Methods, Measures and the Mental

Some Remarks on Empirical Philosophy

R.J. Anderson

Horizon Digital Economy

University of Nottingham

W.W. Sharrock

Department of Sociology

University of Manchester

FOREWORD

This is another foray into the confused arena of the relationships between disciplines. In our previous examination of this topic, our concern was to rein back the imperialist ambitions of our sociological colleagues and to inform those who might be subject to this imperialism (and, *mirabile dictu,* even welcome it) what they might be getting into. This time it is not sociology's overweening aspiration we have in view but proposals developed within parts of cognitive science and directed towards the practice of philosophy. We have no doubt that philosophy can deal firmly enough on its own with the philosophical implications of this cross border incursion. It will need no help from us on that front. However, what is being used to justify the need for the infiltration is the contribution certain investigative methods in the social — and to a lesser extent the psychological sciences — could make to improving philosophical methods. For that to be even feasible, two things need to have been secured. First, a commonality of methodological outlook between philosophy and the social and psychological sciences must have been established. Second, the methods which are to be imported must be used in ways that preserve their integrity. Even though we have views on the former, we will not pursue them here. Instead, it is the latter issue we will discuss. Do the examples given of the use of social and psychological science's methods to answer philosophical questions cut the mustard? If they don't, what is the point of importing them? The conclusion we come to is that they do not. We recognise this conclusion is not, of itself, enough to cause cognitive science to revoke its proposals. What it does do, we think, is narrow the debate to the philosophical implications of cognitive science's recommendations. As we say, we are happy to stand back from that discussion and leave the field to the philosophers.

1 THE PROGRAMME OF EMPIRICAL PHILOSOPHY

In his wonderful little book *Dilemmas*, Gilbert Ryle construed many of the central conundrums of philosophy (or, to be more accurate, many of the problems presented to students and other tyros as central to philosophy) as a roster of needless logical litigations between different ways of conceptualising what appear to be 'the same' problem. In Ryle's view, the conundrums evaporate once the provenance of the differing conceptual schemes is laid out. All the sound and fury signifies nothing, or at least nothing of lasting philosophical significance, only philosophers' hyperactive imaginations. Stephen Stich, Eduard Machery, Joshua Knobe, Shaun Nichols and their colleagues appear to agree with him. They too believe many of the central problems in philosophy are more the artefacts of philosophical practice than representations of real issues in metaphysics, epistemology, ethics, the theory of language and so on. However, their diagnosis is somewhat different. They hold that what creates the illusion of substance is the reliance philosophers place on intuitions and, in particular, on their own intuitions. In framing an argument or making a move within an argument, philosophers often propose that 'we' find some interpretation or conclusion 'intuitively' attractive, sound, sensible or secure. Having given 'our' authority to this intuition, they proceed to cement it into the support for their arguments. To Stich et al., these intuitions are actually putative empirical generalisations about what 'we' or 'anyone' would say. Since, on this line analysis, examining what we would say is the method by which philosophy seeks to explicate and organise our concepts and the theories based on them, Stich et al. take it this makes the practice of philosophy empirical. They do not argue it *should* be empirical, simply that it is. However, they go on to deny the generalisations deployed in this practice can carry the weight placed upon them and hence the issues raised and arguments offered that utilise them are flawed. To rectify this, philosophical practice must be revised and the whole discipline placed thereby on a more secure empirical footing. This will be done by the utilisation of methods commonly available in the social and psychological sciences. Hence the banner under which they offer their critique. They are proposing an empirical reformation in philosophy. As a consequence, the community making this argument has been dubbed Empirical Philosophy (EP).

The Argument from Linguistics

One way that EP has tried to explain its proposals is by drawing a comparison with Structural Linguistics (MACHERY & STICH 2012). Despite its mathematical sophistication, in Structural Linguistics judgments about the grammaticality of sentences, the bindings of reference and so on depend upon the intuitions of researchers. It is these which are used as 'evidence' for the particular theoretical conclusion being pressed. The security of this evidential base (i.e. the extent to which we should be prepared to accept it) turns entirely on the presumption that the intuitions of the researchers are generalisable to all speakers of the language (or, for some orders of claim, to all speakers of any language). Machery and Stich suggest there is no good reason to believe this generalisation strategy is well grounded. Ordinary speakers routinely accept as grammatical and meaningful sentences which theorists reject as ill formed, ambiguous or meaningless. If linguistics is an empirical discipline offering theories which describe the structure of our language use, then relying on researchers' intuitions alone leave it on shaky ground. However, as they are pleased to point out, since the mid-1990s there has been an increasing use of survey methods and other techniques to acquire data on people's views of language use. These methods go some considerable way to obviating the dangers inherent in the reliance on researcher intuitions. To begin with, there is less likelihood of such intuitions being shaped by a researcher's own theoretical commitments. Second, attending to actual language use opens up the possibility of observing variability in usage to which the researcher might otherwise be blind.

