Experimental Philosophy is Cognitive Science[1]

Joshua Knobe

[For Sytsma, J. & Buckwalter, W. (eds.) (forthcoming).
A Companion to Experimental Philosophy. Blackwell.]

One of the most influential methodological contributions of twentieth century philosophy was the approach known as *conceptual analysis*. Research using this approach yielded numerous specific discoveries about the use of specific concepts, but it also led to an interest in questions at a more metaphilosophical level. Philosophers began asking, 'What exactly does it mean to analyze a concept?' 'How do we know whether a conceptual analysis is correct or incorrect?' 'What implications might conceptual analysis have for questions that are not directly about concepts?' Existing work on these metaphilosophical questions has given us some important insights into the methods and aims of conceptual analysis.

Now, in the twenty-first century, we find the emergence of a new approach known as *experimental philosophy*. The result has been a new series of discoveries about the use of specific concepts, as well as a new series of metaphilosophical questions. Philosophers have begun asking, 'What exactly is experimental philosophy?' 'What is work in this field aiming to achieve?' 'What implications might it have for more traditional philosophical issues?'

Given this background, it seems only natural to try to answer questions about contemporary experimental philosophy by drawing on insights from metaphilosophical work on conceptual analysis. In fact, one might well be tempted to reason as follows:

> It's not as though we have to start all over from scratch. We already know a lot about how to do metaphilosophy. We have developed sophisticated theoretical frameworks, and these frameworks have proven extraordinarily successful in helping us to understand twentieth century conceptual analysis. Of course, experimental philosophy differs in certain ways from previous approaches, but all the same, the best way to proceed at this point is probably just to take some of the key ideas from existing work and do our best to apply them to this new form of philosophical research.

I will argue that this strategy is a misguided one. Experimental philosophy, I will suggest, is deeply different from conceptual analysis. Thus, the frameworks that proved so helpful in making sense of conceptual analysis tend only to distort our understanding when applied to experimental philosophy.

Ideally, the effort to understand experimental philosophy would proceed in exactly the opposite way. We would not start out with any preconceptions inherited from work on conceptual analysis. Instead, we would simply pick up a series of experimental philosophy papers, read them carefully, and try to understand what

[1] I am grateful for comments on an earlier draft from Wesley Buckwalter, Joachim Horvath, Eddy Nahmias, Justin Sytsma and John Turri.

they were doing. Then we would construct theoretical frameworks designed specifically to aid us in this task.

Unfortunately, this ideal method is no longer available to us. We already know a lot about conceptual analysis, and we cannot simply unlearn it. The best option at this point is therefore to take up the problem explicitly. We need to look in detail at the ways in which recent research in experimental philosophy differs from traditional research in conceptual analysis. We will then be in a better position to ask whether certain elements of the theoretical frameworks we have inherited might be getting in the way of our attempts to understand this new type of research.

I

Existing metaphilosophical work has identified two possible experimental research programs that could be helpfully understood using frameworks derived from the conceptual analysis tradition. One is a research program that aims to make a *positive* contribution to conceptual analysis; the other is a research program that aims to engage *negatively* by providing evidence against the methodological assumptions of conceptual analysis itself. Work in metaphilosophy has carefully spelled out the key features of these possible research programs and has rigorously explored the philosophical merits of each (Alexander, Mallon & Weinberg, 2010; Alexander & Weinberg, 2007; Kauppinen, 2007; Ludwig, 2007; Sosa, 2007).

The one worry I have about this work is that it seems a bit disconnected from the goals of most actual empirical research in this area. Even a casual glance at recent work in the field would show that the overwhelming majority of the actual empirical studies do not fit neatly into either of these two research programs. Thus, if we debate the philosophical merits of these research programs, we may be learning something of value, but we will not be learning about the merits of the sort of empirical work that most experimental philosophers are actually carrying out.

To get a better sense of what experimental philosophers have actually been doing, Ike Silver and I conducted a simple quantitative analysis. The first step was to put together a dataset of empirical studies conducted by experimental philosophers. To do this, we turned to the PhilPapers database. Silver went through the database and examined all of the papers listed in the category 'Experimental Philosophy' over the past five years. In total, there were 379 papers. He then extracted from these papers all of the actual empirical studies. This method yielded a dataset of 453 studies (Silver, 2014). Once this dataset had been assembled, I classified each of the studies with regard to whether it was presented as participating in one of the two research programs described above.

Some studies are indeed presented as evidence for positive accounts that follow at least broadly in the tradition of conceptual analysis. In such cases, experimental philosophers defend an analysis of a particular concept, and they use empirical results as part of that defense. We can now ask how large a role this sort of work has played in recent research in experimental philosophy. At times, it can be a bit difficult to determine whether a given experiment is best construed as being offered in defense of a conceptual analysis, but even on a very liberal understanding, attempts to defend a specific conceptual analysis account for just 10.4% of the studies reported over the past five years.

Similarly, some studies are presented, in a more negative way, as providing reason to reject the whole tradition of armchair conceptual analysis. Such studies do not merely give us evidence against one or another specific claim made by practitioners of conceptual analysis; they are supposed to provide evidence against the basic methods of conceptual analysis itself (e.g., by showing that people's intuitions are fundamentally unreliable). The question now is how much of the actual empirical work is contributing to this project. The answer is that it accounts for 1.1% of the studies reported over the past five years.

