Accentuate the Negative

Joshua Alexander
Siena College

Ron Mallon
University of Utah

Jonathan M. Weinberg
Indiana University

1. Introduction

The phrase "experimental philosophy" can look as if it is picking out some one thing: at a minimum, some sort of philosophy that makes some use of experimentation. But in fact there are a number of different programs that fall under the umbrella of "experimental philosophy", and our interest here is to drive a wedge of contention between two of them. These two programs concern traditional analytic philosophy's practice of appealing to philosophical intuitions as either evidence for (or against) philosophical claims or data both about the nature of our folk philosophical concepts and judgments and about the nature of the domains in which we make those judgments.[1] According to what is sometimes called experimental philosophy's "negative program", experimental philosophy challenges the well-functioning of this practice.[2] According to experimental philosophy's "positive program", experimental philosophy is (at least an indispensable part of) the proper methodology for this practice.[3] Seeing these programs juxtaposed like that, one might well wonder whether the practice of appealing to intuitions once modified by the positive program can withstand the challenges that the

[1] For our purposes, philosophical intuitions are propositional attitudes generated in response to hypothetical cases in philosophy which are "minimally foundational" (a person may appeal to them as evidence without having to provide evidence for them), non-inferential, and fallible.

[2] The terms "negative" program and "positive" program are now in common use. We are unsure of their origin though they may have been coined by Farid Masrour.

[3] For additional discussions of experimental philosophy's negative program, see J. Alexander and J. Weinberg (2007); A. Kauppinen, (2007); T. Nadelhoffer and E. Nahmias (2007); and J. Weinberg (2007). For additional discussions of experimental philosophy's positive program, see J. Alexander and J. Weinberg (2007); A. Kauppinen (2007); and T. Nadelhoffer and E. Nahmias (2007).

negative program issued against their armchair-bound predecessors. In this paper, we will contend that the answer is, shall we say, in the negative.[4]

2. The Positive Program

In order to canvass the problems for the positive program, we need first to recognize that there really are a range of positive programs extant in this still-young literature. In an earlier taxonomy, Thomas Nadelhoffer and Eddy Nahmias distinguished between two forms of positive experimental philosophy: "experimental analysis" and "experimental descriptivism".[5] For proponents of experimental analysis, philosophy is (at least in part) concerned with understanding the nature of such things as knowledge, justification, meaning, moral responsibility, and morally right action. For proponents of philosophical descriptivism, philosophy is (at least in part) concerned with understanding the nature of folk concepts - how people think about these things. Both experimental analysts and experimental descriptivists think that intuitions provide an important source of evidence for philosophy. What distinguishes both experimental analysis and experimental descriptivism from more traditional philosophical programs is the way in which we are supposed to go about gathering this evidence. According to proponents of more traditional philosophical investigation, we can determine what intuitions are (or would be) generated in response to particular cases simply by determining what our own intuitions are about those cases (e.g. Jackson 1998). Assuming that our own intuitions

[4] We want to be clear that we are not at all challenging the positive program's status as *philosophy*. We take that to be a question that has been conclusively answered in the (no pun intended) positive. Our concern, rather, is that the positive program may fall prey to concerns that we would just as much raise for more traditional philosophical methods.

[5] T. Nadelhoffer and E. Nahmias (2007).

are appropriately representative – or alternatively, assuming that because of philosophical training or acumen, they are superior to folk intuitions (e.g. Ludwig 2007) – we need nothing more than our own intuitions about particular cases in order to determine what intuitions people would (or should) have about those cases. Proponents of both experimental analysis and experimental descriptivism think that we would do better to actually empirically ascertain what intuitions people have about those cases.

While the taxonomy provided by Nadelhoffer and Nahmias is a significant move towards understanding the contours of experimental philosophy's positive program, we need to refine that initial taxonomy somewhat, so that it will be clearer which of our concerns will apply most squarely to which positive program arguments.

The most fundamental question for any program of positive experimental philosophy is whether the ultimate philosophical payoff is meant to be *mentalist* or *extramentalist*. This distinction is originally drawn by Alvin Goldman and Joel Pust (1998): "views about philosophical analysis may be divided into those that take the targets of such analysis to be in-the-head psychological entities versus outside-the-head nonpsychological entities. We shall call the first type of position *mentalism* and the second *extra-mentalism*" (184). Goldman and Pust introduced the distinction in terms of possible rationales for armchair deployments of intuition, but it applies equally well here: if the experimental results of positive experimental philosophy are meant to tell us something of philosophical import, what type of thing is it supposed to be? Where any given account stands in terms of this distinction will determine the most basic theoretical burdens that it must shoulder.

Mentalist positive experimental philosophy can further be divided between what

we will call *conceptualist* and *mechanist* approaches. Perhaps we are interested in what the actual conceptual structure it is that is instantiated in people's heads, for various concepts of philosophical interest, such as INTENTIONAL (Knobe 2003) or INNATE (Machery, Griffiths, and Lindquist ms). Alternatively, we may be interested more in the nature of certain processes, and in answering philosophical questions about them. Can our folk psychology be understood in primarily prediction-and-explanation terms, or is it deeply entwined with our moral and evaluative cognition as well? (Knobe 2003, 2007b) To what extent do affect and rules contribute to the difference between normative evaluations that are moral and those that are not? (Nichols and Mallon 2006; Mallon and Nichols, forthcoming).