Notice this is not an argument about whether to use intuitions rather than, say, observations about corpora of collected speech, but about *whose* intuitions we should use. Nor is it a claim about who can or cannot be reliably expected to articulate their intuitions in sufficient detail to make grammatical and semantic judgments. It is simply an argument about whether researchers' intuitions can be taken to be representative of the population of speakers as a whole and whether we are quite sure their conclusions are not contaminated by their prior theoretical commitments.

Underpinning all this is a picture of the relationship between language use and the descriptions provided by linguistics which goes as follows. The infinite totality of sentences and other expressions of meaning which can be properly articulated in a language is governed by a defined set of rules. These are the rules that people use. That these rules exist and, at least in principle, are describable is an empirical fact and the descriptions themselves delineate empirical facts. The rules that people actually use *are* the rules that linguistics seeks to describe. The aim is to have a tight mapping between the rules that are used and the rules that linguistics describes. Examining our intuitions (linguists' and lay persons') regarding test sentences is the only way we have for determining the validity of the rules we follow. Moreover, when found to be (sufficiently?) valid, such rules exist independently of the methods by which linguistics makes them visible. They are 'real' in an epistemological sense. The question is simply which is the best/better way to test validity? Samples of one or samples of many? In other words, the issue EP is pointing to is more usually described as a matter of sampling error and measurement error.

Small samples give high sampling error; the reliance on their own intuitions means linguists may mismeasure/misinterpret or distort the mapping between the rules (as they really are) and our intuitions about them.

To sum up. Linguistics discovers the facts about the rules governing language use. These rules are real and valid statements about them are true. We determining validity (and hence truth) by consulting our intuitions. Given the possibilities of sampling error and measurement error, if we want linguistics to be soundly based as an empirical discipline, we would be advised to use large samples of language users rather than small and certainly not samples of one. By adopting some of the methods of the social and psychological sciences, linguistics is beginning to address this issue. In drawing the comparison between philosophy and linguistics, EP is suggesting that philosophy's empirical methods are as loosely constructed as linguistics' once were and precisely the same remedy should be applied.

We are not going to take up the question of whether philosophy's practice is or should be empirical in the sense that linguistics may or may not be empirical. It seems to us that before this matter can be settled, we have to be clear whether we are being asked to discuss a 'sociological' finding given by studies of the practice of philosophy or a normative judgment about philosophy, namely that philosophy ought to be empirical and at present is not nearly so. Until we know what is actually being said, we don't know how to respond. However, what we can respond to is the way that EP proposes to carry out the empirical studies it insists should be undertaken. It invokes the methods of the social and psychological sciences, but from the studies it presents as evidence of the need for its reformation, it appears to have little knowledge of the extensive debates in the social and psychological disciplines over the way surveys and similar studies should be framed and implemented and, in particular, of the numerous detailed methodological analyses of 'the interview' and 'the survey' setting. Our discussion summarises some of these issues and sketches their implications for EP's proposals.

The Format of the Studies

The studies EP carries out fall into two broad types which use a common format, one that is familiar in the social and human sciences. The format sets up a stimulus scenario or story and asks participants to say how far they would agree with summary statements about it. Alternatively, they are asked to make non-directed judgments (i.e. without the benefit of a summary statement). Participants' answers are the data of the study. This data is assumed to be an expression of views, beliefs, understandings, knowledge or, as in the cases we will discuss, intuitions the respondents have regarding the issues under investigation. The studies differ among themselves only in the selection of the participants. Do they come from panels of volunteer undergraduates? Or are they collected using a web tool such as Survey Monkey? Both approaches are standard.

Our concerns lie with how these studies and the issues framed within them are set up as well as with the findings which can be offered on the data so generated. Nothing we have to say is

new or particularly insightful. The issues are standard fare in current debates over method and research technique in psychology and the social sciences. We draw attention to them simply because much of the authority EP claims for its studies rests on the fact that they have been borrowed from and are successful in the social and psychological sciences. And this is true. In those disciplines, there are very many well-crafted examples which yield dependable data and on which qualified generalisations can be made. It is simply that since EP's studies are not well crafted, the data generated by them cannot be assumed to be either dependable or generalisable.

