopher considers it of paramount importance to explain, in terms of one's target phenomenon (*e.g.* knowledge), people's responses to it (*e.g.* intuitions about knowledge).

ings suggest that, if the disagreement is real, there is an important onus to explain away the offending intuition among the East Asians:

Obviously, half of them are getting it right, and half wrong. Of those who get it right, now, how plausible can it be that their beliefs constitute or derive from rational intuition, from an attraction to assent that manifests a real competence?

Not that it is logically incoherent to maintain exactly that. But how plausible can it be, absent some theory of error that will explain why so many are going wrong when we are getting it right? Unless we can cite something different in the conditions or in the constitution of the misled, doubt will surely cloud the claim to competence by those who ex hypothesi are getting it right. (Sosa 2007a, 102)

Sosa here suggests that, assuming the disagreement is actual, one's rational confidence in the standard Gettier judgment hangs on the providing of an explaining-away. (Sosa also provides reasons to dispute whether the apparent disagreement is actual.) Stephen Stich enthusiastically agrees with this suggestion:

It is worth emphasizing the enormous importance of this point, on which Sosa and I apparently agree. For 2500 years, philosophers have been relying on appeals to intuition. *But the plausibility of this entire tradition rests on an unsubstantiated, and until recently unacknowledged, empirical hypothesis*—the hypothesis that the philosophical intuitions of people in different cultural groups do not disagree. Those philosophers who rely on intuition are betting that the hypothesis is true. If they lose their bet, and if I am right that the prospects are very dim indeed for producing a convincing theory of error, which explains why a substantial part of the world's population has false intuitions about knowledge, justice, happiness and the like, then a great deal of what goes on in contemporary philosophy, and a great deal of what has gone on in the past, belongs in the rubbish bin. (Stich forthcoming, emphasis in original).¹⁴

It is helpful to consider some analogies. I know that the Earth is more than one million years old; I also know that not everyone knows that; some people think that the earth is only several thousand years old. Suppose someone did a survey and discovered that the distribution of people who believed the earth was more than one million years old correlated with certain demographic variables. Christians might be especially likely to disagree with me about the age of the Earth. Perhaps residents of Louisiana. For symmetry's

¹⁴ An exegetical note: I suspect that Stich may be exaggerating the extent to which Sosa agrees that failure to provide an explaining-away would be catastrophic; one may be troubled by doubt without chucking philosophical mainstays into the rubbish bin.

sake, let us suppose that our survey results indicated, somewhat surprisingly, that East Asians are more likely to think the Earth less than one million years old than are Westerners. Here, now, from my armchair, I do not know what explanation I could attempt to offer for such surprising data, should I be offered it. I would be at a loss to provide a substantive and plausible theory of error to explain the wrongness of so many East Asian judgments. (Relatively trivial and unilluminating theories might be available: “East Asians, unlike Westerners, tend to be disposed to judge that the Earth is less than one million years old; since the Earth is more than one million years old, East Asians are likely to have unreliable intuitions about the age of the Earth.” This theory, even if true and known, presumably does not satisfy the call for a “theory of error”.)

In this case at least, I think that the obvious thing to say is that we’ve discovered that members of a certain demographic group are not reliable judges about the age of the Earth; the proper response in this case is surely not skepticism.

(Consider also the Sonic and Mario case discussed above. In this instance, the view that Sonic did not collect gold coins is well-supported, for those of us who remember the game well enough, or for those of us who have looked it up recently—even if many people falsely believe that Sonic collected gold coins. Possession of the particular error theory is an additional epistemic good; but lacking it does not in any way impugn the theory that Sonic did not collect gold coins. If I know that many people believe that he did, then I may just know that many people are wrong about Sonic, without being able to articulate from what source their error derives.)

No doubt, it will be thought by many that this case is importantly different from the philosophical case. My knowledge about the age of the Earth comes from a rich education and contact with expert scientists whom I have independent reason to trust on these matters. True enough—but is the case in philosophy so different? My philosophical education is rich too—in my own case, much richer than my scientific one!—and I’m in contact with experts about knowledge. Indeed, my Ph.D. thesis was in epistemology, and I hope that I can, without hubris, claim a certain degree of expertise in this matter myself, at least relative to the folk.

So what relevant difference between the geological case and the philosophical case could obtain? Here is an obvious difference: my Gettier judgment is plausibly *a priori*; my judgment about the age of the Earth is not. If someone believes that the Earth is less than one million years old, it is because he lacks important evidence; the person who thinks Jones knows has all the relevant evidence. So goes one argument.