On a mentalist approach, what we learn about are mental entities, and the basic relevance of the experiments may be clear enough since empirical investigation of intuitions can tell us things about the mind. And we do not doubt for a moment that many questions about the mind's structure, contents, and operations will count unproblematically as philosophical (Knobe 2007a). For example, questions about the nature of moral judgment or the relation between the qualitative and representational aspects of perceptual experience are paradigmatically philosophical. On the other hand, it is simply not the case that *all* philosophical questions can be obviously and directly answerable by experimental psychological methods. For example, while it's easy to imagine experiments that illuminate the concepts of freedom and responsibility involved in concerns about free will, it is difficult to imagine how any experiment that could directly illuminate the nature of free will. And while people's intuitions concerning the application of "knows" can tell us at best what people *take* knowledge to be, this may

leave untouched the question of what knowledge *really* is. Thus, while mentalist programs are vindicated by the relevance of empirical evidence about intuitions to conclusions about our mental workings, they can fall short of establishing a host of more ambitious philosophical aims.

Extramentalism thus becomes tempting, since there are so many philosophical questions that we want answered that go beyond the psychological. But how is it that experiments can tell us about matters outside the mind? This is the pressing question for extramentalism, and extramentalist forms of positive experimental philosophy can be further articulated in terms of their main strategies for answering this challenge. We will distinguish between *direct* and *indirect* extramentalist arguments, depending on whether or not claims about the workings of human psychology play an intermediate role in establishing the extramental claims.

Direct extramentalist claims are those that draw conclusions about nonmental entities from premises that include empirical claims about folk intuitions or judgments but do not include premises about human psychology arrived at by via those empirical claims. For example, one might take it that philosophical positions that are intuitive to a large majority of ordinary people, and that are not matters of technical expertise, should be given a significant default positive epistemic status. So, if most folks are intuitive compatibilists, then incompatibilists should have the burden of proof in debates over free will – and vice versa. A practitioner of this version of direct extramentalist positive experimental philosophy might hope that proper survey work could then uncover which of those views has that argumentative burden (Nahmias, Morris, Nadelhoffer, and Turner 2006).

With *indirect extramentalism*, on the other hand, the experimental work is meant *first* to reveal to us important facts about our underlying psychology, and only then can some further inferential story be told about how those facts can help shed light on the extramental philosophical facts of interest. There seem to be two strains of indirect extramentalism currently extant. In *conceptualist* versions, the experiments provide evidence for claims about the structure of some concepts of philosophical interest, and those claims then serve as premises in some further philosophical argument.[6] Following the program of philosophical analysis of folk concepts that runs through David Lewis (1970, 1972) and Frank Jackson (1998), some positive experimental philosophers have insisted that the evidence that some proposition is a folk platitude be empirically supported – or at least empirically scrutinized (see, e.g., Glasgow 2008, Ulatowski 2008).

Another indirect extramentalist strategy is to take an area of philosophy in which we have had conflicting intuitions, and deploy a psychological theory of those intuitions' production in order to help referee which should be trusted, and which merely explained away (e.g., Greene 2003; Nichols 2006).

3. The Pitfalls of Positive Experimental Philosophy

A. The Empirical Challenge from Negative Experimental Philosophy

Positive programs reject the view that it is appropriate to determine what intuitions are (or would be) generated in response to particular hypothetical cases by determining what our intuitions are about those cases. While this marks a significant turn away from traditional analytic philosophy, positive programs continue to share with more

[6] The repeated use of "conceptualism" for both a form of mentalism and a form of extramentalism may seem to risk confusion, but this is not so, as both will fall prey to the same worries in section 3.B. below.

traditional analytic philosophy a number of commitments. Among these are: that intuitions are an important source of philosophical evidence for (or against) philosophical theories or data both about the nature of our folk philosophical concepts and judgments and about the nature of the domains in which we make those judgments; that intuitions are a trustworthy source of evidence or data; and that intuitions about a particular hypothetical case will, by and large, be shared.

But, recent empirical work conducted by philosophers and psychologists gives us reason to worry that philosophical intuitions might be neither trustworthy nor shared. A series of recent empirical studies suggest that some particularly prominent, and commonly appealed to, philosophical intuitions are sensitive to facts about who is considering the hypothetical case[7], the presence or absence of certain kinds of content[8], or the context in which the hypothetical case is being considered.[9] This sensitivity is problematic because such facts have not traditionally been thought to be relevant to the truth or falsity of the claims for which philosophical intuitions are supposed to provide evidence or data. Since evidence is trustworthy and data valuable only to the extent to which their sensitivity is limited to those things relevant to the truth or falsity of the claims for which they are supposed to provide evidence or data, these results call into question the trustworthiness of these intuitions.

Additionally, when these studies are coupled with our inability to either explain what it is about any of these intuitions that make them problematically sensitive or

[7] J. Weinberg, S. Nichols, and S. Stich (2001) and E. Machery, R. Mallon, S. Nichols and S. Stich (2004).

[8] S. Nichols and J. Knobe (2007) and Pizzaro *et al.* (manuscript).