The structure of our discussion is as follows. First, using a particular example, we will look at the way the way the studies EP undertakes are constructed. We will suggest that what appear to be straightforward protocols may, in fact, be deeply ambiguous. Second, we will review the standard requirements for statistical inference. We will suggest that EP's studies do not seem to fulfil them, or at least do not seem to acknowledge that there is currently debate over whether studies carried out in the way these are can fulfil them. Third, we will take a particular EP investigation of an instance of a philosophical topic (Judith Jarvis Thomson's "trolley problem") and show that the forms of reasoning presented in 'the instance' and in 'the experiment' are not isomorphic. We take this to threaten, if not completely undermine, any claim that the experiment renders 'the instance' otiose. We finish with some summary conclusions.

2 AMBIGUITIES AND CONTEXT

One of EP's main allegations about conventional philosophising is that philosophers are prone to ignore the narrative context of the scenarios they consider. Were they not to do so, there would be less inclination to generalise intuitions about the acceptability of particular inferences. Construct the stories differently and the intuitions might well be different. It comes as a surprise, then, to find studies such as that of Knobe and Nichols (2008) being so insensitive to the contextual framings they provide for the scenarios they want to test. They are challenging the idea that the incompatibility of free will with determinism is an intuitive matter, that it is just intuitive to us that our actions are free and not, therefore, determined. Their aim is not to resolve the determinism/free will debate, but to undermine the idea that free will and determinism are *intuitively* incompatible. If evidence shows that the same people can, under varying conditions, both accept and deny that free will and determinism are incompatible with each other, then

neither compatibilism nor incompatibilism can be treated as intuitively apparent.[1] Their case is that whether people will assent to compatibilism or incompatibilism depends upon the way in which the issues are set up. When the issues are stated abstractly, people tend to hold we live in a wholly determined world and hence have incompatibilist views about moral freedom and causal determinism. When they are stated concretely (that is, in terms of some specific and highly recognisable and ethically charged issue), people exhibit compatibilist reactions.

To explain how this contradiction comes about, Knobe and Nichols offer study participants two scenarios; one in which all aspects of the world are determined and one where the only difference is that human action is free. They acknowledge that debates over determinism are myriad and differ in many complex ways. To present all positions in full (or even a summary of them) would require technical concepts and language which the participants in their study would find both unfamiliar and, probably, off-putting. Nonetheless, Knobe and Nichols are certain that they can present matters in sufficiently cogent and transparent ways to stimulate simple (and hence straightforwardly countable) responses. These responses will reveal the respondents' intuitions about the acceptability of the positions in question.

Here are the scenarios.

> *Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next and so on right up until the present. For example, one day John decided to have French fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decisions, then it had to happen that John would decide to have French fries.*
>
> *Now imagine a universe (Universe B) in which almost everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it did not have to happen that Mary would decide to have French fries. She could have decided to have something different.*
>
> *The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision—given the past, each decision has to happen the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each*

[1] As an aside, the conundrum of moral freedom and causal determinism is one of the dilemmas Ryle disposes of.

> *human decision does not have to happen the way that it does. (Knobe & Nichols 2008 pp110 - 111)*

Having read the stimulus, participants are asked to say which description best fits our world and to explain their answer. For Knobe and Nichols, the answers provide statements about the participant's beliefs.

> *Across conditions nearly all participants (over 90%) judged that the indeterministic universe is more similar to our own (ibid p.111).*

The study then goes on to test intuitions about compatibilism and incompatibilism. This time, a half the participants are given a further story about events in Universe A. It provides the concrete (or "emotive") narrative. Here it is:

> *In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and three children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family.*
>
> *Is Bill fully morally responsible for killing his wife and children?*

The other half are simply asked:

> *In Universe A, is it possible for a person to be fully morally responsible for their actions?*

As Knobe and Nichols hoped, those participants given the concrete version of Universe A, offered overwhelming support for compatibilism. Those offered the abstract version gave equally overwhelming support for incompatibilism.

Knobe and Nichols take these results to be clear cut. As a result, they provide a long and elaborate account of the psychological mechanisms that must be at work to produce this 'paradox'. We are not so sure. We think there may be grounds for suspecting that a number of ambiguities might well be in play in the participants' responses as a result of the way the scenarios have been framed. The possibility of these ambiguities means that while Knobe and Nichols have been able to produce impressive counts of responses, *it is less than clear what those counts mean*. To illustrate what we have in mind, we will set out just three of the possible lines of ambiguity we have discerned.

