The Rise and Fall of Experimental Philosophy

Revised draft

Please do not quote without permission before publication, as some changes may be forthcoming.

Antti Kauppinen

July 7, 2006

Abstract

In disputes about conceptual analysis, each side typically appeals to pre-theoretical 'intuitions' about particular cases. Recently, many naturalistically oriented philosophers have suggested that these appeals should be understood as empirical hypotheses about what people would say when presented with descriptions of situations, and have consequently conducted surveys on non-specialists. I argue that this philosophical research programme, a key branch of what is known as 'experimental philosophy', rests on mistaken assumptions about the relation between people's concepts and their linguistic behaviour. The conceptual claims that philosophers make imply predictions about the folk's responses only under certain demanding, counterfactual conditions. Because of the nature of these conditions, the claims cannot be tested with methods of positivist social science. We are, however, entitled to appeal to intuitions about folk concepts in virtue of possessing implicit normative knowledge acquired through reflective participation in everyday linguistic practices.

1. Conceptual Analysis and Intuitions

Conceptual analysis has made a sort of comeback in recent years. For a while, pressure from Quinean attacks on analyticity and Kripkean arguments for a posteriori metaphysical truths led many to keep a low profile about the aprioristic character of their claims, but the tide seems to be turning. Frank Jackson and his fellow Australians have made a strong case for the claim that you cannot do serious metaphysics unless you get clear on exactly what our ordinary talk of mental properties, for example, commits us to (Jackson, 1998, Jackson, Pettit, and Smith, 2004). In the same spirit, contextualists in epistemology have refocused their attention on the ordinary use of 'knows' – to the extent that Keith DeRose, for instance, now identifies himself as a practitioner of the once-despised ordinary language philosophy (DeRose, 2005). In general, there is a growing recognition that the first step in resolving many philosophical problems is still laying out just what we talk about when we use the concepts we do, and this is just the business of conceptual analysis.[1]

Nonetheless, one might be sceptical about the prospects of success in such an endeavour. Whether it is the concept of free will or the nature of moral judgment, competing accounts have been slugging it out not just for decades, but for centuries, even millennia. If one philosopher says, for example, that it is a conceptual truth that an agent who sincerely makes a moral judgment is necessarily motivated accordingly, and another denies this, how is the argument to be settled? What counts as evidence one way or the other? The usual answer is to talk of

[1] There are, of course, some philosophers, like Jerry Fodor (1998), who believe that most or all of our (lexical) concepts are atomic and nothing like analysis is possible or necessary. These philosophers provide deflationary accounts of seemingly conceptual truths to explain these appearances away (see Margolis and Laurence 2003). I will simply ignore these views in the following, since the existence of conceptual truths is accepted on both sides of the debate I am addressing. I am afraid those who deny their existence will find little of interest here.

intuitions: one account matches our intuitions or 'ordinary usage' better than the other.[2] Conceptual intuitions – or 'Socratic intuitions', as they are sometimes called (Margolis and Lawrence, 2003) – are, roughly speaking, pre-theoretical dispositions to apply concepts to some particular cases or scenarios and refuse to apply them to others. There are countless examples of philosophers making claims to the effect that 'Intuitively, we…' or 'We would say…' or 'Ordinarily, we would not describe X as…' or 'It is a platitude that…', and so on. Often, intuitions are appealed to as counterexamples to a proposed analysis: 'Davidson's view would have the implication that X φ-s intentionally in S, but we would not in such and such a case of type S say that X φ-ed intentionally, so we must reject the analysis…'. Sometimes such claims are made in the language of possibility and necessity: 'Intuitively, it is not possible for water to be XYZ'. It is a matter of major metaphilosophical controversy whether these claims express modal intuitions that are distinct from conceptual intuitions; however that may be, I will only discuss concepts here.[3]

A remarkable feature of these claims is an appeal to a 'we'. It is rare to appeal to one's own judgment, and if one does so, the implication in context is that it is not only I who would judge this way, but other speakers would do so as well. For the appeal to intuition to serve its purpose, these others, the extension of 'we', must be those who are not partisans of this or that philosophical position. After all, the intuitions in question are supposed to serve as neutral data against which the competing analyses are assessed. So, it seems that the evidence that settles

[2] A related formulation is that the favoured account explains the *platitudes* involving the concept, the platitudes being those statements whose acceptance is necessary for competence with the concept. This is the 'Canberra' version of conceptual analysis (see Jackson, Pettit, and Smith, 2004) . I'll focus here on the intuition talk, though I believe most of what I'll say could be formulated in the language of platitudes as well.

[3] For a defense of the view that modal intuitions are cases of linguistically unmediated knowledge of metaphysical necessity and possibility, see Williamson, 2005. For a view that links metaphysical possibility to conceivability in terms of two-dimensional intensions, see Chalmers, 2002. My sympathies lie with linguistic approaches to metaphysical possibility (and I think Kripke and Putnam are ultimately in this camp as well), but for the purposes of this paper, it does not matter which meta-metaphysical view one adopts, as long as it is acknowledged that *some* questions about conceptual possibility are legitimate philosophical questions.

philosophical disputes about ordinary concepts is ultimately the particular judgments of non-philosophers. Sometimes this is made explicit; John Hawthorne, for example, frequently appeals to what 'people are inclined to say' about particular cases of knowledge (e.g. Hawthorne, 2004, 71). It is our shared, ordinary concepts that we talk about when we do conceptual analysis. Moral responsibility, for example, is not a technical notion, though some terms that philosophers use in explicating it may be. Indeed, why should anybody care about what philosophers do if they just argued about their own inventions? People want to know if they have moral responsibility or knowledge of other minds in the very sense in which they ordinarily talk about responsibility or knowledge, and to get at that sense one must work with the folk's own concepts. By and large, philosophers oblige; revisionism is a last resort, to be used only when one is convinced that the folk concept is hopelessly confused or too imprecise for one's purposes. To be sure, we are sometimes willing to discard individual intuitions in favour of theoretical unity to achieve a reflective equilibrium – perhaps, in the light of general considerations, we should after all agree not to call an agent in certain circumstances morally responsible, for example. But without a good understanding of folk concepts the whole process of reaching reflective equilibrium would and could not get going in the first place.

How, then, do we get at the intuitions that serve as evidence for the content of those shared, common concepts? The traditional view was that a priori reflection by a philosopher would suffice – conceptual analysis and aprioricity went hand in hand, and rejecting one meant rejecting both. It is this connection that is challenged by a new school of thought about philosophical methodology, sometimes called 'experimental philosophy'[4]. Experimentalists interpret claims about intuitions as straightforwardly empirical and therefore testable predictions

[4] See e.g. the eponymous introductory article by Joshua Knobe (forthcoming a). 'Experimental Philosophy' is also the name of the blog devoted to these and related issues (http://experimentalphilosophy.typepad.com/, coordinated by Thomas Nadelhoffer).

about how ordinary people will answer when presented with actual or hypothetical cases. Experimentalists present themselves as providing much-needed hard, objective data, and consequently use detached, non-participatory social scientific research methods, above all surveys, to obtain it. They deny, at least implicitly, that reflective participation in concept-using practices yields knowledge about what people would say – otherwise, as we shall see, they would not have a case against a priori reflection. In recent years, a welter of survey-based studies has been published on such central concepts as knowledge (Weinberg, Nichols, and Stich, 2001, Swain, Alexander, and Weinberg, MS), reference (Machery et al, 2004), moral judgment (Nichols, 2002, Knobe and Roedder, MS), intentional action (Knobe, 2003, 2004, forthcoming b, Nadelhoffer, 2004), and free will (Nahmias et al, 2004, Nichols, 2004a, Nahmias et al, forthcoming). Other philosophers who have not themselves conducted polls, such as Frank Jackson, Gilbert Harman, and Brian Leiter, have expressed support for them in principle.[5]

So far, experimentalists have put the data generated by their studies to two different kinds of uses. Some, like Weinberg and Stich, highlight the effect of cultural and socio-economic background as well as framing of the questions on people's responses and their consequent variability and instability, raising doubts about the utility of appealing to intuitions in philosophy. Others, like Knobe and Nahmias, are more optimistic and find support for particular philosophical views in their results. To represent this division within the experimentalist school, I will separate the negative and positive theses of experimentalism:

> (EXPERIMENTALISM^{-}) Armchair reflection and informal dialogue are *not* reliable sources of evidence for (philosophically relevant) claims about folk concepts

[5] Jackson talks about his readiness to take polls if needed, though he considers his own judgments as in fact representative (Jackson, 1998, 31). Harman says that we make 'inductive and fallible' inferences from data of the form 'people P have actually made judgments J about cases C as described by D' and suggests that analyses are 'defended in the way one defends any inductive hypothesis' (Harman, 1994, 44). Leiter calls experimental philosophy 'the most important recent development in philosophy' in his widely read blog (http://leiterreports.typepad.com/blog/2004/06/new_experimenta.html).