This disanalogy cannot stand up as presented. As philosophers well know, not all *a priori* investigations are easy; there is no guarantee that just anyone will get these questions right. (Compare the analogous situation when the folk disagree with a mathematician about an *a priori* mathematical fact.) To recognize Gettier cases as cases of non-knowledge is a cognitive achievement; it is entirely possible that, without philosophical training, some people might fail to achieve it. Just as people who have studied chemistry and physics are more likely to make correct judgments about the constitution of tables and chairs, so too are people who have studied epistemology more likely to make correct judgments about knowledge. This is just as it should be.¹⁵

Another example in the domain of the *a priori* can, I think, make this even clearer. People regularly and systematically err in probabilistic reasoning. In a famous set of experiments by Kahneman and Tversky, subjects were quite prone to judge some conjunctions (Linda is a bank teller and a feminist) as more likely than one of their conjuncts (Linda is a bank teller). Suppose for the sake of symmetry that this tendency was found to be much more widespread in East Asians than in Westerners. Would the axioms of probability theory have to be thrown into the rubbish bin, unless someone could come up with a suitable theory of error? Surely not: we have every reason to be more confident that the axioms are true than that most East Asians should be good at reasoning probabilistically.^{16,17}

Sometimes, the appropriate response to disagreement—including widespread disagreement within a particular demographic group—is to conclude on the basis of the disagreement that one's interlocutors are unreliable. This even in cases where no obvious (and non-trivial) suggestion for a theory of

¹⁵ Cf. Williamson's (2007) suggestion that describing a judgment as *a priori* does little by itself to explain how it is known. Thanks to an anonymous referee for pointing out the relevance of this point.

¹⁶ An interesting distraction in this case is that a plausible theory of error is available: the heuristics humans use to estimate probability track imaginability; sometimes, conjunctions are easier to imagine than one of their conjuncts (these are the times when the second conjunct gives you a clue as to how to imagine the first)—when subjects find something more easily imaginable, they judge it more likely. This theory also explains why English speakers are likely to think there are more words that begin with 'k' than there are words with 'k' as the third letter, and why so many people think they're more likely to die on an airplane than in a car. See (Kahneman and Tversky 1973). But of course, this explanation would be only partial; it would explain why people err as they do, but not why, as stipulated, there is cultural regularity with respect to the prominence of the mistake.

¹⁷ Incredibly, Stich (1990) dissents from this truism, suggesting instead that we consider it an open question whether the probability axioms or the surprising intuition about Linda is correct. This, I think, goes much too far; it is a gadarene open-mindedness, and the terminus of its short course is an austere skepticism.

error is forthcoming. This should be obvious about the age of the Earth case; my suggestion is that there is no great reason to treat Gettier cases differently.

One need not, then, if one fails to explain an intuition away, throw everything into the rubbish bin. Humans are funny creatures, and sometimes do and think funny things. That someone, or some demographic group, disagrees with an implication of my theory does not automatically mandate that I explain away their disagreement, on pain of irresponsible dogmatism.

5. Value for Explaining Away

But I do not mean to suggest that there is no value for philosophers in engaging in these explainings-away. I suggested above that there is at least some value in capturing data in general, including the case where the datum is a fact about someone's intuitions. But there are reasons to explain away intuitions beyond that rather anemic one.

The cases I have been focusing on so far are cases where a subject has conclusive reason to accept some conclusion, and is then faced with someone else who finds that conclusion counterintuitive. As I have emphasized, this is sometimes the case—perhaps more often than some experimentalists with skeptical proclivities think. But there are other important cases to consider, including some in which explaining away intuitions can play a more prominent role. Sometimes, for instance, a philosopher may be deliberating about a particular view, without being at all sure what to think. I find in myself conflicting intuitions, and do not know which to endorse. If I can see that one of those intuitions is a member of a class that I'm likely to find appealing even if false, this might provide me with some reason to prefer the other. The Horowitz case provides a nice example: if I am in internal tension between (a) the thought that it is better to do that which results in more lives being saved, and (b) the thought that it is wrong to kill somebody in a way over and above the way it is wrong to let somebody die, I may, if I'm convinced by her explaining-away, discount (b) as the product of a general error in rationality.

The phenomenon is, as I've been emphasizing, not limited to philosophy. When confronted with Müller-Lyer lines, I am inclined to judge one longer than the other, but I know enough about the psychology of perception to discount that inclination. Similar reasoning is at play in the management of documented biases. If I am hiring, and a candidate appears well-qualified, but I also have a bad feeling about him, I may, even if I generally place probative value on such intuitive feelings about applicants, discount it if the candidate is black, and I know (either on general inductive grounds, or specifically) that I'm likely to be biased against black candidates, discount that feeling. Explaining away intuitions, then, can be a very helpful thing to

do when deliberating.

Similarly and relatedly, explaining away intuitions can play a powerful dialectical role. If you and I are arguing, you can gain considerable traction with me if you can convince me that my position is the sort of position that I would be likely to have, even if it were not true. Suppose I'm in love with a woman, and you do not think she is trustworthy. I'm arguing with you: I know her best; you do not know what she's been through; I'm a good judge of character and can look into her eyes and tell that she is trustworthy. Perhaps the first-order evidence you can point to against her character is totally unconvincing for me. You may, however, make inroads if you stop telling me things about her, and start telling me things about me: you can remind me that I tend to fall in love without regard to a woman's trustworthiness, and also that, when I love someone, I tend to think she is trustworthy even when she is not. Of course, you might still fail to convince me—I can be very stubborn in love—but you do at least have a chance, if you can establish these psychological facts about me. You explain away the evidence I thought I had. So explaining away can also have a significant dialectical force.