To sum up, existing metaphilosophical work has focused on two possible research programs in experimental philosophy, but only a small minority of the actual empirical work being done by experimental philosophers falls within these programs. In saying this, I don't at all mean to criticize existing metaphilosophical work. This work has articulated and defended research programs that do in fact exist and are amply worthy of investigation; the point is simply that the vast majority of research in experimental philosophy does not fall neatly into those programs. To the extent that certain philosophers believe otherwise, my sense is that they are being influenced by an *a priori* belief about how work in this area should proceed, rather than by an impartial examination of what actually gets done in published papers in the field.

II

Well then, what are experimental philosophy papers actually doing? My answer should come as no surprise. The majority of experimental philosophy papers are doing *cognitive science*. As such, they are doing precisely the sorts of things one would expect cognitive science papers to do. They are revealing surprising new effects and then offering explanations those effects in terms of certain underlying cognitive processes. If we want to makes sense of this work, the obvious approach would be to look not so much to the frameworks developed in the tradition of conceptual analysis as to the frameworks developed in the tradition of cognitive science.

At this point, however, one might well object that our dismissal of the conceptual analysis framework has been a bit too quick. More specifically, a person could object as follows:

> Yes, it's true that if you just pick up a bunch of experimental philosophy papers and try leafing through them, you will find a lot of material that looks like cognitive science and very little that looks like conceptual analysis. But this is a highly superficial way of exploring the issue. What you really need to do is to look in depth at recent research in experimental philosophy and try to get a better understanding of what this research aims to accomplish. Once you engage in this more careful examination, you will see that there is a deeper sense in which work in experimental philosophy is fundamentally continuous with the conceptual analysis tradition.

What I want to show now is that this objection is mistaken. In fact, I will try to show that the facts of the matter are just the opposite. The more carefully one looks at what contemporary work in experimental philosophy is doing, the more one comes to understand how fundamentally different it is from traditional conceptual analysis.

III

To properly make a case for this claim, we need to begin by introducing a rough characterization of conceptual analysis itself. Of course, the conceptual analysis tradition is a rich and complex one, and it would be impossible to capture all of its nuances in a brief chapter like this one. For present purposes, however, it should suffice just to take note of five salient facts about most existing work in the field.

1. Research in conceptual analysis proceeds in part by appealing to judgments of a certain sort about hypothetical cases. The usual way to describe this method is in terms of 'intuitions.' A great deal of controversy remains about what exactly an intuition is, or even whether that is the best way of describing the relevant sort of judgment, but those issues will play no real role in the discussion that follows. Let us simply put them to one side.

2. Conceptual analysis then uses facts about people's intuitions to arrive at conclusions about *concepts.* This point will prove absolutely central in what follows, and it is worth taking just a moment to discuss it.

First, it should be noted that there is a difference between studying people's intuitions and studying their concepts. People's intuitions are determined in part by their concepts, but people's intuitions are also affected by numerous other factors. For example, people's intuitions about knowledge are determined in part by their concept of knowledge, but they are also affected by people's working memory capacity, by their ability to engage in counterfactual thinking, and so on. (If a person suffers a deficit in her working memory capacity, she might still have a perfectly intact concept of knowledge, but she would no longer arrive at the same intuitions about knowledge in particular cases.)

Second, it is important to distinguish between people's concepts and the things in the world that these concepts are about. For example, it is important to distinguish between people's concept of knowledge (a concept) and knowledge itself (the thing this concept is about). It is not at all surprising that the study of intuitions can help us understand people's concepts, but defenders of conceptual analysis typically try to go beyond this unsurprising claim. Typically, they claim both (a) that the study of people's intuitions can help us understand their concepts and (b) that a proper understanding of these concepts can show us something important concerning the things the concepts are about.

One of the most salient aspects of the program of conceptual analysis is this idea that the study of people's intuitions can somehow provide us with an understanding of real things in the world, and for obvious reasons, this aspect of conceptual analysis has been a major focus within existing metaphilosophical work. I should emphasize, however, that it will not be my focus here. Rather, my focus will be on the more straightforward point that conceptual analysis involves the study of concepts.

3. The aim of research in conceptual analysis is to develop a specific kind of account of a concept. An account of the relevant kind is usually referred to as an 'analysis.' To give just one example, here is an early attempt to provide an analysis of the concept of knowledge:

> A person knows that p if and only if:
>
> 1. The person believes that p
> 2. p is true
> 3. The person is justified in believing that p

In this particular example, the analysis consists of a list of conditions that are alleged to be individually necessary and jointly sufficient, but philosophers have often proposed accounts of concepts that take some other form. I will count all of these accounts equally as conceptual analyses.