> (EXPERIMENTALISM$^+$) Survey studies *are* a reliable source of evidence for (philosophically relevant) claims about folk concepts

Both pessimistic and optimistic experimentalists accept the negative thesis. However, while optimists embrace the positive thesis, pessimists reject it (at least the part about philosophical relevance).

The point of departure for my critique of experimentalism is that the proponents of this type of experimental philosophy[6], whether pessimistic or optimistic, ignore the fact that typical philosophical claims of what people would say are elliptical. I identify three characteristic assumptions that philosophers implicitly make about the responses that count as revealing folk concepts – competence of the speaker, absence of performance errors, and basis in semantic rather than pragmatic considerations. I argue that in virtue of these assumptions, intuition statements cannot be interpreted as straightforward predictions, and therefore cannot, for reasons of principle, be tested through the methods of non-participatory social science, without taking a stance on the concepts involved and engaging in dialogue. For example, when philosophers claim that according to our intuitions, Gettier cases are not knowledge, they are not presenting a hypothesis about gut reactions to counterfactual scenarios but, more narrowly, staking a claim of how competent and careful users of the ordinary concept of knowledge would pre-theoretically classify the case in suitable conditions. The claim, then, is not about what I will call *surface intuitions* but about *robust intuitions*, which are bound to remain out of reach of the Survey Model of experimentalists, or so I will argue. Thus, I reject the positive thesis of experimentalism. This leaves the negative thesis. The key challenge for those who, like myself,

[6] In practice, what is called experimental philosophy has been limited to the kind of surveys I discuss in this paper. Insofar as there could be other sorts of experiments yielding philosophically relevant data, my title is somewhat misleading. In the absence of a stable nomenclature, drawing a line between experimental philosophy and other forms of naturalistic, scientifically informed philosophy is somewhat arbitrary.

reject the experimentalist epistemology of concepts, is to explain the source of our entitlement to make claims about laypeople's responses under possibly counterfactual conditions and thus about folk concepts. I argue that this authority is grounded in normative knowledge gained through reflective participation in ordinary concept-using practices. This knowledge is more like our knowledge of how far it is polite to stand from a conversational partner than like our knowledge of what percentage of people believe in angels. It explains the reliability of what I will call the Dialogue and Reflection Models of the epistemology of folk concepts.

2. Testing Intuitions Empirically

I will take the following as the canonical form of philosophical appeals to conceptual intuitions:

> (I) S; In S, we would (not) say that X is C.

Here S is a description of a particular scenario or case, imaginary or real, X an element of that case, and C the concept that applies (or fails to apply) to X. Different concepts call for different grammatical constructions, but I will ignore such complications here, as well as modal formulations (such as 'Intuitively, it is not possible for X to be C in S'). Now, the main question is: how do we find out whether claims of type (I) are true or not? The general schema that experimentalists use in rephrasing intuition claims is something like the following:

> (E) 'In S, we would (not) say that X is C' is a prediction that (most) non-specialists will (not) say that X is C if the case S is presented to them.

Appropriately filled in, a claim of type (E) is a hypothesis that is obviously empirically testable. If it is a correct operationalization of philosophical appeals to intuition, all that remains for a responsible researcher to do is to present a vignette of the case to a statistically representative sample of non-specialists and record their reactions. If a clear majority of the respondents answer as predicted, the intuition claim is (at least probably) true; if not, it is false. If responses are found to vary depending on background factors like socioeconomic class, the utility of appeals to intuition is placed in doubt, as pessimistic experimentalists argue. There's little reason to think that the truth or falsity of predictions of this kind could be reliably decided from an armchair.

How does this work in practice? I will take as my primary example an appeal to intuition that is typical in discussions of moral judgment internalism:

> (I-MJI) Suppose that George frequently says that everyone has a moral duty to make sacrifices during wartime. However, he lacks any motivation to make sacrifices himself, although he is well aware that the war is on, and goes on living just as he always did. In this situation, we would not say that George has made a sincere moral judgment.

The moral internalist uses such intuitions as data for a theory that postulates a necessary conceptual connection between making a sincere moral judgment and being motivated accordingly. The sort of situation in which George makes a judgment and fails to be motivated is not conceivable (in the relevant sense of conceivability). According to the moral internalist, the explanation for why we would not say that George has made a sincere judgment (or why it is not conceivable in this situation) is that the application conditions of our concept of moral judgment, and correspondingly the truth conditions of the thoughts or assertions of which it is a constituent, incorporate the agent's being motivated accordingly, at least to an extent.[7]

In keeping with (E), the experimentalist transforms (I-MJI) into a testable hypothesis along the following lines:

[7] For varieties of internalist theories, see e.g. Hare, 1952, Smith, 1994. Internalists go on to explain *why* our concept of moral judgment includes a motivational component. In short, the reason they offer is that the point of making moral judgments is making a difference to how we act.

> (E-MJI) '[In the case as described above in I-MJI] we would not say that George has made a sincere moral judgment' is a prediction that (most) non-specialists will not say that George has made a sincere moral judgment if the case is presented to them.

(E-MJI), obviously, can be tested by presenting a suitable version of the case to a representative sample of non-specialists. And indeed, something like it has been tested by Shaun Nichols (2002). He gave the following 'probe' to "philosophically unsophisticated undergraduates":

> John is a psychopathic criminal. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. John has hurt and indeed killed other people when he has wanted to steal their money. He says that he knows that hurting others is wrong, but that he just doesn't care if he does things that are wrong. Does John really understand that hurting others is morally wrong? (Nichols, 2002)

According to Nichols, nearly 85% of the subjects responded that the psychopath does really understand that hurting others is morally wrong (which Nichols takes to be the same as making the judgment that hurting others is morally wrong[8]), in spite of entirely lacking motivation. If this is the case, then (E-MJI) is (most likely) false; that is, non-specialists do not seem to think it is impossible to make a moral judgment while lacking motivation. That is, if (E-MJI) is the right way to construe (I-MJI), it is not a *platitude* the grasp of which is necessary for possessing the concept that genuine moral judgments have an internal connection to motivation or a conceptual *intuition* that someone counts as making a moral judgment only if she is motivated accordingly. If so, the thesis of judgment internalism in moral psychology is not a conceptual truth.

A more complex and ambitious variety of experimentalism uses polls not just to settle whether people really have the sort of intuitions that philosophers assume they do, but also to challenge conceptual assumptions that philosophers routinely make. Joshua Knobe's studies on

[8] The moral internalist thesis is not always clearly formulated in the literature, and the formulation of Nichols's question reflects this ambiguity. Properly understood, the thesis has two parts: a person who *understands* what it is to make a moral commitment *and undertakes* such a commitment will have some motivation to act accordingly.

folk psychological concepts are paradigmatic examples of this variety of optimistic experimentalism. Knobe's ingenious idea is to take two cases that differ from each other only with respect to a variable that prevailing views predict to be irrelevant to people's judgments, and then show that there is, in fact, variation in responses depending on changes in the variable. Thus, mainstream views differ about whether foreseen side effects of actions are brought about intentionally or not, but agree that the applicability of the folk concept of intentionality depends exclusively on the agent's psychological states. Consequently, they predict that folk conceptual intuitions about particular cases of intentional action are not affected by factors external to the agent's psychology. To show that the assumption is problematic, Knobe has run a series of experiments pairing cases in which the side effects brought about by the action are morally bad and morally good, respectively. Here is his first scenario:

> The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.’ The chairman of the board answered, ‘I don’t care at all about harming the environment. I just want to make as much profit as I can. Let’s start the new program.’ They started the new program. Sure enough, the environment was harmed. (Knobe, 2003, 191)

In this "harm condition" in which the anticipated side effect is bad, 82% of Knobe's respondents (random people in Central Park) said that the chairman of the board harmed the environment intentionally. His second scenario differs from the first only with respect to the moral status of the side effect[9]:

> The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.’ The chairman of the board answered, ‘I don’t care at all about helping the environment. I just want to make as much profit as I can.

[9] At least, it is Knobe's goal to present a case that differs only in one respect. Finding a precise counterpart case is far from trivial, but I am going to grant here that it is possible.