Three Unsystematic Ambiguities

Although the study appears to set up A and B as apparently equivalent universes, in fact this is not so. First of all, the compatibilist option associated with Universe B is the more fully explicated, a feature which will become important when we discuss the task being set. Second, although the cases appear equal in what we might think of as argumentational weight, it is the unstated but implied consequences of determinism as described in Universe A which shape the study.

Universe B and the compatibilist position are framed in terms of the incompatibilist account implied in Universe A. To compare the viewpoints, the participants are being invited read the following implications into Universe A and thus into the incompatibilist position.

(i.) In a determinist universe every event is caused:
therefore;

(ii.) No one has any choice in respect of what they do.

(iii.) If choice over what to do is a condition of moral responsibility, and if no one can possibly choose to do anything:
then;

(iv.) It is not possible for anyone to be morally responsible for anything.

This piece of reasoning generates a number of ambiguities. How participants individually resolve them might well effect the findings of the study.

1 Methodological Petitio Principii?

The experiment is designed to allow free choice over the answers to the questions asked. But, if so, that implies a judgment about propositions we have just set out. If our universe were indeed as the determinist Universe A says, participants in the survey could not make decisions on their own behalf. Their responses (as human actions) would also be determined. But, since the concept of a fully deterministic universe actually provides no guidance for how to understand human action, if we do live in Universe A it is not clear that that the experimental set up could yield any data about intuitions over moral choices regarding Universes A and B. For Knobe and Nichols to believe that their study does do yield such data, they have to assume Universe B (or something like it) prevails. They present their scenarios as if they are agnostic with regard to A and B, when in fact they have to hold to B for the study to make sense. As a consequence, the survey displays a tacit pretence to an equal weighting between the incompatibilist and compatibilist positions. As such, we are reminded of Hume's acid comment about those who argued for Pyrrhonian scepticism:

> *If they be thoroughly in earnest, they will not long trouble the world with their doubts, cavils, and disputes; if they only be in jest, they are, perhaps, bad railers, but can never be very dangerous... (Dialogues Concerning Natural Religion p 7)*

2 What's the task? And what are the examples doing?

These might seem a strange questions but they are not. Knobe and Nichols are clear that their first question is designed to elicit *beliefs* about our world. But is that how it might appear to participants? Knobe and Nichols give no advice to participants about what constitutes an answer to their question. They have to work that out for themselves. Given the phrasing ("Please briefly explain your answer"), it would not be unreasonable to think one was being tested by being

asked to analyse the two universes and set out logical arguments for why A or B is similar to our world. This is even more likely when the subjects are all (as they were) undergraduates and hence used, if not predisposed, to taking tests of one kind or another. Setting out logical arguments in answer to a test is not the same as giving one's personal beliefs as an answer to an enquiry. Whichever conclusion participants eventually come to regarding the task set requires them to work out what determinism actually means in Universe A. Only when that is done are they in a position to provide assessment of the (logical or intuitive) acceptability of either position. But the only way this can be done is by consulting the stories or examples given. Yet, as we have just described, these could well be read as implying the unacceptability of the determinist position.

The specification of determinism given in the scenarios and in the explications admits of no exceptions, caveats or options. All action is determined. And yet in the initial account of Universe A, John is said to have decided to have french fries for lunch. Unless we want to say that what John did wasn't 'really' making a decision,[2] how can this act of John's be compatible with determinism? Since it is in the description of Universe A, it must be. But that can only mean that propositions (ii) and (iii) set out above are somehow misinterpretations of what is meant by determinism. Since there is nothing in the set up to explain how John's selection of french fries can be a *choice* motivated by a *decision* and still be determined, we are left with the puzzle of aligning how when someone has no choice over what they do, they can be said to make a decision to do it. For most people, we would think the obvious way this can be done is by assuming the illogicality of the determinist position.

Now consider the example in the concrete scenario. From the point of view of the determinist, the story is an irrelevance. If it is not possible for anyone to be responsible for anything they do, then, despite any surrounding contextual detail, there can be no moral responsibility. However, 'responsibility' is a cluster concept and no all elements of the cluster primarily rest on the implication of *moral* responsibility. For example, responsibility is often associated with the identification of a perpetrator. 'Who is responsible for this mess on the floor?' the teacher asks, wanting to know which one of the children made it. In the narrative there is only one possible answer to the question: Who was responsible for killing Bill's wife? Bill was. So is Bill responsible but not morally responsible? Is that what the story is telling us? But if that is so, the description of Universe A does not rule out the possibility of describing people as responsible for actions in one of the ordinary ways we do. Even in a deterministic universe, we can feel comfortable describing Bill as the perpetrator of the killing and death of his wife and children — after all, no one else was! If this is so, then there seem no reason not to evaluate that killing as a murder, in respect of which the perpetrator is indeed ordinarily deemed blameworthy. Once again, the only way this conclusion can be aligned with the principles of determinism is by assuming the determinist position actually espouses an inconsistency.