To conclude: two themes have emerged in this discussion of explaining away intuitions. The first, illustrated by consideration of the case studies, is that explaining away intuitions is more difficult than is often thought; one relies, in an attempt to explain an intuition away, on particular psychological generalizations, and these generalizations are answerable to, and supportable only by, the psychological facts. This is not necessarily to say that one must be engaging in rigorous psychological methodology in order responsibly to attempt to explain away—sometimes, informal consideration of intuitive responses to cases is plausibly sufficient. But philosophers proposing psychological theories to explain away intuitions ought to reflect, at least minimally, on whether the generalizations upon which they rely are plausible; too often, as in the cases proposed by Hawthorne and Stanley, they are not.

The second theme ameliorates, to some extent, the first. Although it is more difficult than has sometimes been recognized to explain away intuitions, it is also less important. There are counterintuitive truths; finding one need not be cause for embarrassment. In cases where counterintuitive consequences of otherwise appealing theories are discovered, my advice to philosophers is to be upfront. Gild the pill with an explaining-away if you have a plausible one to offer; if you do not, then admit that you have a counterintuitive consequence to swallow, and explain why it is worth it to do so. Weak attempts to explain away recalcitrant intuitions only further muddy the issue.

Acknowledgments

Thanks for helpful comments to Derek Ball, Jessica Brown, Cameron Buckner, Herman Cappelen, Yuri Cath, Torfinn Huvenes, Jennifer Nagel, and Jonathan Weinberg. I presented drafts of this paper to the Indiana University Experimental Epistemology Laboratory seminar and to Arché's Intuitions and Methodology seminar; I am grateful to the participants there for helpful discussion. Thanks also to those who participated in several helpful comment exchanges on the Arché Philosophical Methodology weblog, and to an anonymous referee for this journal.

Bibliography

- Bealer, G. and Strawson, P. (1992). The incoherence of empiricism, *Proceedings of the Aristotelian Society* **66 (Supplement)**: 99–138.
- Conee, E. and Feldman, R. (2004). Making sense of skepticism, *Evidentialism*, Oxford University Press, New York, pp. 227–321.
- Dreier, J. (forthcoming). Queer pigs and the web of belief, in R. Joyce and S. Kirchin (eds), *A World Without Values: Essays on John Mackie's Moral Error Theory*, Springer Press.
- Hawthorne, J. (2004). *Knowledge and Lotteries*, Oxford University Press, New York.
- Hintikka, J. (1999). The emperor's new intuitions, *The Journal of Philosophy* **96**: 127–147.
- Horowitz, T. (1998). Philosophical intuitions and psychological theory, *Ethics* **108**: 367–385.
- Ichikawa, J. (forthcoming). Quantifiers, knowledge, and counterfactuals, *Philosophy and Phenomenological Research*.
URL: <http://jonathanichikawa.net/papers/qkc.pdf>
- Ichikawa, J. (manuscript). *Intuitions and Begging the Question*.
URL: <http://jonathanichikawa.net/papers/ibq.pdf>
- Kahneman, D. and Tversky, A. (1973). Availability: A heuristic for judging frequency and probability, *Cognitive Psychology* **4**: 207–232.
- Kment, B. (2006). Counterfactuals and the analysis of necessity, *Philosophical Perspectives* **20**: 237–302.
- Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches, *Midwest Studies in Philosophy* **31**: 128–159.

- Nolan, D. (1997). Impossible worlds: A modest approach, *Notre Dame Journal for Formal Logic* **38**: 535–572.
- Sosa, E. (1999). How to defeat opposition to Moore, *Philosophical Perspectives* **13**: 141–153.
- Sosa, E. (2007a). Experimental philosophy and philosophical intuition, *Philosophical Studies* **132**: 99–107.
- Sosa, E. (2007b). *A Virtue Epistemology*, Vol. 1, Oxford University Press, New York.
- Stanley, J. (2005). *Knowledge and practical interests*, Oxford University Press, New York.
- Stich, S. (1990). *The Fragmentation of reason*, The MIT Press, Cambridge, MA.
- Stich, S. (forthcoming). Reply to Sosa, in D. Murphy (ed.), *Stich and his critics*, Blackwell, Malden, MA.
- Weinberg, J., Nichols, S. and Stich, S. (2001). Normativity and epistemic intuitions, *Philosophical Topics* **29**: 429–460.
- Williamson, T. (2004). Philosophical ‘intuitions’ and scepticism about judgement, *Dialectica* **58**: 109–153.
- Williamson, T. (2007). *The Philosophy of Philosophy*, Blackwell, Malden, MA.