> Let’s start the new program.’ They started the new program. Sure enough, the environment was helped. (Knobe, 2003, 191)

In this "help condition", 77% of the people asked said that the chairman of the board did *not* help the environment intentionally. Thus, while prevailing views predict symmetry in people's responses to such cases, Knobe's studies suggest that the responses are in fact *asymmetrical*, driven by factors external to the agent's psychology, namely moral considerations. From this and other similar studies Knobe concludes, further, that moral considerations play a role in people's *concept* of intentional action (Knobe (forthcoming b)).

The studies by Nichols, Knobe, and other experimentalists differ in details and aims, but all of them presuppose that something like (E) is the correct operationalization of appeals to intuition. For optimists, responses to surveys yield data against which competing philosophical views and analyses can be assessed. For pessimists, the fact that we can find ordering effects and cross-cultural variations in responses to surveys shows that the whole practice of appealing to folk intuitions is dubious.[10] Both subscribe, nevertheless, to the Survey Model of the epistemology of folk concepts, and endorse the claim that testing folk intuitions is an a posteriori enterprise on par with empirical science. It promises to put philosophy on a path of progress and put an end to vain quarrel. This promise, surely, explains the rise of experimental philosophy.

3. Ellipsis and the Implicit Assumptions of Intuition Claims

[10] For example, Swain, Alexander, and Weinberg (online MS) present data that shows that people’s responses to a putative reliabilist case of knowledge vary depending on what sort of cases they’ve been presented with before asking the question; Weinberg, Nichols, and Stich (2001) discuss various survey results showing a systematic divergence between responses of American and Southeast Asian as well as high-SES and low SES subjects to Gettier and other cases.

Where does the Survey Model go wrong? I believe that construing appeals to intuition as (E) is a natural mistake to make – it is one way to read literally what is often said. If what philosophers claim really was that people are inclined to say x in S, period, the experimentalist construal would be correct. But I will maintain that it is not what we do, in spite of the surface grammar. Instead, philosophical claims about intuitions are typically *elliptical*. Ellipsis is a common linguistic phenomenon; when it is taken to be obvious in the context, people say things like "I love the City" instead of "I love New York" (or London or Helsinki or whatever). Given the purpose of the discourse and shared background assumptions, there is no need to spell out explicitly what is being claimed – indeed, doing so would violate the Gricean maxim of not giving unnecessary information (see below). Similarly, when making claims about intuitions or platitudes, about what "we would say", philosophers take for granted certain background assumptions that, when made explicit, show that (E) is *not* the right way to spell out what is asserted in (I). At least, this is how charity requires us to conceive of their claims. Moreover, as I will argue, these background assumptions are justified in light of the goals of conceptual analysis.

My alternative explication of (I) is the following:

> (A) 'In S, we would say that X is C' is a hypothesis about how (1) *competent users* of the concepts in question would respond if (2) they *considered the case in sufficiently ideal conditions* and (3) their answer was *influenced only by semantic considerations*.

It is central to my case against the positive thesis of experimentalism that requirements (1)–(3) rule out surveys as a method for accessing the semantic application conditions of folk concepts. In this section, I will discuss these requirements and the rationale for them, and address the issue of their testability in section 4.

3.1 Surface Intuitions and Robust Intuitions

There are really two steps in the inquiry into folk concepts. First, if we are asking non-specialists, we want to find out what their individual representations of the concept (or whatever constitutes their grasp of it) are – what Larry's or Anne's or Lily's concept of knowledge is, for example. And we are interested in this because, second, we ultimately want to know what the folk's *shared* concept of knowledge is (or whether there is one in the first place). This latter step does not require that the linguistic behaviour of the folk is completely uniform, but only that they aim to conform their thought and speech to the same constraints as others.[11] We talk about *the* concept of knowledge or *the* concept of moral judgment, after all, and claims about conceptual truths are claims about how proper applications of such public concepts are related to each other.[12] What I will argue is that the sort of information that surveys yield does not warrant taking *either* step. A person's response does not reveal what the extension or intension of her concept is if it results from some other factor such as inattention or pragmatic considerations that surveys do not control for, and it does not, in addition, reveal what the public concept is if she has a poor grasp of it or if she simply makes a mistake in a certain (type of) case. This is why her 'intuitions' count only when the sort of conditions listed in (A) are met. One way to put this is to say that surveys can only inform us of *surface intuitions* that do not help us in the project of finding out the folk concepts. For that purpose, we need *robust intuitions* that are elicited only when conditions in (A) obtain – that is, when failures of competence, failures of performance, and influence of irrelevant factors are ruled out.

[11] For an account of the sort of shared rule-following that makes 'commonable thought' possible, see Pettit, 1996, 180–190.

[12] In Frank Jackson's terms, this amounts to asking whether claims made in one vocabulary (such as that of justified true belief) are made true by the same facts as claims made in another vocabulary (such that of knowledge) (Jackson, 1998).

3.2 Competence and Normativity

It should be obvious that when philosophers appeal to 'us' in making their claims, the extension is limited to those who are competent with the concept in question. After all, what *incompetent* users of a concept say about a given case does not tell us anything about the concept we are interested in – someone who has no relevant pre-theoretical knowledge about the concept cannot manifest it. Nobody would test a Gettier analysis by asking a small child whether the person in the case described knows or not, or count the child's response as a counterexample. And children are only the most obvious example. On many theories of concept possession, competence with a concept is a matter of degree and context.[13] This is to deny that there is, strictly speaking, such a thing as a 'competent speaker of English', for example.[14] To be sure, normal speakers are able to latch on to patterns of proper use and extrapolate correctly to new cases, as long as the similarities and differences between the cases are salient enough. For many practical purposes such ability suffices for competence. But some will be less and some more successful at grasping the rationale guiding application to new cases and thus discriminating between scenarios.[15] Some concepts will be harder to grasp than others – perhaps most people with normal physiological capacities will be able to tell, when presented with a visual scenario, whether an object is white or not, but it is not as easy to tell whether an argument is compelling or whether a person in a

[13] See for example Brandom, 1994.

[14] Here I disagree with Williamson, 2005.

[15] An anecdote may be helpful. The other day my girlfriend was on her way to sauna, and asked me to bring a towel from the bedroom. In the bedroom, I saw two towels, and shouted her "Which one is yours?" She responded with "The peach one." I looked at the towels, neither of which was obviously non-peach to me. So I had to ask: "Which one of these is peach?" The moral is that while we both grasped the concept of being peach-colored, my grasp was not as good as hers. If the other towel had been black or blue, I would have immediately known which one was peach, so clearly I have some idea of what counts as peach. But I could not discriminate between peach and non-peach finely enough to distinguish it from another color closer to peach, such as salmon or amber or even pink (see http://en.wikipedia.org/wiki/Peach_(color) for samples if you are similarly challenged). Given that classification is an important part of what we do with concepts, I plausibly have a less perfect grasp of the concept than my girlfriend and other more color-competent people.

counterfactual scenario should be described as morally responsible or not, if one is to accord with the correct pattern of applications of the concept.

Importantly, as my talk of 'patterns of proper use' and 'correct extrapolation' already suggests, talk of competence brings in *normative* questions.[16] To say that someone is a competent user is to say that she is able apply the concept correctly to a sufficient number of cases (where what counts as 'sufficient' will surely depend on context), and thus to take a stand on what counts as correct use. As discussions inspired by Wittgenstein's remarks on rule-following have shown, such normative claims about correct use (and thus competence) cannot be derived from facts about actual usage or (simple) dispositions to apply the concept. Meaning or conceptual content in effect lays down a rule dividing scenarios into those in which it is appropriate to apply the concept and those in which it isn't. In Kripke's example, to say that one means the addition function by '+' is to say that one *should* respond '125' when the task is to compute '57+68', not that one *will* or *would* so respond.[17] Without such normative constraints, the notion of content would vanish. If a person who applied *red* to pure fallen snow or a cucumber, for example, would not be making a *mistake*, her concept would not have the same content as ours, or perhaps no content at all – anything would equally fall in the extension of the speaker's concept *red*, provided that she was somehow led to apply it to an object.[18] For talk of mistakes to make sense, the concept must set some normative constraints for its use – there must be a distinction between what seems right and what is right.

It follows from the normativity of content that we cannot simply look at what situations someone applies a concept to and infer what the criteria or rules guiding her use are, since that

[16] For this connection, see Kripke, 1982, esp. 31n22.

[17] See Kripke, 1982, 37 and passim. Kripke is, of course, developing a line of argument in Wittgenstein (1953). The distinction between linguistic content and mental content is not relevant to the issue of normativity, as Paul Boghossian (1989, 510) notes. I will use both kinds of examples for simplicity of exposition.