---

[2] Notice no-one is saying John had his hand forced, was being held to ransom or otherwise constrained — the circumstances we normally cite when we say someone "had no choice" or "didn't really" make a decision.

It seems, then, that the way the study is set up is deeply ambiguous. Its questions could be understood as solicitation of personal opinions or, equally, they could well be understood as tests of the reasoning capabilities of participants, in the latter case with the narratives to be taken as instructions for what an answer might be. In reviewing the responses, can we be sure the participants are expressing their own inconsistent views with respect to compatibilism rather than responding to what they perceive to be inconsistencies in the logic of the materials are presented to them?

3 When is a decision not a decision?

The ambiguities we have discussed so far hinge on a lack of clarity over the relationship between determinism and moral responsibility. In a determinist universe, what actually causes actions? Because everything is determined from the beginning of time, the individuals described cannot do other than buy the french fries or kill the wife and children; they cannot choose to act differently. It seems as if the whole weight of cosmic history is propelling *everything* that they do. Any example we consider must be yet another case of the same axiom: what people do next has already been decided well before they get to the point of doing it. Even if we suppose that the whole weight of the entire causal history of the universe is propelling events onward, there is still obscurity as to how this inexorable causal influence is effected (other than through the tautology that if an infallible predictor predicts that something must happen then it must happen, otherwise the predictor is not 'infallible'). At the very least, it is obscure what the causal source of someone's behaviour is. Is their behaviour their own doing or something that just happens to them?

The latter interpretation is what Knobe and Nichols want us to take determinism to mean. But if, from the beginning of the universe, it is decided (in the determinist sense that John will buy and eat french fries, then how can John now decide — in the ordinary sense of select between two possible alternatives, either of which can be chosen? Isn't it already decided in the first sense? And yet the descriptions of John's buying the french fries and Bill killing his wife and children state they do decide to do these things. But, if they are both deciding and not deciding, what on earth are we supposed to *think* they are doing?

Summary

As anyone who has ever constructed stimulus material for a class, an examination or even a survey knows, what may seem transparent to the author can often seem opaque and confusing to readers. For this reason, designers of surveys take special care in designing their prompts and questions and are particularly sensitive to the possibility that "helpful examples" might actually distract participants and induce more, not less, confusion. They are equally careful to ensure the way materials are presented does not unintentionally predispose interpretations which seem at odds with what the materials themselves seem to imply. Knobe and Nichols commit both gaffes. As a consequence, it is not clear what the results they have gathered actually mean nor, indeed, that the data gathered is actually data of the phenomena they wanted to investigate. They wanted

examples of beliefs and intuitions. As we have shown, they could well have been given the tentative conclusions of informal course of reasoning about the logic of their materials.

3 Experimental Philosophy's Inferential Strategy

The inferential strategy which EP adopts is, of course, very familiar. To determine the distribution of an attribute in a population, measure that attribute in a defined sample and then, using the methods of statistical inference, generalise from the sample to the population. Our question is simple. Given the samples EP compiles, can we be confident this strategy will work? Will well formed, valid conclusions relating to the populations at large be derived from the data collected? We think a moment's serious reflection would lead one to have considerable reservations on a number of fronts:

a. Ecological validity. Are we convinced that the responses generated are fair replications of how 'we' would actually respond when faced with 'real world' versions of the scenarios in the study set ups? How likely would we be to respond differently if we were really faced by the circumstances summarised in the stories as opposed to being asked to respond on an 'as if' basis to manifestly invented stories?[3]
b. Construct validity. Are we convinced that attitudes, intentions, feelings and the like are the kinds of things which are amenable to measurement and, if so, to measurement in this way? Do we have a theory of measurement (let alone a good theory of measurement) for this category of phenomena?
c. External validity. Are the samples used in the studies fair ones and fully representative of 'us', the ordinary members of society?
d. Statistical validity. Do the data satisfy the requirements of the statistical instruments and methods deployed on them?

In this section, we take each in turn and explain what our reservations are.