4. As researchers gain an ever deeper understanding of the relevant intuitions, their analyses tend to become ever more complex. The basic trajectory will be familiar to anyone who has participated in this sort of research. A philosopher comes up with a relatively simple analysis that appears to do the trick, and at first, it appears that all is well. But then, inevitably, a problem arises. Someone is able to identify a surprising intuition that shows that the simple analysis isn't quite right. Reacting to this initial difficulty, philosophers set about developing a slightly more complex analysis that is able to handle the counterexample. But to no avail; someone then manages to come up with a counterexample to the more complex analysis, which leads to an even more complex one… until, ultimately, we arrive at an analysis of truly monstrous complexity. Here, for example, is an analysis of the concept of knowledge introduced by Swain (1974):

> *S* knows that *h* iff (i) *h* is true, (ii) *S* is justified [by some evidence *e*] in believing *h*…, (iii) *S* believes that *h* on the basis of his justification and…(iv)…there is an evidence-restricted alternative Fs* to *S*'s epistemic framework Fs such that (i) '*S* is justified in believing that *h*' is epistemically derivable from the other members of the evidence component of Fs* and (ii) there is some subset of members of the evidence component of Fs* such that (a) the members of this subset are also members of the evidence component of Fs and (b) '*S* is justified in believing that *h*' is epistemically derivable from the members of this subset. [Where Fs* is an 'evidence-restricted alternative' to Fs iff (i) For every true proposition *q* such that '*S*is justified in believing not-*q*' is a member of the evidence component of Fs, '*S* is justified in believing *q*' is a member of the evidence component of Fs*, (ii) for some subset C of members of Fs such that C is maximally consistent epistemically with the members generated in (i), every member of C is a member of Fs*, and (iii) no other propositions are members of Fs* except those that are implied epistemically by the members generated in (i) and (ii).]

The example here happens to come from the study of the concept of knowledge, but one finds a quite similar trajectory in work on the concepts of causation, intentional action, and so on.

5. This gradually increasing complexity is widely seen as evidence that something is going seriously wrong. Conceptual analysis was not supposed to

deliver a giant mishmash of clauses and subclauses; it was supposed to capture the relevant intuitions in a theory that displayed a certain elegance or simplicity. (Indeed, the complex analysis of knowledge reproduced above was offered by Lycan, 2006, to show that work on this topic had gone completely off the rails.)

With this brief characterization in the background, we can now return to the topic of experimental philosophy. It is hard to deny that contemporary experimental philosophy resembles conceptual analysis at least in certain superficial respects. Experimental philosophers clearly do study something about people's intuitions. Moreover, they clearly do sometimes draw on whatever it is that they discover regarding intuitions as part of an argument that arrives at conclusions regarding real things in the world. (For example, experimental philosophers clearly do study something about intuitions concerning knowledge, and it is equally clear that they sometimes argue for conclusions regarding knowledge itself.)

For this reason, it may be tempting just to take the entire metaphilosophical framework that has been developed for understanding conceptual analysis and apply it to experimental philosophy. But before we go ahead and do that, we should pause for a moment to look more closely at what experimental philosophers actually do. To begin with, we need to ask ourselves whether it is in fact the case that experimental philosophers are engaged in an attempt to develop analyses of concepts.

IV

On one level, the answer to this question is perfectly obvious. Just try picking out an experimental philosophy paper at random and taking a look at what it says. Almost certainly, you won't find that it makes any attempt at all to develop an analysis of a concept. Instead, you will find something quite different.

Most typically, what you will find is an attempt to identify and explore a specific *effect*. In the paradigmatic case of this sort of work, a researcher is studying people's application of a concept and comes upon some specific pattern in the results that seems highly surprising and counterintuitive. Then other researchers explore this effect further, trying to get at the cognitive processes underlying it. Throughout this whole process, the emphasis is always on one particular effect and its psychological underpinnings; no one ever proposes anything that looks like an analysis of the concept as a whole.

To take just one example, consider an important recent paper by Danks, Rose and Machery (forthcoming). Danks and colleagues show that people actually arrive at different judgments depending on how the relevant information is presented to them. In particular, it makes a great deal of difference whether the information is presented in summary form (as a vignette) or in a more experiential form (through causal learning). The authors demonstrate the existence of this effect in an elegant experimental study, and they make a strong case for the claim that it is showing us something truly fundamental about the way people ordinarily assess causation. But here is the thing. *There is no defense of any general theory about people's causal judgments or about the concept of causation.* One cannot ask whether the authors' overall theory of the concept is complete or incomplete, simple or complex, because no such theory is ever presented. The entire paper is about this one specific effect.

With this point in mind, we can return to an observation that might have seemed puzzling or mysterious when we first introduced it. We noted above that only 10.4% of the empirical studies conducted over the past five years are offered in defense of an analysis of a concept. This fact might at first have seemed surprising, but the reason is actually quite simple. An enormous percentage of the studies are presented as evidence for some claim about how people use a concept; it's just that almost all of them are presented as evidence for a claim about one specific effect, not about the use of the concept as a whole.

One might think at first that this point is a rather superficial one. After all, one can always imagine a person responding roughly as follows:

> Yes, it's true that each individual paper does not defend a general theory about the concept it investigates. Still, each of these individual papers can be seen as just one part of a larger research program. It is this larger research program that is gradually progressing toward a characterization of the concept as a whole. For example, there might be a whole series of different papers on causal intuitions (each exploring a different effect), but one can see those various papers as together contributing to a larger research program that aims to characterize people's concept of causation more generally.

What I want to suggest now is that this response too is mistaken. Papers in the tradition of conceptual analysis were indeed embedded in a research program of roughly this kind, and it is natural enough to start out with the assumption that experimental philosophy papers must be doing something at least vaguely similar. The one problem with this assumption is that it begins to look highly dubious as soon as one begins examining the progress of actual research programs in the field.