[18] Better yet: *nothing* would fall into a concept's extension, if there were no normative constraints for its application.

would amount to excluding the possibility of making mistakes. What, then, determines which applications count as correct according to a speaker's grasp of a concept? A tempting response is to say that correct applications are those that one is *disposed* to give under suitable conditions. This allows for the possibility of mistakes, since it can be true that I am disposed to do something I do not actually do. It is, however, clear that at least simple forms of dispositionalism do not solve the problem, since, as Kripke points out, we can also be disposed to make mistakes. To take his example of following the rule for addition, he points out that there are some people who are disposed to forget to 'carry' when adding large numbers, so that their answers do not accord with the addition function. The simple dispositionalist will have to say that what these people mean by '+' is some other, more complex function (Kripke calls it 'skaddition'), with regard to which they make no computational mistakes. But this is most implausible. The correct description of the case is, as Kripke puts it, that "for them as for us, '+' means addition, but for certain numbers they are not disposed to give the answer they *should* give, if they are to accord with the table of the function they actually *meant*" (Kripke, 1982, 29). The question about what, if anything, makes it the case that the skadders actually meant to add thus remains open after we have considered their dispositions. Sceptics about meaning, like Kripke himself, say that there is, in the end, no fact of the matter; instead, when we say that someone means *addition* by '+' in spite of making occasional mistakes, we express acceptance into the community as an adder, as someone whose responses can be expected to "agree with those of the community in enough cases, especially the simple ones" (Kripke, 1982, 92).[19] Non-sceptics try to provide for a 'straight solution' that would show that there is, after all, some fact about a speaker or her community that makes it the case the she should give a particular answer, even if she does not actually do so.[20]

[19] This is Kripke's 'sceptical solution'; obviously, there could be sceptics who rejected that as well.

[20] For important papers on the topic, see Miller and Wright (eds.), 2002.

This is a live controversy, and I do not want to enter it here. What is important for my purposes is that the same points apply on the communal level, as many have pointed out. Communities can make mistakes by their own lights, and even be disposed to do so. Arguably, to make sense of the notion of normative constraint, we must allow for the possibility that the rules for our concept *red* determine whether or not it applies to a future case independently of what we will, then, judge – in other words, it may be that we should, by our rules, call something 'red' even if *everybody* in actual fact judges otherwise.[21] If this is the case, there is no way to derive the conceptual norms in force in a community, and thus standards of competence, from the responses or simple dispositions of a majority.

Let me recap the argument of this section. For the purpose of understanding a folk concept, only the responses of competent users count. Competent users are those whose application of the concept generally matches the conceptual norms prevailing in the linguistic community. To sort out incompetent users, one must therefore identify at least the most important norms governing the concept. These norms cannot be derived from either actual use or simple dispositions, individual or collective, since the very notion of normative constraint opens a gap between what people are inclined to say about a particular case and what they should, by their own lights, say about it. It is important to bear in mind that the norms in question are not imported from the outside by the philosopher – rather, they are rules that concept-users at least implicitly are committed to, even if they follow them 'blindly'. If you point out to the person who is systematically forgetting to 'carry' that she is not following the addition rule she thought she was

[21] This is emphasized by McDowell, who argues that communitarians about rule-following cannot account for the crucial distinction between being out of step with the judgments of one's peers and failing to conform to the normative constraints set by the concept. This amounts to losing sight of the commonsense notion of objectivity, which requires that "the patterns to which our concepts oblige us are ratification-independent" (McDowell 1984, 232). McDowell's dense paper aims to show that Wittgenstein provides a non-platonistic vision of how we can grasp a pattern of application extending to future cases independently of the actual outcome of any future investigation.

and explain why, it is most unlikely that the response will be along the lines "but I'm not trying to add, I'm just skadding!". Given the potential gap between actual response and correct response, it will not be a simple task to determine which speakers are competent users.

3.3 Ideal Conditions

I argued in the previous section that appeals to intuition are appeals to the judgment of competent users of concepts. But competence is no guarantee of getting it right. Even competent users can make mistakes, and mistakes do not serve as support or counterexamples to proposed analyses. The conditions in which judgments are made must be conducive to avoiding performance errors. For short, I will call such conditions *ideal*. They are conditions in which there are no perturbing, warping or distorting factors or limits of information, access or ability (Pettit, 1999, 32). There is no single substantively specified set of ideal conditions for applying concepts, since such conditions may vary with the concept in question. Rather, as Philip Pettit notes, we find these conditions implicit in the practices of resolving discrepancies across time or subjects – as we notice differences in responses and look for an explanation for them, we come to discredit judgments made under certain conditions (Pettit, 1999, 29–33). For example, we do not treat judgments about colours made in certain kinds of lightning or certain judgments about responsibility made in an agitated state as authoritative – we understand that there are circumstances in which people are tempted to blame somebody even if in a cool hour they themselves would acknowledge that nobody is to blame.

There are, to be sure, some general things to say about conditions that are favourable for intuitive judgments. When a philosopher says that competent speakers would say certain things, she does not predict that they will respond in a certain way off the cuff. Nor would such response support a philosophical thesis. Appeals to intuition are not appeals to gut reactions, but simply to pre-theoretic judgments that may require careful consideration. It is not always obvious whether a concept applies to a case, nor are users always attentive to relevant details. Giving the answer that reflects one's concept is, naturally, the more difficult the more unusual the case. Being asked to apply a concept to a hypothetical situation that is a remote possibility by ordinary lights can call for advanced skills in counterfactual reasoning – for example, what indeed would we say about responsibility if someone committed the same crime over and over again were the universe re-created over and over again with the same initial conditions and laws of nature?[22] In other cases, there may be considerations weighing in different directions that need sorting out. Concepts form webs and clusters, and it will often be necessary to look at several cases to find patterns, connections, and contrasts. To get it right by one's own lights can take hard thinking and time, and the attempt could be thwarted by passions or loss of interest. There is a general requirement to think through the implications of individual judgments – a hasty judgment or simply a judgment that fails to fit with one's other uses of the concept will not count as one's robust intuition about the case.

3.4 Semantic vs. Pragmatic Considerations

Even if we limited ourselves to responses by competent speakers in ideal conditions, what they would say about particular cases would not necessarily reveal us what we are interested in, namely the *semantic* contours of the concept at hand or the contribution is makes to the truth conditions of sentences in which it is used. The core mistake of early ordinary language philosophy was assuming such a direct link between proper use and meaning (see Soames 2003,

[22] This is the sort of case that Nahmias, Morris, Nadelhoffer, and Turner (forthcoming) ask people to consider in order to find out whether people find the compatibility of free will and determinism intuitive.

esp. chapter 9). This is because the appropriateness of what we say also depends on various pragmatic factors that are not part of the meaning or semantic content of the expression. For example, some things are too obvious to say, others would give a wrong impression in the context. To take a classic example, Ryle claimed in *The Concept of Mind* that 'voluntary' and 'involuntary' are used only for actions which ought not to be done: "We discuss whether someone's action was voluntary or not only when the action seems to have been his fault" (Ryle, 1949, 69). Even supposing that this observation on ordinary use is correct – and there is certainly no reason to doubt Ryle's competence and attention! – the conclusion does not follow. For it is a solid Gricean pragmatic principle that cooperative speakers try to give just the right amount of information given the purposes of the conversation, no more and no less.[23] In most contexts it would be *unnecessary* to say, for example, that I voluntarily had lunch yesterday, unless there was something exceptional to it – as a result, it would typically conversationally implicate that I usually have lunch only involuntarily, or indeed that I accept whatever blame there may be forthcoming for having had lunch yesterday. Consequently, talk of voluntariness would be *misleading* in ordinary contexts, in which the implicatures do not hold, and therefore pragmatically inappropriate, something we would not ordinarily say. But that does not mean it would be *untrue* that I had lunch voluntarily; it would still be semantically appropriate to say so. In general, it is not easy to separate the contribution of semantic and pragmatic considerations to what people say (and what it is *proper* to say) – excellent, trained philosophers have made major blunders – and in surveys of amateurs it is practically impossible.

4. The Failure of the Positive Thesis of Experimentalism

[23] E.g. Grice 1989, 26–27. Compare Mates 1958, 129.

In the previous section, I have argued that the correct explication of (I) is (A) rather than (E). As long as the requirements (1)-(3) above are not tested for, testing for (E) amounts, in effect, to testing for (E*):

> (E*) 'In S, we would say that X is C' is a prediction that (most) non-specialists who (1') *appear to understand the question* will say that X is C if the case S is presented to them (2') *however they consider it in whatever conditions they find themselves in* and (3') *whatever kind of considerations influence* their response.