Ecological Validity

Imagine we wanted to see if it is possible to grow wine producing grapes 300 metres above sea level in our garden in the Staffordshire Moorlands. We take sample cuttings of a small number of healthy vines from a well-known wine producing area, plant them in compost with a pH of 7.0 and nurture them in our heated greenhouse. Two years later the resultant vines bear grapes. Should

[3] It is no rebuttal to say that this question applies across the board to this type of study no matter which discipline uses it, which it does. By extension, neither is it a rebuttal to respond that it holds just as well for philosophy's own accounts of the stories. The latter would be true if and only if philosophy was actually offering the summaries as empirical generalisations.

this experiment convince us that it is possible to grow wine bearing grapes in the Staffordshire Moorlands? In one sense it should. We have planted them and they have grown. We have data; the number of grapes that resulted. But are these data really representative of how grapes would fare under normal conditions in our region? The answer to that is obviously "Of course not!" The conditions normally found in the Staffordshire Moorlands (soil pH, temperature variation, rainfall, wind, etc. etc.) inhibit all but the hardiest plants and certainly would not encourage grape and wine production. This is not an issue about sampling. The vine cuttings are fair samples (precisely what that means we will discuss below). It is about inferring from one cluster of conditions to another on the assumption, if not identical, they are sufficiently similar.

The example just given raises two issues relevant to the studies carried out by EP. The first is what we might think of as *context calibration* — how far does the set up and operation of the study task match the target scenario? In EP's studies, this scenario is the addressing of 'epistemological' and 'ethical' puzzles in daily life. Do the study tasks map well on to this? The second is *semantic calibration* — how far do the meanings of key terms used in the story, or to explicate it, align with how these same terms are used in normal discourse? Can we assume a homogeneity of meaning?

The studies aim to ascertain how 'we' would respond when faced with philosophical conundrum or moral dilemma. Context calibration is about how far the set ups described reproduce the circumstances in which philosophical and ethical judgments normally get made. This is important because EP does not want to make claims about how the population makes judgments when run through psychological experiments but rather how we would normally or routinely make those judgments in the rest of our lives. A first and perhaps most important observation is that we do not normally make such judgments on an *as if* basis. Rather when we need to make judgments like these, we know the way our lives will be lived out will turn on them. The choices have consequences for us; there will be things we will now do differently, ways we will have to behave differently. Moreover, these judgments are made as part of a flow of action set in circumstances in which we are immersed and absorbed in managing. This backward and forward contingent character is central to the context in which we make these kinds of judgments. It is entirely distinct from the abstracted and reflective modality which the experimental or survey set up creates. Does the fact that these decisions will matter to us in real life and are non-consequential for us in the study make any difference to how we form judgments and what those judgments are? Do we know? Can we be certain it doesn't?[4]

In addition, the decisions we routinely make in daily life are just that, routine decisions. We are familiar with their character and scope. This is precisely what is not the case in the experimental set ups. The stories presented there are philosophical toys created to draw out and

[4] In response EP researchers could say that they feel the parallels are sufficiently strong. It seems to them they make decisions in the ways the set up requires. But that leaves us with the irony that they are relying on their intuitions to refute the reliance of intuitions in philosophy.

emphasize specific philosophical conceits. In working out whether the lady playing the organ at a funeral is the same person we bumped into in the supermarket yesterday, we don't immediately worry about the possibility of identity theft. Equally, when deciding whether anyone is at fault for spilling the milk or upsetting the relatives, we do not start with the presumption that that a supercomputer might have made a Laplacean prediction about all future events, or that someone has had their brain re-wired. Identity theft, Laplacean supercomputers and re-wired brains are not normal features of our determinations of who is whom or who is to blame for what.[5]

The point is that in wanting to confront participants with issues which are of relevance to philosophy, the designers of EP experiments import examples and conditions which are only of interest to philosophers. And, to allude to David Hume once more, philosophy is not normal life. Asking participants in a study to solve mock philosophical puzzles is neither identical with nor similar to the normal conceptual or moral judgments we make in the course of living our lives. The experiments EP runs are neither idealised versions of the ways such judgments are made nor abstractions of the epistemological and ethical puzzles philosophy worries about.