V

The best way to get a sense for actual research programs in experimental philosophy is to look in detail at one particular example. So let's pick out one specific effect and take a look at a few of the hypotheses that have been developed to explain it. We can then ask how those specific hypotheses were incorporated into larger programs of research.

Let us take as our example the effect of moral considerations on intuitions about intentional action. A series of studies have shown that in cases of a certain type (so-called 'side-effect' cases) people are more inclined to regard an agent's behavior as intentional when they see it as morally bad. We can now consider three different hypotheses that were developed to explain that one effect:

- The *deep self* hypothesis posits a process whereby people attribute to the agent a 'deep self' and then check to see whether the agent's actions concord with his or her deeper mental states (Sripada, 2010; Sripada & Konrath, 2011).

- The *blame validation* hypothesis posits a cognitive process whereby people are motivated to shift their beliefs in such a way as to justify attributions of blame (Alicke & Rose, 2010; Nadelhoffer, 2006).

- The *counterfactual* hypothesis posits a process whereby people's moral judgments impact the alternative possibilities they consider when trying to make sense of the agent's actual state (Knobe, 2010).

Looking at these hypotheses, one can easily imagine a whole range of different research programs in which each of them could be embedded. The question we want to address now is about which of those research programs people actually ended up pursuing. That is, after each of these hypotheses appeared, what did researchers actually do to extend or build on them?

Consider first the deep self hypothesis. This hypothesis was originally proposed in a series of papers by Chandra Sripada and colleagues (Sripada, 2010; Sripada & Konrath, 2011). At least in principle, one could imagine how subsequent research might aim to build on the insights from these papers and gradually work toward the construction of a complete theory of intentional action intuitions. But the fact is: neither Sripada nor anyone else has actually tried to do anything of the kind. There has never been any serious attempt to take this work and integrate it into a larger 'theory of intentional action intuitions.' On the contrary, all of the actual research has taken a very different direction.

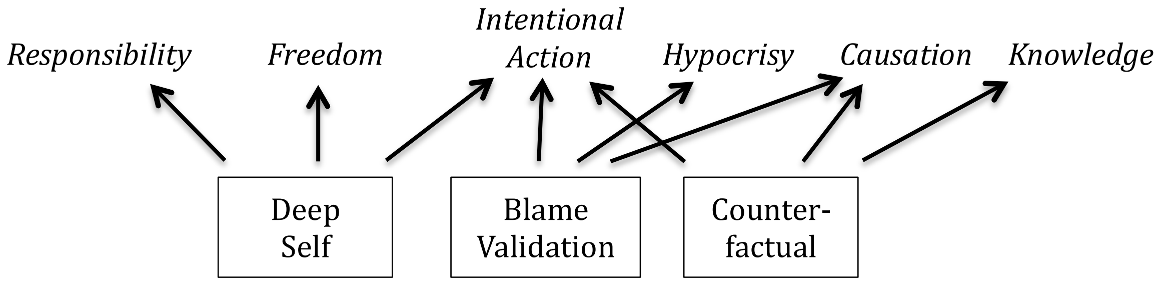
In the years since Sripada's first paper on this topic, he has developed an increasingly refined and sophisticated account of deep self attributions. Then he has taken this account and used it to explain a wealth of other surprising effects, including effects on people's judgments of moral responsibility (Sripada, 2010) and freedom (Sripada, 2012). Spurred on by this work, other researchers have then contributed their own attempts to understand the process of deep self attribution and its impact on various further phenomena (Leben, 2014; Newman, Bloom & Knobe, 2014; but see Rose, Livengood, Sytsma & Machery, 2012). At this point, these papers clearly constitute a dynamic and highly successful research program. However, it is not a program devoted to examining intuitions about intentional action; it is a program devoted to examining the underlying cognitive process of deep self attribution.

Now take the blame validation hypothesis. This hypothesis was initially proposed in a paper by Thomas Nadelhoffer (2006) and, independently, in work by Mark Alicke and David Rose (2010). Looking just at those original papers, one could well imagine, at least in principle, how they could have formed one part of a research program devoted to gradually working out all of the various factors that can influence people's intentional action intuitions. But here again, it turns out that no one has actually been trying to do anything of the kind.

Instead, subsequent work has aimed to more fully explore the nature of the blame validation process and to look at the ways in which this same process could be impacting people's application of various other concepts. Thus, Alicke and Rose argued that blame validation also plays a role in people's application of the concept of causation (Alicke, Rose & Bloom, 2011) and of the concept of hypocrisy (Alicke, Gordon & Rose, 2013). Then Turri and Blouw (forthcoming) extended their original theory to include a notion of 'excuse validation,' which Turri used to examine people's intuitions about assertion, especially insofar as they relate to the concept of knowledge (Turri, 2013). Once again, we see the development of an impressive research program, but in this case too, it isn't a program aimed at understanding people's intuitions about intentional action; it is a program aimed at understanding the underlying cognitive process of blame validation.