[9] S. Swain, J. Alexander, and J. Weinberg (2008) and L. Petrinovich and P. O'Neill (1996).

predict which other intuitions may or may not be problematically sensitive, they challenge the trustworthiness, not just of the class of intuitions that have so far been studied, but of the whole class of philosophical intuitions. (Alexander and Weinberg (2007), Weinberg (2007)) Just as these recent empirical studies call into question the trustworthiness of philosophical intuitions, they also call into question whether there is, in fact, something like a shared intuition about a particular hypothetical case that can be appealed to either as evidence or data. These studies show that particular hypothetical cases can give rise to a number of different intuitions, thereby calling into question any claims as to what *the* folk intuitions are – a significant problem for positive programs, each of which views getting at *the* folk intuitions to be either a significant philosophical insight in its own right or a necessary step towards achieving a significant philosophical insight. It also raises the question of how we should proceed when confronted with conflicting intuitions. At a bare minimum, anyone who wants to select one from among those intuitions that are generated in response to a given hypothetical case needs to explain why the other intuitions should be discounted. The trouble is that determining just what to do when confronted with conflicting evidence or data is not especially straightforward - as the growing literature in the epistemology of disagreement demonstrates (see, e.g., Christensen (2007), Elga (forthcoming), Feldman (2006), Feldman and Warfield (2007), Kelly (2005, 2007, and forthcoming), White (2005)).

These findings thus pose a clear challenge to direct extramentalist positive experimental philosophy, inasmuch as that sort of project attempts to deploy the intuitions themselves as evidence for philosophical claims. We think that direct extramentalism may be *almost* as imperiled by negative experimental philosophy as are

the armchair methods themselves. One can see the various other positive experimental philosophy projects as making all making epistemologically more modest uses of their findings, by tying them in one way or another to claims about the human mind. But as we shall see, this modesty will not be sufficient to allow other extant forms of positive experimental philosophy to evade the critiques of their negative cousins.

B. The Quine-Machery Problem

Conceptualist approaches to positive experimental philosophy have proved popular, and are perhaps the most common sort of experimental philosophy today. Although conceptualists have not been motivated by the worries we reviewed in the previous section, nonetheless conceptualism seems to hold out the *prima facie* promise of some resources that would make it better able to withstand the challenge from negative experimental philosophy. First, conceptualism allows for a modicum of relativization, which may go some way towards defusing the threat of cross-group differences in intuitions – if Asian and Western subjects have different intuitions, then perhaps they just have different concepts (though see Mallon et al. forthcoming). Relatedly, one may hope under conceptualism to be able to disregard some intuition variation and instability as mere noise, not reflective of the underlying "conceptual competence". The two moves are related, as the first one is only possible if the second one can enable us to distinguish conceptually-based differences in intuition from non-conceptually-based differences.

Conceptualist positive programs thus rely on the idea that we can use empirical evidence to establish what is and what is not constitutive of a given concept of

philosophical interest. But it is, in fact, far from clear how to do so, a point made famous by Quine's "Two Dogmas of Empiricism" (1951). On the one hand, Quine noted, it is not at all clear in virtue of what facts statements about meanings are true, and on the other, it is clear that statements do not entail, as part of their meanings, commitments to particular observations being one way or another. Applying these concerns to experimental philosophy, it is useful to think about how the data regarding the usage of particular terms by the folk is supposed to support claims about the meanings of those terms. Taking the second dogma first, we can follow Quine in noting that no isolated set of observations regarding judgments employing a term *t* observed folk judgments employing a term *t* necessitates any claim including *t*, and this includes claims regarding the meaning of *t*. Rather, our statements about the meanings the folk attach to terms, like other empirical claims, "face the tribunal of sense experience not individually but only as a corporate body" (41). Given any set of survey studies, for example, it always remains open to hold that, because of some distorting factor, the actual folk judgments employing *t* do not really reflect the folk meaning of *t* but represent some other force: for example, theoretical conviction, pragmatic considerations, or confusions or biases.

Of course, this sort of epistemological holism faces any empirical project, but it does not necessarily undercut it. Rather, we engage in inference to the best explanation, given the totality of our background beliefs and evidence, and we gather additional evidence that attempts to decide between the most plausible hypotheses. But in the case of hypotheses about word meanings, this move is not at all straightforward. Simply put, we have no idea to which facts claims about meaning are responsible, and so we have no method of empirically resolving competing hypotheses about meaning. Indeed, the idea

that there are such facts about meaning is empiricism's first dogma.

This sort of point has been made pointedly in the context of experimental philosophy by Edouard Machery (2007) in his discussion of the debate over the so-called "Knobe-effect" or "side-effect effect." In probably the most famous finding of experimental philosophy, Joshua Knobe showed that whether or not a foreseen side effect is judged to be intentional is influenced by whether or not the side effect is bad. Knobe presented subjects with two versions of the following vignette:

> **Harm Condition**
>
> 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*.

In the second, Help Condition, the vignette was the same, except the word "harm" was replaced with "help." In each case, subjects were then asked whether or not the chairman harmed/helped the environment intentionally. But the conditions produced sharply divergent results. Most subjects in the Harm Condition (82%) said the chairman harmed the environment intentionally, while most in the Help Condition said the chairman did not help the environment intentionally. Knobe (2003a) concluded that this asymmetry was not a mistake by subjects, but rather reflected subject's competence with the concept *intentionally*, but other commentators have disputed this, alleging that the effect emerges from considerations extrinsic to the concept, for example, a desire to blame the perpetrator of foreseen harm (see, e.g., Nadelhoffer 2004a, b; Adams and Steadman xx). Machery, however, rightly pointed out is that the debate seems to

hinge upon the appropriate individuation of the concept *intentionally,* and that there is simply no way to resolve this debate absent some idea of how to individuate concepts. And there is absolutely no reason to think that such an idea is forthcoming. This is Quine's point about meaning writ across the landscape of contemporary philosophy of psychology.

It is worth considering, however, whether a more sophisticated psychological inquiry might solve this problem however. For example, "theory" theorist psychologists aim to discern deep principles of "core knowledge" or commitment that might pull apart confounding factors, perhaps revealing the semantic structure of ordinary concepts. For example, there is now a widespread literature discussing folk essentialist construals of various natural kind concepts (e.g. S.A. Gelman 2003). Given these important research programs it may seem presumptuous, and downright *unQuinean*, to try to use a philosopher's armchair argument to attack a scientist's way of arguing.