The extent to which words taken from ordinary discourse become terms of art in philosophy has been widely commented on. 'Cause' is just one prominent example. Equally, when philosophers talk about 'identity', 'freedom' or 'responsibility' what they have in mind are refined and highly attenuated forms of the cluster concepts used in daily life. Possibly the most obvious place to see this is in the distinction that philosophy draws between 'knowledge' and 'belief' and in the conundrum whether 'justified true belief' is 'knowledge'. As Simon Cullen (2010) points out, in order to make the disjunction between knowing and believing explicit, Weinberg et al. (2006) had to qualify these two concepts as 'really knows' and 'only believes' thereby suggesting participants see the one disvalued in terms of the other. This is not to say that ordinary people do not see a difference between knowing something and believing something. It is just that they do not necessarily see them as standing in an epistemic order. Equally, when study set ups ask if our actions are fully free, say, or what it is about a person that identifies them, notions of 'causally determined' and 'identity' are very different from how these words are used in ordinary life. Unless we give a direct steer on how to interpret such terms (and in the previous section we saw the difficulties that might generate), all participants have to draw on to make sense of the puzzles which have been set is how the words are used in ordinary life.

Construct Validity

Construct validity requires that the attribute being investigated be validly measured by the instruments deployed to measure it. For EP, this means that the statements of agreement or disagreement given by the study participants must be a valid measure of their intuitions, beliefs, feelings or intentions concerning how they would respond or what they would do. This

[5] These are all examples used in various EP studies.

requirement has two components. First, the attribute (intuitions, feelings, intentions etc) must be measurable. Second, the responses must be valid measures of the measurable attribute.

It might be thought that establishing the measurabilty of things like attitudes, intuitions, intentions and the like concerning conceptual issues would be a first and necessary thing for EP to undertake. It would then have some warrant for constructing its study set ups. However, as is usual in many social and human sciences, EP simply takes this step as read. In psychology in general, it is an unexamined assumption that intentions, intuitions and so on simply are measurable. In his classic analysis, Joel Michell (2004) characterises this presumption as follows:

> *In their quest for mental measurement, psychologists have contrived devices (tests or experimental situations) which, when appropriately applied, yield numerical data. These devices are treated as windows upon the mind, as if in the fact of yielding numerical data they revealed quantitative attributes of the mind. However, the windows upon the mind presumption dissolves the distinction between cause and effect, in this case the attributes of the mental system causing behaviour and attributes of the effects this behaviour has upon the devices contrived. That the latter possess quantitative features in no way entails that all of the former must. Hence, the windows on the mind presumption is questionable and, so, in the absence of additional, relevant evidence, not a sound basis for accepting the conclusion that the numerical data procured via the contrived devices is a measure of anything. (Michell. 2004, pp. 21-2)*

What would it take to refute Michell and demonstrate intentions and intuitions etc. are measurable? To start with we would have to show that they are quantifiable. Non-quantifiable attributes are binary. This cup either is or is not blue. Of course we can discuss the hue and intensity of the blueness and we might have difficulty (some of us have extreme difficulties!) telling shades of blue apart. But whether the object has the colour is not quantifiable. Now take the mass of the same cup. If presented with an array of cups, we could line them up in their order of mass. Their massness (so to speak) has an ordering. Quantities are orderable and have a relationship to one another. Saying this cup is 100 grams means it is twice as heavy as another cup of 50 grams and 100 times heavier than the standard unit gram. All these weights fall within the possible range of masses. The range of masses which objects can have constitutes the class of mass. Such classes are called *attributes.* The point about the properties of a range of an attribute is that they are mutually exclusive. This cup cannot be both 100 grams and 5 grams. Moreover, different members of the range stand in numerical relationship to one another. Measures are numerical relationships.

Furthermore, numerical relationships have an *additive* structure. If the paperweight is 300 grams and the cup is 100 grams, when both placed on a balance together they will need a

counterweight of 400 grams. In additive numerical structures, if a, b and c are ranges of the property of an attribute then:

1. For any value a and b, only one of the following can be true:
    a. a = b
    b. There exists c such that a = b + c
    c. There exists c such that b = a + c
2. For any values a and b, a + b > a
3. For any values a and b, a + b = b + a
4. For any values a, b and c, a = (b + c) = (a + b) +c

Two final conditions for measurement are required. First, if an attribute is measurable there can be no upper or lower limit of measurement and, second, the system of measurement must be continuous — that is, there are no gaps in the range. If an attribute is additive, unbounded and continuous, then it is measurable. The "scientific task", as Michell terms it, is to demonstrate that the attributes under investigation are measurable. This, he argues, is precisely what has not been done in regard to many psychological concepts. There is no empirically grounded theory of measurement which demonstrates that many psychological attributes are measurable. In particular, there has been no first principle demonstration of the measurability of our intuitions, feelings, intentions and the like which could underpin the measures of them operationalised in the social and human sciences.