The truth or falsity of (E*) is surely quite irrelevant to whether our shared concept C properly applies to X, the question that the philosopher is asking. There is no support to be had from responses of those non-philosophers who only appear to understand the question, who may have an imperfect grasp of the concept in question, who may or may not think hard about the application of the concept in circumstances that may or may not be conducive to avoiding conceptual mistakes, who may or may not rush in their judgments, and who may or may not be influenced by various pragmatic factors. Nor do such surface intuitions provide data that is to be explained or explained away, since some of them may be mere noise that does not have to be accommodated in an account of the folk concept. Moreover, these responses are an unreliable guide not only to the public concept, but to the individual respondents' concepts as well, since they apply their own rules fallibly, and do not only respond to semantic factors. My first criticism of the positive thesis of experimentalism – that surveys are a reliable source of evidence for philosophically relevant claims about folk concepts – can then be formulated as follows: the *actual studies* conducted so far have failed to rule out competence failures, performance failures, and the potential influence of pragmatic factors, and as such do not yield the sort of results that could support or raise doubts about philosophical appeals to conceptual intuitions.[24]

[24] It would not be fair to say that experimentalists are not at all sensitive to these issues, and some have tried to ensure that responses go beyond what I am calling surface intuitions. I will discuss some of these attempts shortly.

The crucial question, however, does not concern the studies conducted so far. It is whether it would be, at least in principle, *possible* to test claims of type (A) empirically in the sense that experimentalists recognize, that is, in terms of non-participatory social scientific methods, and if so, how.[25] This is very doubtful. First, as I noted, the question about who is a competent user is a normative question, a question about who gets it right, and it is very hard to see how one could answer it from the detached stance of an observer. To begin with, it seems that experimentalists assume – and must assume – that meaning or conceptual content supervenes on actual use or simple response-dispositions of speakers. After all, that is what they are testing for. Taking polls is a more or less reliable way to discover facts about actual use, and thus, they implicitly assume, a more or less reliable way to discover facts about what people mean by their words or what their concepts are. But as I already noted, what counts as correct use for the folk cannot be derived from the folk's actual use of the concept in question, so that standards of competence cannot be established with reference to majority response. Further, though one may presume that respondents have general mastery of a language, that is not enough, given that there are local variations in competence. One may be a minimally competent user of a concept, having the sort of rough grasp that enables one to converse about central cases, but lack sufficient understanding to apply it to philosophically interesting cases. (Think about someone for whom subjective certainty is a central element in the concept of knowledge.) To be sure, it is possible to ask control questions to rule out, for example, the responses of people who identify knowledge with subjective certainty, as Weinberg, Nichols, and Stich (2001) did. But to make this commendable move is already to take a stance on what the folk concept of knowledge is – in this case, to distinguish between two different everyday senses of 'knowledge' and take one of them to be philosophically relevant. Control questions amount to presupposing that

[25] This question was pressed by anonymous referees for *XX*.

certain answers will not reflect the folk concept, and these presuppositions cannot, by definition, be justified by means of surveys. (I agree, of course, that they can often be otherwise justified, but that is to deny the negative thesis of experimentalism.) The burden is on the experimentalist to present a neutral test of who is a sufficiently competent user of the folk concept of knowledge or moral judgment, for example.

Second, testing for ideal conditions and careful consideration does not seem to be possible without engaging in dialogue with the test subjects, and that, again, violates the spirit and letter of experimentalist quasi-observation. What is needed is a way of checking whether the test subject is making a performance error by her own lights. We can imagine a researcher going through a test subject's answers together with her, asking for the reasons why she answered one way rather than another, making sure she really did correctly understand the counterfactual scenario involved and did not read more or less into it than described in the test, pointing out similarities and disanalogies with other cases of, say, knowledge or moral judgment, and trying to get her to reflect on whether her response is really what she wants to say in the case in point – whether she is really following her own rules. But this is no longer merely 'probing' the test subjects. It is not doing experimental philosophy in the new and distinct sense, but rather a return to the good old Socratic method (see below). Again, the burden is on the experimentalist to show how we could ascertain that ideal conditions obtain for each of the test subjects without leaving detachment behind. Otherwise, we might as well skip the superfluous and unreliable survey and go straight into dialogue.

Finally, testing for the influence of pragmatic considerations is no simple matter either, though here the problems seem to be practical rather than principled. The distinction between semantics and pragmatics is a matter of much contention in contemporary philosophy of language[26], but one could perhaps roughly say that the semantic content of a sentence is determined by the standing meaning of the lexical items, syntactic rules for their combination, and those elements of the non-linguistic context that are needed to resolve the reference of lexical items in accordance with their standing meaning.[27] (The last clause allows indexicals, for example, to contribute to the semantic content of a sentence.) In short, semantic content comprises what is required for a sentence to express a truth-evaluable proposition in context. Depending on the context of utterance, a speech act may express a number of other propositions over and above its semantic content. Gricean conversational and conventional implicatures are at least a central class of these non-semantic – that is, pragmatic – contents. Now, while Grice provides criteria for what counts as an implicature and some tests for differentiating between what is said and what is meant, it is not at all clear how to apply them to a survey situation. For example, Fred Adams and Annie Steadman (2004) have suggested that in the Knobe study, people say that the CEO brings about the side effect intentionally in the harm condition because they want to blame the CEO and think that blaming and intentionality go together. In other words, saying that the CEO did *not* bring about the side effect intentionally would implicate that he is not to blame for it. Adams and Steadman claim that this and the converse implicatures in the help condition show that pragmatic considerations explain the asymmetry in Knobe's results.[28] The question is how to test for this. On Grice's view, conversational implicatures are

[26] See for example the papers in Szabo (ed.) 2005, and Cappelen and Lepore 2005.

[27] For an extended discussion on alternative ways of making the distinction, see Stanley and King (2005).

[28] Knobe (2004) is an ingenious attempt to circumvent the problem by substituting an armchair-equivalent (!) phrase, "in order to", that allegedly lacks the pragmatic implicatures of "intentionally"; the obvious response is that if the phrase really has the same content as the original term, it is no surprise if it has similar implicatures. This is something to be settled case by case. Indeed, sharing implicatures is plausible in the case in question, since if there is a conversational implicature at play, it is a generalized rather than a particularized one – that is, if saying that A did not do something intentionally implicates that A is not to blame, it does so normally or in most contexts – and as Grice argues, generalized conversational implicatures have a high degree of nondetachability: "Insofar as the calculation that a particular conversational implicature is present requires, besides contextual and background information, only a knowledge of what has been said (or of the conventional commitment of the utterance), and

essentially such that they can be worked out, given the assumption of conversational cooperation and facts about the context, including mutual beliefs (Grice 1989, 31). But when a person responds to a yes/no survey question (or rates assent on a Likert scale), just what is the conversational context? Who is he or she conversing with, and how do we work out what he or she assumes about the hearer's beliefs? Frankly, this is a baffling task.[29] Once again, an actual dialogue would help, but it would mean leaving behind the Survey Model and its pretension to scientific objectivity.

The Argument from Disagreement

At this point the experimentalist may well say: “But surely there is *something* about the test subjects’ psychology that explains why most of them say psychopaths make moral judgments or why the moral status of side effects makes such a dramatic difference to attributions of intentionality!” This is, of course, true as far as it goes. The question, however, is whether these results tell us something relevant about folk *concepts*, as experimentalists claim. So far, I have argued that for that to be the case, certain conditions would have to be fulfilled, and that experimentalist methods cannot test for them. But there is also a different reason to be suspicious about the inference from poll results to the contours of the folk concept. I will call it the Argument from Disagreement. Its starting point is that agreement and disagreement about a subject matter presuppose that both parties are talking about the same thing, and this in turn presupposes that they share the same concept. For the purposes of this argument, it does not

insofar as the manner of expression plays no role in the calculation, *it will not be possible to find another way of saying the same thing, which simply lacks the implicature in question*" (Grice 1989, 39, my emphasis).

[29] In the case of conversational implicatures, one might be able to test for cancellability – that is, whether people regard it possible to say without contradiction that, for example, the CEO did not intentionally harm the environment, but he is still to blame for it, since he knew it was going to happen. If they agreed to this, and still went on to say the CEO did what he did intentionally in the harm condition, this would support Knobe's case.

matter what counts as sharing a concept, but for simplicity, I will say that the concepts of two people are identical if they make the same contribution to the truth conditions of claims involving that concept in all possible worlds.[30] Thus, if my concept of redness is the concept of a surface reflectance property R_1 and your concept of redness is the concept of surface reflectance property R_2, when I say that something is red and you say that it is not red, there is no disagreement between us; by calling the object 'red' I have attributed one property it and you another. We are merely talking past each other, and both our claims may be true at the same time. Agreement and disagreement thus both presuppose that our talk expresses a shared concept.