Still, such an objection assumes that the psychologists' projects and the philosophers' projects are the same. If, instead, psychologists are typically just trying to map out what the psychological structures and processes are that implement our abilities to categorize, with no attendant commitments to any aspect of those structures being meaning-constitutive, the problem doesn't arise.

C. Competence, Performance, Marr, and The Limits of Surveys

But such psychological projects can seem to offer the possibility of significant philosophical payoffs, and both mechanist and indirect extramentalist varieties of positive

experimental philosophy can try to extract them, even while perhaps ducking the challenges that face conceptualist varieties. For example, even as negative experimental philosophy has frequently demonstrated unexpected and unwanted variation in people's intuitions, this observed variation in intuition would no longer pose a problem if we possessed a means for discerning epistemic wheat from chaff. Shaun Nichols and Joshua Knobe (2007) have attempted to do just that, with regard to divergent intuitions concerning free will and determinism, by trying to argue that some of the observed variation is a matter of *performance errors* in one of the studied conditions.

To see how this might work, consider an example from linguistics. Linguists use intuitions about grammaticality as data to construct the grammar of a natural language, but they distinguish between the *competence* involved in producing judgments from the factors that influence *performance.* As Robert Cummins puts it: "competence is ideal... performance, that is, the performance that the system would exhibit but for resource limitations, physical breakdown, and interference from other processes" (Cummins (1996), p. 44). So, in one famous instance, subjects find multiply-center-embedded English sentences like

The man the boy the woman saw heard left.

to be ungrammatical, but linguistic theory says that they comply with the syntax with which we work (viz. they are grammatical according to our competence). The apparent ungrammaticality of these sentences is often explained away in terms of limits on working memory in the parser (e.g., Marcus (1980)). Thus, we take all the evidence we have and construct a model of the cognitive mechanisms that operate to produce judgments in a domain, and we determine the borders of a folk domain by looking at the

mechanism in the model that produces the paradigmatic judgments we are concerned with. The workings of that mechanism determine the competence of the subject within that folk domain. Judgments that are influenced by factors outside that mechanism represent performance errors.

Although we are not troubled by many of the deployments of the competence/performance distinction throughout cognitive science,[10] we do not think that it is a distinction that can – yet – do the work that some positive experimental philosophy practitioners have hoped to have it do. Simply put, experimental philosophy currently lacks the experimental and theoretical resources to make a good use of that distinction for its purposes.

First, experimental philosophy in general has mostly made use of, and continues to deploy, survey methods. Subjects are given a questionnaire, and their judgments are elicited regarding some range of scenarios, with the experimenters typically manipulating the substance of the scenarios but also possibly their order or other contextual elements. Such methods can generate a set of extensional data: given scenario x under conditions y, a certain percentage of subjects give answer z. Such data can at best operate only at the first of Marr's (1982) three levels, the theory of computation -- that is, an input/output account of what function the system computes. However, explanations in terms of performance error most plausibly operate at either the second or third of Marr's three levels of explanation – the level of the algorithm, and the level of the physical implementation.

[10] In addition to its original home in linguistics, the distinction has also done important work in other parts of cognitive science, e.g., in the developmental folk psychology literature (Surian & Leslie 1999; Bloom & German 2000; Scholl & Leslie 2001).

This distinction thus depends on a construal of the actual workings of the system in question.[11] One cannot separate competence from performance with only input/output data, but rather one requires, at least in the background, some sort of account as to what the idealized operation of the system is supposed to be like, such that performance errors can be explained away in terms of the system falling short of that idealization in some way. In the absence of any processing or physical accounts, we just cannot know how the requisite idealization is supposed to go. Such explanations can only succeed, though, given a reasonably clear idea of what resources are being strained, and preferably also how that resource might be limited in the first place. This is why the standard performance error account of center embeddings works so well – short-term memory has a pretty good track record in both regards, as revealed in the general popularity of "cognitive load" as an experimental manipulation. But these are not questions at the level of which inputs produce which outputs, for they require some story about the inner workings of the system. They are therefore not the sorts of questions that can be addressed via survey methods.

It may be easier for positive experimental philosophy to apply a competence/performance distinction in terms of one process interfering with another, and this is indeed what we see with Nichols & Knobe (2007). [12]

Nichols & Knobe explored two different factors that could influence subjects' willingness to attribute the possibility of moral responsibility in a hypothetical case: (i) whether or not the case was described as being in a deterministic universe or an

[11] It also presupposes at a minimum that it will be possible to decompose the relevant cognition into mechanisms with individually discernible functions. We note this commitment without taking issue with it here.

[12] To our knowledge, no one has explored "physical breakdown" as a candidate source of performance errors in XPhi.

indeterministic one, and (ii) whether or not the case was affectively engaging. (We grant here for the sake of discussion that they have correctly characterized the way their experimental materials map into these distinctions.) Unsurprisingly, they found that subjects were generally more willing to attribute the possibility of responsibility in indeterministic universes than in deterministic ones. Perhaps more unexpectedly, they found a similar increased willingness in high-affect cases over low-affect ones. So in a low-affect, deterministic case only a minority of subjects (23%) judged moral responsibility to be possible. A slender majority of subjects judged moral responsibility possible in the high-affect, deterministic case (64%), with more robust majorities in the low-affect, indeterministic case (89%) and most of all in the high-affect, indeterministic case (95%).