My own work has taken more or less the same path. Dean Pettit and I originally proposed the counterfactual hypothesis as a way of understanding a specific effect in people's intentional action intuitions (Pettit & Knobe, 2009). But no one has offered any real suggestions about how this hypothesis could figure in a larger theory of intentional action intuitions. Instead, all of the actual research has been elsewhere. Recent work has seen the development of more formal theories that explain more precisely how the impact of moral judgment on counterfactuals is supposed to work (e.g., Knobe & Szabó, 2013), and a series of studies have used these theories to explain people's application of numerous other concepts, including freedom (Phillips & Knobe, 2009) and causation (Hitchcock & Knobe, 2009). So once again, the real research program that came out of that early paper is not about the concept of intentional action per se; it is about an underlying cognitive process.

The pattern of research discussed here can be illustrated in the following figure:



On the top row, we have a series of concepts; on the bottom row, a series of underlying cognitive processes. The arrows then show cases in which one of the processes has been hypothesized to influence applications of one of the concepts.

The key point now is this: None of these actual research programs aim at systematically investigating the concepts. All of them are investigating the cognitive processes. Predictions about people's applications of the separate concepts then simply fall out of these theories about the underlying cognitive processes.

VI

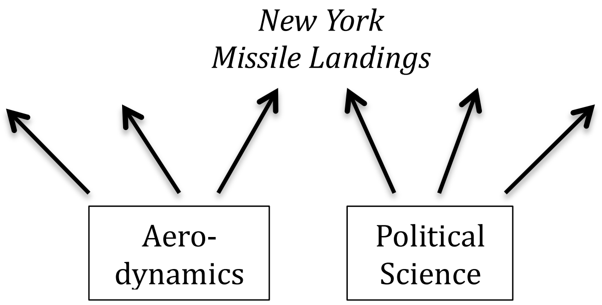
Here again, it might be thought that the claim we are making is merely a superficial one. One can well imagine a person responding:

> Clearly, the research you are describing provides us with information on two different levels. On one level, it provides us with information about underlying cognitive processes; on the other, it provides us with information about the use of individual concepts. Now, you claim that the actual researchers working in this field have been more concerned with the cognitive processes than with the individual concepts. But is that an issue of any real significance? It seems to be more a claim

> about of the sociology of the field than a claim about the substance of what has been learned.

I now want to argue that this response is also deeply mistaken. It is indeed the case that we are gaining information on two different levels, but there is a big difference between merely gaining information about a topic and pursuing a research program devoted to investigating it. The key difference for present purposes is that a proper research program is not just supposed to deliver a hodgepodge of unrelated facts; it is supposed to offer us something with a certain beauty, elegance or simplicity. In short, it is supposed to embody the relevant *theoretical virtues*. There is then a substantive question as to whether we should be trying to develop something that embodies these virtues at any given level.

To see this point more clearly, let us consider an especially extreme case. Suppose we are engaged in a war and it is of crucial importance to us to be able to predict whether our enemy's missiles are going to strike New York City. To do this, we might turn to scientific theories from a diverse array of fields (aerodynamics, political science, etc.). Each of these theories would then also enable us to predict other things that had nothing to do with missiles. So we end up with a structure like this:



In each of the fields at the bottom of this figure, we should presumably be aiming for a full-blown theory that would be expected to display the relevant theoretical virtues. But should we expect the same from our account of New York City missile landings? The obvious answer would be no. This just doesn't seem like the sort of area in which it would be appropriate to demand a proper theory. When it comes to a topic like this one, it would be appropriate to expect something more like a hodgepodge of facts.

We can now apply this same kind of reasoning to work on people's intuitions. It is clear that work in this area should provide us with information about people's use of particular concepts. Still, there is a substantive question, one worthy of serious debate, whether we should be aiming to develop anything worthy of being called 'a theory of causation intuitions,' 'a theory of intentional action intuitions,' 'a theory of knowledge intuitions.' The question is whether this is a level on which proper theory is possible.

Perhaps the best way to address this question is to look at the results of existing efforts to construct such theories. To take just one example, researchers within the conceptual analysis tradition have spent decades studying intuitions about knowledge. We can now ask what sorts of results this work has delivered. There can be little doubt that it has taught us many interesting things about people's intuitions. That is, it has revealed a number of important and very real effects that are amply worthy of further study. Yet, at the same time, there is a widespread feeling that work on this topic has not converged on anything even remotely resembling a 'theory of the concept of knowledge.' One natural response to this outcome would be to conclude that this simply isn't the sort of area in which proper theory is possible. For each of the surprising effects that researchers have uncovered, we should of course be seeking deeper theoretical understanding, but there is no reason to demand at the outset that this understanding must come from a theory that has anything to do with knowledge in particular. It might well come from a theory at some other level.

Be that as it may, it seems that contemporary work in experimental philosophy has not been in the business of constructing theories about the use of individual concepts. Thus, if this work is to display theoretical virtues, it cannot manifest those virtues in precisely the manner familiar from the aspirations of conceptual analysis. Whatever virtues it might embody must be understood in a somewhat different way.

VII

We noted above that theoretical work tends to strive after certain characteristic virtues and, in particular, that it tends to strive after the virtue of *simplicity*. The key question now is how this simplicity is to be understood.