Now, let us consider the results of the surveys I mentioned. They are without exception *mixed*: a certain percentage of people are ready to apply the concept in question, while others will refrain. What can we say in such a situation? In the Nichols survey on moral internalism, 85% of those surveyed answered that a psychopath with no moral motivation can still understand that hurting others is wrong, while 15% disagreed. Nichols took this lopsided result to show that the internalist account of the folk concept of moral judgment is mistaken. But the argument from disagreement shows that this inference is illegitimate. It can be formulated as a dilemma. Either a test subject's response to a survey question reveals whether the case falls under her concept, or it doesn't. If it doesn't, the response is obviously uninformative and running the survey for this purpose pointless. But what if it does? Then those who answer in the negative will not and can not *disagree* with those who answer in the affirmative. Were this the case in the moral internalism survey, the minority's concept would incorporate the internalist motivation condition, so that attributions of moral judgment to people who are not appropriately motivated would be

[30] This is an extensional criterion, and in need of amendment to deal with necessarily co-extensional but distinct concepts, if such exist.

false, while the majority's concept would lack this condition and thus contribute differently to truth conditions of claims about moral judgment. In other words, there would be two distinct concepts at play, moral judgment-$_{\text{EXT}}$ and moral judgment-$_{\text{INT}}$, the first employed by the majority and the second by the minority, and consequently no disagreement between the two groups – they would just talk past each other when using the phrase 'moral judgment'. But this is absurd. We know that in most such cases the people in fact disagree and consequently that they share a single concept, regardless of how they respond to the survey question. What, then, could we learn about the concept from asking the question? This is the argument from disagreement.

The obvious fallback position for the experimentalist is that the majority response reveals the correct application of the shared concept, and those in the minority are simply wrong in their application of the shared concept (that is, either lacking in competence or making a performance error). This is a step in the right direction to the extent that it acknowledges a gap between one's dispositions to apply a concept and the proper application of one's concept. But as we already saw, correctness is not a matter of going with the crowd – it is certainly possible that a majority is mistaken in the application of the shared concept. It is not a priori true that minorities cannot be more competent, more careful, or more aware of pragmatic influences in concept application than majorities. And why would the responses of those who happen to be in the majority not only reveal their concept but also that of those in the minority, while the responses of those in the minority would reveal nothing about anybody's concept? Whence the asymmetry? What entitles us to throw out the responses of the minority when we are trying to get at a concept shared by those in the majority and those in the minority? We need an explanation of why we should prioritize the majority's responses, and the mere fact that they are a majority is no such thing. Thus, we have another reason to believe that the experimentalist approach leads to a dead end. It cannot tell us what the folk's shared concepts are, because it cannot tell us which responses count as revealing them and which do not.

It may be worth pointing out that neither the argument from disagreement nor the conditions on robust intuitions show, or try to show, that the results obtained by experimentalists do not call for an explanation of *some* sort. There must indeed be a reason why people have the surface intuitions they do. That reason will no doubt be different in different cases. Some responses will indeed result from correct application of a shared concept. Others may be driven by a desire to blame, for example, and yet others by charity or some other factor that leads test subjects to read more into vignettes than they explicitly state.[31] In some cases, the explanation may be philosophically interesting, even if not for the purposes of conceptual analysis. Mostly, however, the explanation will be psychological, and nothing I have said bears on the possibility of experimental *psychology*.

5. From Surveys to Dialogue and Reflection

In the previous sections I have argued that the Survey Model of the epistemology of folk concepts and with it the positive thesis of experimentalism fails, in spite of its initial promise and plausibility. This leaves the negative, sceptical thesis. How can we confirm or disconfirm appeals of the form (I), if not by running polls? How do we gain traction with the concepts of ordinary folk? How do we elicit robust intuitions instead of surface intuitions? I have argued that this is a question about assessing counterfactuals of type (A) about what competent, careful speakers would say in favourable conditions if they abstracted away from pragmatic considerations – in

[31] I argue for a charity-based explanation of the results in Knobe and Roedder (MS) in my contribution to the first Online Philosophy Conference, 'Lovers of the Good: Comments on Knobe and Roedder' (http://garnet.acns.fsu.edu/~tan02/OPC%20Week%20Three/Commentary%20on%20Knobe.pdf).

other words, what the semantic rules of our language say about the application of the concept in question. For example, when we ask whether genuine moral judgments conceptually involve corresponding motivation according to ordinary folk, we are in effect asking whether competent, careful speakers guided only by semantic considerations would describe someone who says he thinks something is wrong but is not disposed to refrain from doing it as having made a sincere moral judgment. This is why the fact, if it is such, that 85% of those queried in a college classroom or a Manhattan park say a psychopath can make a moral judgment while being entirely unmoved does not contradict the strong internalists' claim about what non-specialists *would* say, since that claim is only about the folk's responses in the kind of conditions that I have outlined – about robust rather than surface intuitions, in the terminology introduced above. And there is no way for a philosopher to ascertain how people would respond in such a situation without breaking the fourth wall to create the relevant conditions – that is, without entering into dialogue with them, varying examples, teasing out implications, presenting alternative interpretations to choose from to separate the semantic and the pragmatic, and so on. I will call this approach the Dialogue Model of the epistemology of folk concepts.

5.1 The Dialogue Model

How and why does philosophical dialogue work? I will take a concrete example. In the moral internalism case, one might present non-specialists with the sort of scenario in which someone is trying to convince a friend to adopt a moral stance, say become a vegetarian, and ask if the friend has really become convinced of the wrongness of eating meat as long as he has no inclination of refraining from doing it. If they answer negatively, the philosopher could point out that this seems inconsistent with attributing genuine moral convictions to the psychopath, perhaps try to disambiguate in which sense one might be inclined to say the psychopath ‘understands’ something is wrong, and so on. If we want to get people to respond in ways that represent the rules they actually are committed to, we cannot pretend to be ignorant of common distorting factors – to use the Kripkean example, if we want to find out whether someone’s concept of addition is the same as ours, we are surely entitled to point out an obvious failure to carry in a particular case and so perhaps get her to see the application as a mistake by her own lights. This process of making sure that a particular response genuinely reflects the respondent’s concept is hard work, and never free of the danger of leading the witness in the direction favoured by the questioner. To engage in it is to give up the detached stance of the experimentalist observer and involve the folk in the fray of philosophical debate instead of reaching for a magic foothold outside of it. It is to do philosophy pretty much as it has always been done.

Suppose that through dialogue or otherwise one manages to create circumstances in which one can be reasonably confident that the conditions listed in (A) are fulfilled and the responses of competent non-specialists do in fact reveal their concepts. Perhaps the majority still responds to the original case in the same way. We could then say that the survey results have turned out to be robust, since they remain unchanged after potentially distorting factors have been pruned off. Robust intuitions, we might say, are represented by those responses of non-specialists that are *stable* under arbitrary increases in consideration of relevantly similar situations, ideality of circumstances, and understanding of the workings of language (centrally, the semantics/pragmatics distinction). In idealized terms, these responses are those that infinitely patient and focused respondents would give at the end of a dialogue with a Super-Socrates, who never misleads but engages in maximally skilful midwifery that consists in bringing about conditions (A). In practice, responses are more or less robust, depending on how closely the ideal dialogical situation is approximated. Since we use language to communicate with each other and sharing concepts is necessary for agreement and disagreement, there is strong a priori reason to believe that people's robust intuitions will line up with each other, at least in central cases.

Now, it is clear that as long as we are talking about the folk's own concepts, robust intuitions of competence speakers must count. This really is data that must be explained or explained away, but it is not obtained by surveys, at least not surveys alone, since they cannot discriminate between responses that are robust and responses that are not. Yet it is the only kind of data worth having. For example, if there are robustly varying patterns of response to a number of cases between, say, Americans and South-East Asians, we can legitimately infer that there are two or more concepts at play and thus no straightforward agreement or disagreement between the groups.[32] And if a clear majority of robust intuitions support counterexamples to internalism, the internalist must withdraw his claim of having intuitive support and capturing the folk concept of moral judgment.[33]

5.2 The Authority of Reflection

Although I have defended dialogue as a privileged means of access to folk concepts, I do not think that it is strictly speaking necessary for philosophers to go around bothering their friends and acquaintances. In practice, assessing the truth of intuition claims can remain a relatively armchair business that begins with our own considered reactions to the case at hand. We are entitled to have confidence in such reflection, since we take a lot of real life experience of using

[32] This would indeed be the case if the results cited in Nichols, Stich, and Weinberg (2003) were robust – which we do not learn from their surveys alone, as I have argued.