So we have some diversity of intuitions across different cases here. Philosophers may not worry as to whether the determinism/indeterminism differences track a real difference in subjects' competence in attributing agency, but one may worry about what to make of the influence of affect on these judgments. Nichols & Knobe consider two possible interpretations. On the "affective competence" interpretation, our emotions are properly part of our agency-attribution system, and their tendency towards compatabilism (as revealed by the majority response in the high-affect, deterministic case) thus reveals a real commitment of our psychology of responsibility. On the "affective performance error" interpretation, our emotions interfere with the more properly incompatabilist judgments of agency. Here's how they argue for the latter interpretation:

> We think that the affective performance error model provides quite a plausible explanation of our results. What we see in the [low affect] case is that, when affect is minimized, people give dramatically different answers depending on whether the agent is in a determinist or

> indeterminist universe. On the performance error hypothesis, these responses reveal the genuine competence with responsibility attribution, for in the low affect cases, the affective bias is minimized. When high affect is introduced… the normal competence with responsibility attribution is skewed by the emotions; that explains why there is such a large difference between the high and low affect cases in the determinist conditions.
>
> Now let's turn to the affective competence account. It's much less clear that the affective competence theorist has a good explanation of the results. In particular it seems difficult to see how the affective competence account can explain why responses to the low-affect case drop precipitously in the determinist condition, since this doesn't hold for the high affect case. Perhaps the affective competence theorist could say that low affect cases … fail to trigger our competence with responsibility attribution, and so we should not treat those responses as reflecting our normal competence. But obviously it would take significant work to show that such everyday cases of apparent responsibility attribution don't really count as cases in which we exercise our competence at responsibility attribution. Thus, at first glance, the performance error account provides a better explanation of these results than the affective competence account.

Yet there is a problem here. To describe one process as interfering with another presupposes an individuation of the processes involved, which is again not something that can be done purely with the sort of survey data that Nichols & Knobe have (like almost all practitioners of experimental philosophy). If we already possessed a well-worked out account of the particular mechanisms operating in these domains and their various interactions, then such an account could maybe provide a framework within which such studies could do the required work. But no such account is currently on offer that can help tell us whether the affective influence on people's judgments a component part of, or an extraneous to, the system producing those judgments. In the absence of any individuation of mechanisms at either the algorithmic or the implementational level of explanation, we cannot tell, The question becomes particularly messy for Nichols & Knobe, as they want to opt for a "hybrid" account in which some of the affect is *part* of

the competence, while other parts of it present an interfering factor. We are not arguing about the truth or falsity of that claim, which strikes us as *prima facie* plausible. Rather, we are addressing whether positive experimental philosophy's survey methods are sufficient to establish such claims as true or false, and we are concerned that they cannot.

One way to see this problem more sharply is to consider that the two hypotheses that Nichols & Knobe consider have to compete with a number of other hypotheses, in which some set of the observed performance is produced by one system and some other set by a different, interfering system. They offer one way of carving things up, but the worry that we're articulating here is that it is extremely difficult, given only the sort of data that they have, to preference any one of those ways over other possibilities. For example, a proponent of an affective competence model could suggest that people's answers in the high-affect cases are fine, but some other mechanism interferes with people's judgments in the low-affect cases; perhaps the description of the determinist universe triggers some sort of explanation-detection system, which competes with the responsibility-attribution system, and produces improper interference. Or perhaps there is just one unified mechanism, and the profile of responses they report is simply the result of its computations, and there are no performance errors to be explained away at all! Such a result would be *philosophically* surprising, but there is no *psychological* reason, given only survey data, to rule such a possibility out.

D. The Proper Domain Problem

One way to get around the worries just articulated would be to already possess a mature theory of the computation for the system in question. Although such a theory tells

us what the input-output function is for the system, it does more than that – it tells us what function it is that we should understand the system as computing. If we already had one of those, then we could use it to help referee between at least some competing accounts at the algorithmic level.

Although such an appeal to the theory of the computation is theoretically possible, it will not help here. For we are considering cases in which positive experimental philosophy is supposed to help *referee* between conflicting accounts of a given domain. As such, these are cases in which the fundamental philosophical facts are dialectically up for grabs. To determine whether Nichols & Knobe's subjects' affect is interfering with their judgments, or a manifestation of their competence in those judgments, we need to know first what function it is that their psychological systems are trying to compute. If their system is meant to judge the world along the lines of compatibilism, then the affect would be part of the competence; and if their system is meant to judge the world along the lines of incompatibilism, then the affect would be an interference.

(Note that this worry, perhaps unlike the worry in the previous section, applies more generally than just to survey-based positive experimental philosophy. For example, it applies equally well to Joshua Greene's attempts to argue for utilitarianism based on his neuroimaging studies of subjects considering trolley cases (e.g., Greene *et al*. 2001). We fear that practitioners of positive experimental philosophy have forgotten the extent to which psychologists who appeal to a competence/performance distinction typically do so with very robust theories of both the system in question and its target domain already in play. We are unfortunately nowhere near achieving either of such types of theories in most areas of positive experimental philosophy.)

Practitioners of positive experimental philosophy might thus look around for other sources of evidence that can help determine what function it is that, say, our moral responsibility system is computing. But this will turn out to be difficult to do for there is no reason to think, even conceding that the mind is comprised of systems discrete enough to be assigned separate proper domains, that such systems will correspond closely enough with domains of philosophical interest to provide acceptable reductions of them. And to lose track of this fact is to distort both the philosophy and the psychology.