Work in conceptual analysis aimed to develop analyses of concepts, and in that context, the obvious proposal would be that we should be aiming for *simple analyses of concepts*. I have been suggesting that work in experimental philosophy is not engaged in anything like this traditional project. What is needed, then, is a different account of simplicity, one that is more appropriate to the kind of research that experimental philosophers have actually been conducting.

Now, the sort of thing that experimental philosophers most characteristically do is to identify new effects and then explain them in terms of underlying cognitive processes. Thus, if we are to have a conception of simplicity that is appropriate to experimental philosophy, it would have to tell us something about how to achieve simplicity in work of precisely this type.

How then is this notion to be understood? As a first stab, one might suggest that an explanation is simple to the degree that it requires relatively little in the way of assumptions about people's cognition. In other words, we start out with an effect to be explained; then we offer an explanation of it using certain assumptions about people's cognition. The less we need in the way of complex assumptions about cognition, the simpler the resulting explanation.

But it takes only a moment's reflection to see that this first stab at articulating the relevant conception of simplicity is no good. After all, explanations in experimental philosophy frequently draw on theories that have already been extremely well supported by existing research. To give just one example, De Brigard (2010) uncovers an interesting new effect and then offers an explanation of that effect using the resources of prospect theory. Now, prospect theory is a quite complex theory, but it has been supported by decades of existing research in cognitive science. Surely, the fact that De Brigard relies on this highly well-established theory does not constitute a lack of theoretical virtue in his explanation.

So what we need is a slightly different conception of simplicity. Let us say that an explanation for a given effect is simple to the extent that it avoids introducing *additional* assumptions about people's cognition, over and above those that would be needed to explain other effects. On this conception, any explanation has to be understood against the backdrop of a larger account of cognition that is needed to explain effects other than the one under discussion. The explanation of a given effect is simple to the degree that it works without having to introduce additional assumptions for which there is no independent evidence.

To illustrate the basic idea here, we can return to our earlier example: the moral asymmetry in intentional action judgments. One way to explain this effect would be to add in to our total account of cognition, on top of everything we already believe, a principle that simply amounts to a description of the effect to be explained. That is, in addition to everything else we believe, we could add the principle:

> When people regard a side-effect as morally bad, they will conclude that it was brought about intentionally, whereas when people regard a side-effect as morally good, they will conclude that it was brought about unintentionally.

Clearly, this would be a terrible explanation. The trouble is that no aspect of it draws on, or could even be confirmed by, independent evidence of any sort. The whole thing is simply posited ad hoc to explain this one effect.

Now consider an approach at the opposite extreme. Suppose we develop a very general theory about how people make judgments. This theory says that certain judgments are impacted by a process we call *Process X*. Without even looking at people's applications of the concept of intentional action, we refine the theory and work it out in considerable detail, so that we end up with a rich understanding of precisely what Process X involves. Then, once the theory is more or less in place, we just add one further assumption:

> Intentional action judgments are impacted by Process X.

Given what we already know about Process X, and what we already know about intentional action judgments, this one additional assumption leads immediately to a host of new predictions. One of these is that intentional action judgments will show a moral asymmetry.

Now, in a certain sense, this latter explanation would be far more complex than the first one we considered. After all, the explanation relies on a complex general theory of Process X, so if we had to write out the explanation in full, it would include some very complex theoretical material. The thing to keep in mind, however, is that none of this theoretical material is being introduced in order to explain this specific effect. We were already committed to it for independent reasons, and the only thing we need to add was a single further assumption. In this sense, the explanation can be seen as impressively simple.

The progress of experimental philosophy over these past few years has involved a striking movement toward simplicity of this type. As research proceeds, we come to have ever more substantive theories about the underlying cognitive processes. The result is that we need to say ever less about each separate effect. That is, we become able to explain each individual effect without positing much of anything that was introduced for the purpose of explaining it in particular.

As we continue down this path, we are moving toward the ideal of an explanation that is *absolutely simple*. In an explanation of this ideal type, one would pick out a surprising new effect and make sense of it while relying only on assumptions for which there is already independent evidence. Thus, the amount of new theory one would need to add, on top of what was required for independent reasons, would be strictly zero.

VIII

At this point, we need to consider just one final objection. Suppose that someone were to say:

> All right, I accept that experimental philosophy differs from conceptual analysis in many important respects. I understand that experimental philosophy does not aim at complete analyses, that it explains effects in terms of underlying cognitive processes, that it operates with a different conception of simplicity. But all the same, I can't help thinking that there is an important sense in which these two traditions are fundamentally continuous. After all, the original aim of conceptual analysis was to give us a better understanding of concepts, and it seems clear that experimental philosophy is doing exactly that. Whatever else one might say about experimental philosophy, surely one would have to agree that it has given us some fascinating insights into people's concepts!

I want to suggest now that even this objection is actually incorrect. Most research in experimental philosophy is so radically different from traditional conceptual analysis that it would be a mistake to think of it as doing anything like what conceptual analysis originally aimed to do.

To illustrate the key points here, it might be helpful to look at one specific case and examine it in real detail. Let us focus, then, on the study of intuitions about *knowledge*. As we will see, there has been a surge of experimental research on this topic, and this research has arrived at a number of fascinating insights. However, it would not be accurate to say that work in this field is in any way providing us with the sort of thing that conceptual analysis originally hoped to produce. In particular, the insights it has arrived at are not properly described as being about the concept of knowledge.