[33] However, lack of intuitive support does not yet mean that the moral internalist, for example, has to give up on her view or even modify it to accommodate legitimate counterexamples. It simply means that if she holds on to it, she will be to some extent a *revisionist* about moral judgment. This is to incur a cost, but if other alternatives are even more costly – say, they make no evolutionary sense – it is a cost that may be worth paying.

concepts to the armchair with us. Having participated in the relevant language games and having been corrected and sanctioned while we were learning the concepts under analysis, we have gained normative knowledge about criteria for their proper application to particular cases as well as about inferential relations among them. We know which moves are acceptable in our linguistic community (that is, acceptable by those who share our concepts), since we have received and given feedback to and from others. By the time we begin to do philosophy, we have accumulated years and years' worth of experience about what counts as proper application of concepts to different cases – we have, as it were, already done the sort of research I sketched above in connection with the Dialogue Model. This experience is what grounds the epistemic authority of the Reflection Model. Since dialogue and reflection do yield philosophically relevant evidence about folk concepts, the negative thesis of experimentalism fails as well. Pessimism about dialogue and reflection is no more warranted than optimism about surveys.

At this point, the experimentalist may reach for her last card: is not the long history of contentious appeals to conceptual intuitions by itself sufficient to show that armchair reflection leads nowhere? To see why this is not the case, we have to look at the practice of philosophical reflection on concepts more closely. To begin with, the normative knowledge that this reflection draws on, like similar knowledge of the rules of etiquette, for example, is implicit, and it is rarely easy to articulate and summarize it in an analysis. Some are better at making it explicit, some worse. There are connections between concepts to be missed and cases that are hard to fit in a pattern. There may be an unexpected but plausible pragmatic explanation for why something is inappropriate. Unbiased reflection is not easy, and the activity of reflecting itself may even destroy pre-reflective knowledge – many philosophers know how it feels to lack confidence in one's own reactions to particular cases. No wonder, therefore, that philosophers disagree among themselves both about appeals to intuition in particular cases and about the implications of particular cases for understanding concepts. But that is neither the end of the story nor a cause for desperation, for it is through mutual correction over time that explications of rules edge closer to the actual norms implicit in how we use terms. As examples and counterexamples mount, some intuitive judgments may well come to be widely accepted as data points that have to be accounted for, and may lead to theoretical convergence as well.

The moral internalism debate may serve as a case study of gradual convergence of intuitions through (philosophical) dialogue and reflection. In a classic article from 1937, Charles Stevenson presented a typical case of moral persuasion:

> "When you tell a man that he oughtn't to steal, your object isn't merely to let him know that people disapprove of stealing. […] If in the end you do not succeed in getting *him* to disapprove of stealing, you will feel that you've failed to convince him that stealing is wrong." (Stevenson, 1937, 19)

Here Stevenson appeals to his readers' intuitions to support his internalist contention that "[a] person who recognizes X to be 'good' must *ipso facto* acquire a stronger tendency to act in its favour then [sic] he otherwise would have had." (Stevenson, 1937, 16). [34] This 'magnetism' of moral judgment is a central plank of his argument for emotivism. Unsurprisingly, externalist critics challenged his contention. Here is Henry David Aiken:

> "I may recognize, for instance, that the music of Tschaikowski is 'good,' since many honest and discriminating people have affirmed its power to move and to please, and yet not in the least be impelled to listen to it. […] Moreover, during periods of weariness or satiety, especially, 'goods' which we believe and gladly acknowledge to have the profoundest import to ourselves often leave us quite cold, and our judgment that they are 'good' has no magnetism or persuasive power whatever." (Aiken, 1944, 461)

[34] To keep the discussion focused on methodological issues, I will here ignore the difference in the motivational relevance of moral value judgments and ought-judgments, though I believe it is very important for serious moral psychology. I am also granting Stevenson that disapproval implies motivation.

From such cases, call them *conventional* and *akratic* uses of 'good', respectively, Aiken drew the externalist conclusion that "there is such a thing as the acknowledgement of truth or falsity of an ethical judgment when no 'magnetism' is involved" (Aiken, 1944, 461). Of course, internalists did not give up. But, importantly, their defense did not involve headbutting intuition against intuition or even outright rejection of the intuitive status of externalist claims like those of Aiken. R. M. Hare's *Language of Morals* gave influential responses to both of the two intuitive challenges I have picked out. First, Hare argued that we sometimes, consciously or unconsciously, use 'good' and 'ought' in an inverted commas sense to indicate, for example, that something is considered to be good by most people or experts, without endorsing the evaluation ourselves (e.g. Hare, 1952, 167). In other words, though Aiken and other externalists are right about what we would say in cases of conventional usage, these cases are not relevant to understanding moral judgment. Hare *accepts* the conceptual intuitions, but provides a different *theoretical* account of their status, appealing to a kind of pragmatic explanation.

In the case of akratic uses, Hare adopts a different strategy that amounts to solving the problem by a more complete description of the case. Here is what he says:

> "If a person does not do something, but the omission is accompanied by feelings of guilt etc., we normally say that he has not done what he thinks he ought." (Hare, 1952, 169)

The point that Hare is making is that, as the externalist points out, we do grant that someone may in some cases lack sufficient motivation to act as she thinks she ought, but, what the externalist fails to notice, we do so *only if* that person experiences guilt or some other residual feelings for failing to follow through on her judgment. We could say that Hare *refines* the case to draw out more clearly what is actually driving the intuitive judgment.

I lack the space to follow this dialectic all the way through to today, but I want to point out some important features of it. First of all, a number of judgments about particular cases have become broadly accepted data points that theories of moral motivation must accommodate. Externalists by and large accept that people are usually motivated by their judgments. As Russ Shafer-Landau, a prominent externalist, coolly puts it:

> "There are two especially relevant items to be entered into evidence. Fact: everyone we know is motivated to some extent to comply with his or her moral judgements. Fact: we suspect the sincerity of someone who proclaims fidelity to a moral code, all the while showing no inclination to abide by it." (Shafer-Landau, 2003, 156)

At the same time, internalists by and large accept that akratics, depressives, and certain kinds of amoralists may nonetheless make genuine moral judgments without being motivated by them (see e.g. Smith, 1994, 133–136). So, there is a *convergence of intuitions* – not a perfect one, to be sure, but clearly noticeable. Second, disagreement in theoretical accounts still persists, suggesting that it is best explained by difference in auxiliary commitments in philosophy of language, philosophy of mind, or metaphilosophy. Externalists think they can explain away the internalist intuitions, or at least show that they are not robust, and vice versa. Correspondingly, there is no reason to think that experimental data, even if accepted on all sides, would take the debate forward any better than the data provided by dialogue and reflection. And finally, though disagreement in philosophical explanations indeed persists, there has been a significant degree of convergence on that front as well. Consider the position of Michael Smith, a leading contemporary internalist:

> "[A]gents who judge it right to act in various ways are so motivated, and necessarily so, absent the distorting influences of weakness of the will and other similar forms of practical unreason on their motivations." (Smith, 1994, 61)

Smith thinks that it is a conceptual truth that moral beliefs motivate agents insofar as they are rational, because, on his view, moral beliefs are beliefs about reasons and beliefs about reasons motivate rational agents (Smith, 1997). Given that he postulates a non-contingent connection between judgment and motivation, his view is in one obvious sense an internalist one. Yet it allows for akratics, depressives, and even certain kinds of amoralists to make genuine moral judgments without motivation, if only at the price of irrationality.[35] For this reason, R. Jay Wallace in a recent article describes Smith as someone who *rejects* moral judgment internalism (Wallace, 2006, 185).[36] Mutually accepted intuitions and counterexamples have here led to a situation in which insofar as the classification of a view is not a merely verbal matter, it hangs on theoretical considerations rather than on capturing cases. Philosophical progress has been made.

There is no reason to think that the debate about moral judgment is an exception. Of course, it does not show that reflection and dialogue will *always* lead to a convergence. But in many cases such failures will be explicable in terms of the nature of the concept in question. Perhaps the ordinary concept is simply too vague for philosophical purposes, so that there is no fact of the matter which of the competing appeals to what we would say better conforms to the rule implicit in everyday use. Or perhaps the use of the concept in question is subtly context-dependent, so that the intuitions that one side draws on are robust in one context and the intuitions of the other side in another, and the appearance of conflict is illusory.[37] Neither sort of case threatens the authority of reflection or the fruitfulness of dialogue.