Philosophers are typically interested in domains that correspond to philosophical concepts of interest, for example, *responsibility*, *morality*, *knowledge*, and so forth. But there is no reason to think that these domains neatly align with, rather than cross-cut, our cognitive architecture. To see this, consider recent claims of a "linguistic analogy" for the moral domain – claims that moral cognition is underwritten by a domain-specific adaptation for morality, on a par with the linguistic faculty posited by Chomsky (e.g. Dwyer 1999, Harman 1999, Mikhail 2000, Hauser et al. 2007).

So this might offer us one way of thinking about proper functions, in terms of evolution: the proper function of a mechanism is the function for which the mechanism was or is selected for, the function that computes solutions to problems in its *proper domain*. While there could be other ways of specifying what exactly determines or constitutes the proper function of a mechanism, the main point is that such proper functions may either be directed at problem domains that are more or less neatly coextensive with philosophical domain of interest like morality or they may not. Indeed, some proponents of a moral faculty are quite clear about this (see, e.g., Hauser et al. 2007): they mean to make the nontrivial claim that there is a species-typical, innate, and

domain-specific mechanism whose proper function is moral judgment, and they understand such proper functions in terms of evolutionary pressures to solve problems within the moral domain.

But when we start thinking about natural selection, it suggests pressures that lead in quite different directions than moral reasoning. To choose a quick and straightforward example, it seems like sound evolutionary logic to think that evolution might favor mechanisms that systematically favor members of one's own family, or members of one's own group, or person's that might assist one's reproductive success, while it is at least arguably correct to think that these considerations are not *morally* relevant. At the very least, one cannot simply assume that the domain determined by the proper function of the mechanism that underwrites morality and the domain of morality are coextensive.[13]

Consider how losing track of the distinction between the domain of morality and the proper domain of the mechanism that implements morality can distort psychological inquiry. Much recent work in moral psychology, including work in the "linguistic analogy" tradition, has engaged in relatively straightforward appropriation of the philosophical technique of eliciting moral intuitions by presenting carefully constructed moral dilemmas (e.g. Mikhail 2000, Hauser et al. 2007). This appropriation, however, makes the substantial assumption that the kind of data of relevance to philosophers will also be relevant to psychologists. Notice first that philosophers carefully construct moral dilemmas as ways of eliciting intuitions of relevance to assessing competing theories of morality. For this reason, philosophers' dilemmas

[13] For further elaboration on this point, see Mallon (2007).

typically *exclude* factors that are widely considered to be morally irrelevant. For example, variations of the famous "runaway trolley" typically aim to probe the circumstances (if any) in which it is okay to bring about the death of a person in order to save five others. But such variations typically do not include versions in which we are asked to weigh the lives of relatives, out-group members, potential sex partners, and so forth, and (to repeat), this is precisely because whatever difference such factors might make are irrelevant to the moral questions at hand. But, there's every reason to think that these factors *are evolutionarily relevant* and so considering them in computing moral judgments may be part of the proper function of whatever mechanism or mechanisms that underlie moral judgment.

When we deviate from this assumption and consider different factors, we may find different answers – answers that suggest a very different sort of faculty than a "moral" faculty is at work. In one of the earliest experimental investigations of trolley dilemmas, Petrinovich, O'Neill, and Jorgensen (1993) report finding that subjects prefer the lives of relatives and friends over strangers in standard trolley scenarios, a finding they take to support socio-biologists' and evolutionary psychologists' suggestions that humans are designed, in part, to be concerned with their own inclusive fitness. Suppose that this data from Petrivonich et al. is correct. Suppose further that much of our moral judgment is underwritten by an evolutionarily designed mechanism M that computes using an internalized principle like:

> (K) The wrongness of an action resulting in an avoidable death is inversely proportional to the subject's relatedness to me.

Such principle may well be morally irrelevant, but it may well be relevant to the

operation of the faculty that underlies moral judgments about trolley cases. Hauser et al. indicate that in contrast with such research that focuses on questions of "evolutionary significance," their research will probe "the computational operations that drive our judgments" (2007, p. 127). But this begs a crucial question, viz. whether the computational process driving our typical moral judgments are themselves biased by evolution in ways that are at odds with our intuitive sense of morality.[14] It is possible that our concept of morality may emerge only when an innate, domain specific mechanism is used in ways that are at odds with its design (e.g. when it is not allowed access to information such as the relatedness of a person to us).

If philosophical and psychological boundaries needn't be even approximately isomorphic, then it goes to show inquiries into those borders are relatively autonomous. One cannot read the borders of philosophically interesting domains off of the psychology, and one cannot read psychological borders off of the philosophy.

4. Conclusion & Prospects for the Future

Let us recap how the worries we have raised here can be seen to afflict the varieties of positive experimental philosophy articulated in section 2 above. All parties face a challenge of figuring out what to do with the negative experimental philosophical evidence of various sorts of unwanted variability in people's philosophical intuitions. Direct extramentalists owe us the same kind of story that traditional armchair philosophers owe us: how do we discern which intuitions count? When different folks

[14] Hauser et al. make this same move more explicitly when they exclude gender as a relevant explanatory dimension, writing that "we find it clear that some distinctions [e.g., the agent's gender] do not carry any explanatory weight" (2007, p. 131). Here again, they make judgments that reflect a judgment about what sort of considerations are properly considered moral ones. But there seems little reason to think evolution would have respected such niceties in constructing us, so it is not clear why such exclusions are relevant to our underlying functional organization.

yield different intuitions, which one do we take to be likely to be tracking the philosophical truth?