To begin with, let's consider again the plight of a conceptual analyst working in this area. She is exploring intuitions about knowledge, when she notices something interesting. It seems that people are reluctant to say that someone knows that *p* in cases where *p* is false. So she decides to write out the first condition of what she hopes will eventually be a successful analysis of the concept of knowledge.

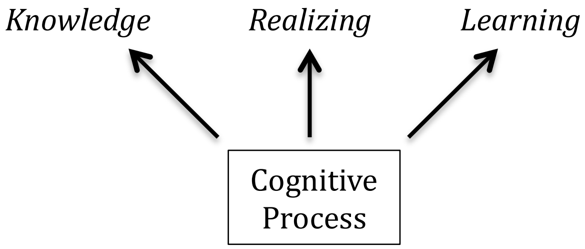
A subject knows that *p* if and only if:

(1) *p* is true

Judged by the standards of conceptual analysis, this opening salvo might appear to be strikingly simple. In fact, one might think that this first condition is a real success and that the problems only began to arise later, as it becomes clear that she will not be able to capture all the nuances of the concept using conditions as simple as this first one.

But seen from the perspective of contemporary experimental philosophy, it seems that the analysis is already too complex. It is not enough just to introduce a principle that directly describes the pattern of people's intuitions. We need to develop a theory that *explains* this pattern. Such a theory would tell us about the underlying cognitive processes that lead people to have the intuitions they do. It would then allow us to offer a far simpler account of this effect.

Buckwalter (2014) reports a series of experiments designed to pursue precisely this strategy. He begins by noting that the concept of knowledge can be seen as just one example from a larger class of concepts. (This class includes the concept *knowledge* but also the concepts *realizing* and *learning*.) He then posits an underlying psychological process that impacts people's use of all of the concepts in this class. Ultimately, he ends up with an account that looks like this:



The resulting theory explains one aspect of people's knowledge intuitions. Specifically, it helps us to understand why people sometimes have the intuition that it is possible to have knowledge of false propositions ('He just *knew* that he was going to hit the jackpot this time…') but more often have the intuition that only true propositions can be known ('He couldn't have *known* that he was going to hit the jackpot – look at what ended up happening!').

What I want to emphasize here, however, is not the details of Buckwalter's theory but its general strategy. The idea is not to add to our account of the concept of knowledge a separate claim of the form: 'In cases of the following type, people will be reluctant to ascribe knowledge of false propositions.' Instead, Buckwalter develops a theory that is not properly described as being about the concept of knowledge at all. This theory describes a particular sort of cognitive process, and Buckwalter is able to test it by looking at the use of a variety of other concepts (*realizing*, *learning*, etc.). Then, once this theory is in place, it becomes possible to explain the key phenomena while introducing only quite minimal assumptions about the concept of knowledge in particular.

Now suppose we turn to a second fact about people's intuitions. It has often been suggested that people's intuitions about knowledge depend in part on judgments about the *relevant possibilities*. People seem to regard certain possibilities as relevant and others as irrelevant, and this distinction appears to play a role in their intuitions about whether a given mental state counts as knowledge.

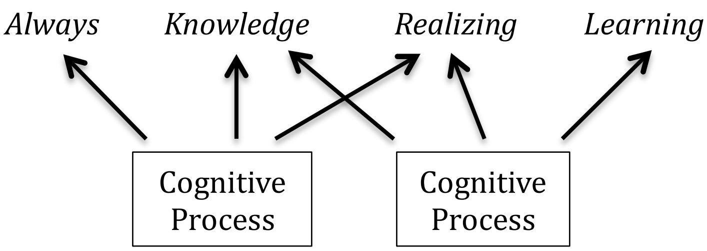
Within the tradition of conceptual analysis, the obvious way to capture this fact would be just to directly add it into the analysis. Thus, we might proceed by writing out a second condition:

> A subject knows that *p* if and only if:
>
> (1) *p* is true
> (2) All possibilities that have properties *F* or *G* are ruled out by the subject's evidence

We would then have to find some way of filling out this condition in more detail, leaving us with a quite complex account.

But here again, this approach would not be considered at all appropriate by the usual standards of experimental philosophy. Experimental philosophers would not be satisfied with an account that proceeds just by directly stipulating that a property will be considered relevant under certain conditions. Rather, they would want to identify the underlying psychological processes that lead people to see different possibilities in this way. Ideally, claims about these processes would then be backed up by independent evidence.

As it happens, Jonathan Schaffer and colleagues have been pursuing an experimental research program along precisely those lines (Buckwalter & Schaffer, 2013; Schaffer & Knobe, 2012; Schaffer & Szabó, 2014). They have been gradually developing a general theory of the way people quantify over relevant situations. This theory aims to provide insight into people's ordinary use of the concept *knowledge*, but it would also help us to understand their use of various other concepts. (For example, it would help us to understand their use of concepts like *realizing* and also concepts like *always*.) Putting this research program together with the previous one, we are thereby left with a picture that looks like this:



Notice what is happening as this research program progresses. We are certainly learning something important, but it would be highly misleading to say that we are gradually adding to our theory about the concept of knowledge. In fact, the real effect of the research program is in exactly the opposite direction. As it continues to progress, we are adding ever more to our theory about how people quantify over relevant situations. This progress then allows us to explain the relevant facts about people's intuitions while building ever *less* into our account of the concept of knowledge in particular.