[35] Externalists like David Brink claim that principled, rational amoralists are conceivable, and Smithian internalism thus fails (Brink, 1997, 18–21). Here there is a genuine open dispute about a particular case, but as it hangs on technical terms (who counts as rational?), it is not a counterexample to convergence in ordinary language intuitions.

[36] Certainly Smith – as well as Simon Blackburn (1998, 61) and other recent defenders of internalism – counts as an externalist by Shafer-Landau's criterion, according to which externalists are those who accept "the conceptual possibility of an agent who on a single occasion fails to be motivated by a moral judgment that he endorses" (Shafer-Landau, 2003, 145).

[37] I argue that this is the case with respect to the concept of moral responsibility in my 'Talk About Responsibility' (in preparation).

In this final section, I have defended the epistemological authority of dialogue and reflection with respect to the content of folk concepts against the negative thesis of experimentalism, and provided a case study to undermine pessimism grounded in presumed lack of progress in debates about folk intuitions. I have charted the rise and fall of experimental philosophy to vindicate the self-confidence of traditional philosophers. In short, as philosophers, we continue to participate in ordinary linguistic practices, but do so reflectively, paying careful attention to what is appropriate and why and drawing on the insights of those who have explored the same paths before. Running a poll provides no shortcut in this business of reaching a better conceptual self-understanding. At best, survey results provide food for thought – but we are better nourished if instead of designing artificial setups we pay close attention to what is said in real life situations of language use, as conscientious philosophers have done at least since Socrates.[38]

[38] In preparing this paper, I have benefited enormously from fair and generous feedback by philosophers who are sympathetic to or committed to the experimentalist approach, including Joshua Knobe, Thomas Nadelhoffer, Eddy Nahmias, Alfred Mele, and Jonathan Weinberg. In the same group belong, I suspect, the two anonymous referees for XX. In addition, I want to thank for valuable comments Joel Anderson, Pekka Mäkelä, Alyssa Ney, Lilian O'Brien, Jussi Suikkanen, and the audience at the 2004 Brown Graduate Conference, where an earlier version of this paper was presented.

References

Adams, F. and A. Steadman. 2004. Intentional action in ordinary language: core concept or pragmatic understanding? Analysis, 64.2, 173–181.

Blackburn, S. 1998. Ruling Passions. A Theory of Practical Reasoning. Oxford: Oxford University Press.

Boghossian, P. 1989. The Rule-Following Considerations. Mind 98, 507–549.

Brandom, R. 1994. Making It Explicit. Cambridge, Mass: Harvard University Press.

Brink, D. O. 1997. Moral Motivation. Ethics 108, 4–32.

Cappelen, H. and Lepore, E. 2005. Insensitive Semantics. A Defense of Semantic Minimalism. Oxford: Blackwell.

Chalmers, D. 2002. Does Conceivability Entail Possibility? In T. Gendler and J. Hawthorne (eds.), Conceivability and Possibility. Oxford: Oxford University Press, 145–200.

DeRose, K. 2005. The ordinary language basis for contextualism and the new invariantism. Philosophical Quarterly 55 (219), 172-198.

Fodor, J. 1998. Concepts. Where Cognitive Science Went Wrong. Oxford: Oxford University Press.

Grice, P. 1989. Studies in the Way of Words. Cambridge, Mass.: Harvard University Press.

Hare, R.M. 1952. The Language of Morals. Oxford: Clarendon Press.

Harman, G. 1994. Doubts about conceptual analysis. In M. Michael and J. O'Leary-Hawthorne (eds), Philosophy in Mind. The Place of Philosophy in the Study of Mind. Dordrecht: Kluwer, 43–48.

Hawthorne, J. 2004. Knowledge and Lotteries. Oxford: Oxford University Press.

Jackson, F. 1998. From Metaphysics to Ethics. A Defense of Conceptual Analysis. Oxford: Oxford University Press.

Jackson, F., P. Pettit, and M. Smith. 2004. Mind, Morality, and Explanation. Selected Collaborations. Oxford: Oxford University Press.

King, J. C. and Stanley, J. 2005. Semantics, Pragmatics, and the Role of Semantic Content. In Zsabo (ed.) 2005, 111–164.

Knobe, J. 2003. Intentional action and side effects in ordinary language. Analysis, 63, 190–193.

Knobe, J. 2004. Intention, intentional action and moral considerations. Analysis, 64, 181–187.

Knobe, J. (forthcoming a). Experimental philosophy. Philosophers' Magazine.

Knobe, J. (forthcoming b). The Concept of Intentional Action. A Case Study in the Uses of Folk Psychology. *Philosophical Studies*.

Knobe, J. and Roedder, E. The Concept of Valuing: Experimental Studies. Online MS (http://garnet.acns.fsu.edu/~tan02/OPC%20Week%20Three/knobe.pdf). Accessed June 17, 2006.

Kripke, S. 1982. Wittgenstein on Rules and Private Language. Oxford: Basil Blackwell.

Lyas, C. (ed.) 1971. Philosophy and Linguistics. London and Basingstoke: MacMillan.

Machery, E., R. Mallon, S. Nichols, and S. Stich. 2004. Semantics, cross-cultural style. Cognition, 92, B1–B12.

Margolis, E. and S. Lawrence. 2003. Should we trust our intuitions? Proceedings of the Aristotelian Society, 103, 299–323.

Mates, B. 1958. On the verification of statements about ordinary language. Originally in *Inquiry* 1, reprinted in Lyas (ed.) 1971, 121–130.

McDowell, J. 1984. Wittgenstein on Following a Rule. Reprinted in Mind, Value, and Reality. Cambridge, Mass.: Harvard University Press, 221–262.

Miller, A. and C. Wright. 2002. Rule-Following and Meaning. Chesham: Acumen.

Nahmias, E., T. Nadelhoffer, S. Morris, and J. Turner. 2004. Surveying free will: folk intuitions about free will and moral responsibility. Unpublished manuscript. Florida State University.

Nahmias, E., S. Morris, T. Nadelhoffer, and J. Turner (forthcoming). Is Incompatibilism Intuitive? Philosophy and Phenomenological Research.

Nichols, S. 2002. How psychopaths threaten moral rationalism: is it irrational to be amoral? The Monist, 85, 285–304.

Nichols, S. 2004a. The folk psychology of free will: fits and starts. Mind and Language, 19, 473–502.

Nichols, S. 2004b. Folk concepts and intuitions: from philosophy to cognitive science. Trends in Cognitive Sciences.

Nichols, S., S. Stich and J. Weinberg. 2003. Meta-Skepticism: Meditations on Ethno-Epistemology. In S. Luper (ed.), The Skeptics.

Pettit, P. 1996. The Common Mind. An Essay on Psychology, Society, and Politics. Paperback Edition. Oxford: Oxford University Press.

Pettit, P. 1999. A theory of normal and ideal conditions. Philosophical Studies, 96: 21–44.

Putnam, H. 1975. The meaning of 'meaning'. In K. Gunderson (ed.), Language, Mind, and Knowledge. Minneapolis: University of Minnesota Press, 131–193.

Ryle, G. 1949. The Concept of Mind. London: Hutchinson's.

Smith, M. 1994. The Moral Problem. Oxford: Blackwell.

Soames, S. 2003. Philosophical Analysis in the Twentieth Century. Volume 2: The Age of Meaning. Princeton: Princeton University Press.

Stevenson, C. L. 1937. The Emotive Meaning of Ethical Terms. Mind, 46 (181), 14–31.

Swain, S., Alexander, J., and Weinberg, J. The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp. Online MS. (http://garnet.acns.fsu.edu/~tan02/OPC%20Week%20Two/Swain,%20Alexander,%20and%20Weinberg.pdf) Accessed June 17, 2006.

Szabo, Z. (ed.) 2005. Semantics vs. Pragmatics. Oxford: Oxford University Press.

Wallace, R. J. Moral Motivation. In J. Dreier (ed.), Contemporary Debates in Moral Theory. Oxford: Blackwell, 182–196.

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

Weinberg, J., S. Nichols, and S. Stich. 2001. Normativity and Epistemic Intuitions. *Philosophical Topics*, 29, 429-460.

Williamson, T. 2005. Armchair philosophy, metaphysical modality and counterfactual thinking. Proceedings of the Aristotelian Society, CV (1), 1–23.