The difficulties faced by direct extramentalism in this regard can make some sort of conceptualism attractive. One resource that conceptualism offers is the legitimacy of some degree of relativism: maybe people with different intuitions just have different concepts, so everyone is still correct. (Here we see a clear case where positive experimental philosophy has an advantage over traditional armchair methods: the latter has no capacity to discern such demographic differences in intuitions, and hence concepts, whereas the former can commission as much cross-cultural research as is needed.) Another resource that conceptualism offers is some means for explaining *away* some of the variation, as due to factors other than the meaning-constitutive elements of the concepts themselves. But that entangles them in Quinean difficulties that they have not (and, we think, cannot) resolve.

Finally, in order to address the problem of unwanted variability, mechanist versions of mentalism or indirect extramentalism seem to require *either* a well worked-out architecture of the different systems involved, *or* some way of fixing the proper domain of those systems. Given the kinds of survey data typically gathered in current positive experimental philosophy research, though, and the problem of philosophical domains cross-cutting psychological domains, mechanist approaches seem to lack both the experimental and theoretical tools needed to advance their programs at this time.

At this point, the prognosis might seem rather grim for experimental philosophy's positive programs. Interestingly, its salvation might ultimately rest on its ability to become *more* experimental – or at least more like experimental psychology. The kinds of

survey methods that experimental philosophers so frequently employ play little role in experimental psychology – and, for good reason: there are better methods available to answer the kinds of questions that are of interest to both experimental psychologists and philosophers (Scholl (2007)). Experimental philosophy's positive programs would do well, we think, to become more like experimental psychology. The hope of experimental philosophy's positive program was to use science to help do some of the work that traditional philosophy hasn't been able to do (or hasn't been interested in doing). Part of the challenge facing the positive programs is to become more scientifically sophisticated. But doing more, and better, *science* will not be enough by itself to fully met the challenge – the positive programs must also do enough *philosophy* to see how to bridge the gap from empirical findings to philosophical payoffs.

Bibliography

Adams, F. and A. Steadman (2004), "Intentional Action in Ordinary Language: Core Concept or Pragmatic Understanding?", *Analysis*, Vol. 64, pp. 173-181.

Adams, F. and A. Steadman (2004b), "Intentional Action and Moral Considerations:

Still Pragmatic", *Analysis,* Vol. 64, pp. 268-276.

Alexander, J. and J. Weinberg (2007), "Analytic Epistemology and Experimental Philosophy", *Philosophy Compass,* Vol. 2(1), pp. 56-80.

Bloom, P. and German, T.P. (2000), "Two reasons to abandon the false belief task as a test of theory of mind", *Cognition,* 77, B25-B31.

Christensen, D. (2007), "Epistemology of Disagreement: The Good News", *The Philosophical Review,* Vol. 116(2), pp. 187-217.

Cummins, R. (1998), "Reflection on Reflective Equilibrium", in M. DePaul and W. Ramsey (eds), *Rethinking Intuition,* (Rowman and Littlefield Press), pp. 113-128.

Dwyer, S. (1999), "Moral Competence", in K. Murasugi and R. Stanton (eds), *Philosophy and Linguistics*, (Westview Press).

Elga, A. (2006), "Reflection and Disagreement", *Nous*, Vol. 41(3), pp. 478-502.

Feldman, R. (2006), "Epistemological Puzzles About Disagreement", in S. Heatherington (ed), *Epistemology Futures,* (Oxford University Press).

Feldman, R., and F. Warfield (2007), *Disagreement,* (Oxford University Press)

Gelman, S. A. (2003), *The essential child: Origins of essentialism in everyday thought,* (Oxford University Press).

Glasgow, J. (2008). "On the Methodology of the Race Debate: Conceptual Analysis and Racial Discourse" *Philosophy and Phenomenological Research,* Volume 76 Issue 2, Pages 333 – 358.

Goldman A. and J. Pust (1998), "Philosophical Theory and Intuitional Evidence", in *Rethinking Intuition,* ed. M. DePaul and W. Ramsey, (Rowman and Littlefield Press), pp. 179-200.

Greene, J.D., R.B. Sommerville, L.E. Nystrom, J.M. Darley, and J.D. Cohen (2001), "An fMRI Investigation of Emotional Engagement in Moral Judgment", *Science,* Vol. 293, pp. 2105-2108.

Greene, J.D. (2003) "From neural "is" to moral "ought": what are the moral implications of neuroscientific moral psychology?", *Nature Reviews Neuroscience,* Vol. 4, 847-850.

Harman, G. (1999), "Moral philosophy and linguistics", in K. Brinkmann (ed),

Proceedings of the 20th World Congress of Philosophy, vol. I: Ethics, (Bowling Green, OH, Philosophy Documentation Center): pp. 107-115.

Hauser, M.D., L. Young and F. Cushman (2007), "Reviving Rawls' Linguistic Analogy: Operative principles and the causal structure of moral actions" in W. Sinnott-Armstrong (ed), *Moral Psychology, Volume 1: The Evolution of Morality: Adaptations and Innateness.*

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

Kelly, S. (2005), "The Epistemic Significance of Disagreement", in J. Hawthorne and T. Gendler Szabo (eds.), *Oxford Studies in Epistemology, Volume* 1 (Oxford University Press): pp. 167-196.

Kelly, S. (2007), "Peer Disagreement and Higher Order Evidence", in R. Feldman and T. Warfield (eds), *Disagreement*, (Oxford University Press)

Kelly, S. (forthcoming), "Disagreement, Dogmatism, and Belief Polarization", *The Journal of Philosophy*.