Let us now sum up. Philosophers have noted certain striking patterns in people's intuitions, and it is natural to seek to capture those intuitions in a philosophical theory. The original aim of conceptual analysis was to capture the patterns in people's intuitions through theories about the corresponding concepts. (One would capture intuitions about knowledge in a theory about the concept of knowledge, intuitions about intentional action in a theory about the concept of intentional action.) On first encountering experimental philosophy, one might well suppose that it is aimed at doing something broadly similar. That is, one might think that it continues the traditional effort to arrive at theories about people's concepts, though this time with the benefit of experimental research methods.

However, a closer examination of the actual research suggests that this is not the case. Instead, experimental philosophy has sought to capture the patterns in people's intuitions through theories about underlying cognitive processes. In actual practice, this does not involve analyzing concepts, or doing something broadly similar to analyzing concepts, or engaging in some preparatory work that would eventually allow us to analyze concepts. It is not a matter of analyzing concepts at all; it is something else entirely.

IX

As we noted at the outset, there has already been a great deal of excellent metaphilosophical work exploring the methods and aims of conceptual analysis. The insights coming out of this work give us an enormous advantage whenever we are trying to understand research that either contributes to or attacks the conceptual analysis tradition. After all, when we are trying to understand research of this type, we can simply turn to the theoretical frameworks developed in existing metaphilosophical work and apply them to the case at hand.

I have argued, however, that research in experimental philosophy does not fall into this category. The vast majority of empirical research in experimental philosophy neither contributes to nor attacks the conceptual analysis tradition. On the contrary, the vast majority of this research is *cognitive science*. It consists of identifying surprising effects in people's intuitions and explaining those effects in terms of underlying cognitive processes.

Thus, if we want to arrive at a better understanding of contemporary research in experimental philosophy, it will not be helpful just to assume that the metaphilosophical questions we face are minor variations on the ones that arose for conceptual analysis. The question we face is not something along the lines of, 'What are the implications for larger philosophical issues of a research program that explores people's concepts?' Rather, the question we face is a different one entirely, namely: 'What are the implications for larger philosophical issues of a research program that explores the cognitive processes underlying people's intuitions?'

References

Alexander, J., Mallon, R., & Weinberg, J. M. (2010). Accentuate the negative. *Review of Philosophy and Psychology*, 1, 297-314.

Alexander, J., & Weinberg, J. M. (2007). Analytic epistemology and experimental philosophy. *Philosophy Compass*, 2, 56-80.

Alicke, M., Gordon, E., & Rose, D. (2013). Hypocrisy: What counts? *Philosophical Psychology*, 26, 673-701.

Alicke, M., & Rose, D. (2010). Culpable control or moral concepts? *Behavioral and Brain Sciences*, 33, 330-331.

Alicke, M., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *Journal of Philosophy*, 108, 670-696.

Buckwalter, W., & Schaffer, J. (forthcoming). Knowledge, stakes, and mistakes. *Noûs*.

Danks, D., Rose, D., & Machery, E. (forthcoming). Demoralizing causation. *Philosophical Studies*.

De Brigard, F. (2010). If you like it, does it matter if it's real? *Philosophical Psychology*, 23, 43-57.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, 106, 587-612.

Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical explorations*, 10, 95-118.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, 33, 315-329.

Knobe, J., & Szabó, Z. G. (2013). Modals with a Taste of the Deontic. *Semantics and Pragmatics*, 6, 1-42.

Leben, D. (2014). Motivational internalism and the true self. Unpublished manuscript. University of Pittsburgh, Johnstown.

Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy*, 31, 128-159.

Lycan, W. G. (2006). On the Gettier problem problem. In *Epistemology Futures*, Hetherington, S. (eds.), New York: Oxford University Press, 148-168.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations*, 9, 203-219.

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value Judgments and the True Self. *Personality and Social Psychology Bulletin*, 40, 203-216.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, 24, 586-604.

Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry*, 20, 30-36.

Rose, D., Livengood, J., Sytsma, J., & Machery, E. (2012). Deep trouble for the deep self. *Philosophical Psychology*, 25, 629-646.

Schaffer, J., & Knobe, J. (2012). Contrastive knowledge surveyed. *Noûs*, 46, 675-708.

Schaffer, J., & Szabó, Z. G. (2014). Epistemic comparativism: A contextualist semantics for knowledge ascriptions. *Philosophical Studies*, 168, 491-543.

Silver, I. (2014). [Empirical studies in experimental philosophy, 2009-2013]. Unpublished raw data.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, 132, 99-107.

Sripada, C. S. (2010). The Deep Self Model and asymmetries in folk judgments about intentional action. *Philosophical Studies*, 151, 159-176.

Sripada, C. S. (2012). What makes a manipulated agent unfree? *Philosophy and Phenomenological Research*, 85, 563-593.

Sripada, C. S., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind & Language*, 26, 353-380.

Turri, J. (2013). The test of truth: an experimental investigation of the norm of assertion. *Cognition*, 129, 279-291.

Turri, J., & Blouw, P. (forthcoming). Excuse validation: A study in rule-breaking. *Philosophical Studies*.