Knobe, J. (2003), "Intentional Action and Side Effects in Ordinary Language", *Analysis*, Vol. 63, pp. 190-193.

Knobe, J. (2007a), "Experimental Philosophy and Philosophical Significance", *Philosophical Explorations*, Vol. 10, pp. 119-122.

Knobe, J. (2007b), "Reason Explanation in Folk Psychology", *Midwest Studies in Philosophy*, Vol. 31, pp. 90-107.

Kauppinen, A. (2007), "The Rise and Fall of Experimental Philosophy", *Philosophical Explorations*, Vol. 10(2), pp. 95-118.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy*. 67:426-446.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50, 249-58.

Ludwig, K. (2007), "The Epistemology of Thought Experiments: First Person versus Third Person Approaches", *Midwest Studies in Philosophy*, Vol. 31, pp. 128-159.

Machery, E. (2007), "The folk concept of intentional action: Philosophical and psychological issues", *Mind & Language*, Vol. 23, pp. 165-189.

Machery, E., S. Lindquist and P. Griffiths (manuscript), "The Vernacular Concept of Innateness"

Machery, E., Mallon, R., Nichols, S., and Stich, S. (2004), "Semantics, Cross-cultural style", *Cognition*, Vol. 92, Vol. 3, pp. B1-B12.

Mallon, R., E. Machery, S. Nichols and S. Stich (forthcoming), "Against Arguments from Reference", *Philosophy and Phenomenological Research.*

Marcus, (1980)

Mikhail, J. (2000), "Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A theory of justice'", *Philosophy*, (Cornell University).

Mallon, R. (2007), "Reviving Rawls Inside and Out", in W. Sinnott-Armstrong (ed), *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, (MIT Press), pp. 145-155.

Mallon, R. and S. Nichols (forthcoming), "Moral Reasoning, Moral Rules, and Moral Dilemmas" in J. Doris, S. Nichols and S. Stich (eds), *The Oxford Handbook of Moral Psychology*, (Oxford University Press).

Nadelhoffer, T. (2004a), "On Praise, Side Effects, and Folk Ascriptions of Intentionality", *Journal of Theoretical and Philosophical Psychology*, Vol. 24, pp. 196-213.

Nadelhoffer, T. (2004b), "Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow", *Journal of Theoretical and Philosophical Psychology*, Vol. 24, pp. 259-269.

Nadelhoffer, T. and E. Nahmias (2007), "The Past and Future of Experimental Philosophy", *Philosophical Explorations*, Vol. 10(2), pp. 123-149.

Nahmias, E., S. Morris, T. Nadelhoffer, and J. Turner (2005), "Surveying Freedom: Folk Intuitions about Free Will and Moral Responsibility", *Philosophical Psychology*, Vol. 18(5), pp. 561-584.

Nahmias, E., S. Morris, T. Nadelhoffer, and J. Turner (2006), "Is Incompatibilism Intuitive?", *Philosophy and Phenomenological Research*, Vol. 73(1), pp. 28-53.

Nichols, S. (2006), "Imaginative Blocks and Impossibility: An Essay in Modal

Psychology", in S. Nichols (ed.) *The Architecture of the Imagination* (Oxford: Oxford University Press), pp. 237-255.

Nichols, S. and J. Knobe (2007), "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions", *Nous*, Vol. 41, pp. 663-685.

Nichols, S. and R. Mallon (2006), "Moral Rules and Moral Dilemmas", *Cognition*, Vol. 100, pp. 530-542.

Nichols, S., S. Stich, and J. Weinberg (2003), "Metaskepticism: Meditations in Ethno-Epistemology", in S. Luper (ed), *The Skeptics: Contemporary Debates*, (Ashgate Press).

Petrinovich, L., P. O'Neill, and M. Jorgensen (1993), "An Empirical Study of Moral Intuitions: Toward an Evolutionary Ethics", *Journal of Personality and Social Research*, Vol. 64, pp. 467-478.

Petrinovich, L. and P. O'Neill (1996), "Influence of wording and framing effects on moral intuitions", *Ethology and Sociobiology*, Vol. 17, pp. 145-171.

Pizzaro, D., E. Uhlman, D. Tannenbaum, and P. Ditto (manuscript), "The Motivated Use of Moral Principles".

Quine, W.V.O. (1951), "Two Dogmas of Empiricism", *Philosophical Review*, Vol. 60(1), pp. 20-43.

Scholl, B.J. and Leslie, A.M. (2003), "Minds, Modules, and Meta-Analysis", *Child Development*, 72, 696-701.

Surian,L. and Leslie, A. M. (1999). "Competence and performance in false belief understanding: A comparison of autistic and normal 3-year-old children." *British Journal of Developmental Psychology*, 17, 141-155.

Swain, S., J. Alexander and J. Weinberg (2008), "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp", *Philosophy and Phenomenological Research*, Vol. 76(1), pp. 138-155.

Ulatowski, J. (2008), "How Many Theories of Act Individuation Are There?" Ph.D. Dissertation, Department of Philosophy, University of Utah.

Weinberg, J. (2007), "How to Challenge Intuitions Empirically Without Risking Skepticism", *Midwest Studies in Philosophy*, Vol. 31(1), pp. 318-343

Weinberg, J, S. Nichols and S. Stich, (2001), "Normativity and Epistemic Intuitions", *Philosophical Topics*, Vol. 29, pp. 429-60.

White, R. (2005), "Epistemic Permissiveness", *Philosophical Perspectives*, Vol. 19, pp. 445 – 459.