We are not debating what intentions, feelings, beliefs and intuitions are. All we are asking is, whatever they are, are they are measurable and quantifiable? Do they pass the additivity, unbounded and continuity tests? Are they of the same type as our ordinary concepts of length, weight, volume and so on? Pointing out that psychology has measures for them is no answer. As Michell points out, if they are non-measurable, then it doesn't matter what measures have been applied to them. At best, they will be an irrelevance; at worst completely distorting.

What does it mean to say that our intention to talk on the phone at the weekend is measurable? Does our intention to talk on the phone stand in an ordered relationship to our intention to go to the pub or watch television? Note, this is not a question of desire. That is another question entirely. Can we concatenate the intentions to call on the phone, watch television and go to the pub to produce a composite intention that is the sum of them all? Can we even do that for repeated intentions to phone each other yesterday, today and tomorrow? What about judgments about right and wrong? If we say that breaking the speed limit is wrong and also say that stealing birds' eggs is wrong, can our judgments be ordered? (Notice, again, this is not the same as asking if the offences are equally wrong). Can we add the judgments together to get another which is the sum of them both (the judgment of a combination of speeding and birds' egg stealing)?

None of these questions serve to rule out any kind of psychological investigation. All they bring out is that before we can worry about whether the set up conditions of any psychological (or any other) investigation accurately capture the quantitative characteristics of the attributes we are investigating, we have to satisfy ourselves that such attributes are quantitative. If we do not, we risk engaging in a lot of what will be misplaced effort (at best). And, to repeat Michell's charge, that is precisely what psychology has not done. In their basic text on method, Bond and Fox succinctly summarised the result of this failure as follows:

> *What then happens in practice is that psychometricians, behavioral statisticians, and their like conduct research as if the mere assignment of numerical values to objects suffices as scientific measurement..... This is evidenced by such widespread practices as summing values from responses to a Likert scale and treating the total score as if it were a measure. The lack of empirical rigor in such a practice is indefensible. Numbers are assigned to response categories to produce ordinal-level data, after which these numbers are summed to produce a total score. This total score then is used in subsequent statistical analyses. The ordinal data are treated as if they were interval-level data, and no hypotheses are tested to acknowledge that this particular assignment of numbers represents a falsifiable hypothesis. Hence, the additive structure of these quantitative attributes is summarily ignored. Quantitative researchers in the human sciences need to stop analyzing raw data or counts, and instead analyze measures. (Bond and Fox, 2001 p 2)*

Let us set aside these difficult problems of quantifiability and accept the assumption that the attributes we are interested in are (somehow) quantifiable. The next question is how to instrument them. The standard method is the self-report. Ask people what they what they think about how they would feel or act and use their answers as indicators or measures of the relevant psychological state. Of course this takes us straight back to the issue of ecological validity which we discussed earlier. However, we will add another assumption to the one we have just made, and assume the attributes revealed by answers to questions in investigative set ups are our intentions, beliefs and so on. Now the only question is about the construction of the questions and the coding of the answers. Are either of these likely to have their own confounding effects? And if so, has EP avoided them?

To recap our earlier discussion. Before they can furnish the data EP is looking for, the investigative set ups require respondents/participants (a) to understand the scenario which is given to them; (b) to understand the questions they are asked; and (c) to understand what appropriate answers to those questions would be. The standard account of meaning in language suggests that understanding involves syntactic, semantic and pragmatic elements. All three are critical to how speakers and hearers understand each other. It was Antti Kauppinen (2007) who first

indicated the lack of attention given by EP to pragmatic or contextual considerations in the resolution of meaning in its investigations. Numerous studies in socio-linguistics and conversation analysis have shown that it is not just the terms used but the ways in which they are deployed (for example their ordering and placement) which are used by respondents/participants to understand the questions asked and work out what kind of answer is looked for. As every survey researcher learns, the question frame is not neutral. Ask questions in a different ways and you will get different answers. EP actually knows this too since a number of its studies are about the ordering of the vignettes, examples and tasks. The argument is that the effect of changed ordering demonstrates that the summaries of intuitions, judgments and so on given by philosophers cannot be treated as the 'ground truth' regarding the population at large. We have already noted the oddity that EP does not see the same considerations apply to its own method.

The net result of disregarding the contextuality of question construction leads to what Cullen (2010) calls "Survey-Driven Romanticism" (in contrast to the Intuition-Driven Romanticism that EP accuses philosophy of exhibiting). Survey Driven Romanticism assumes that

> *.....people's philosophical intuitions are implanted within them in some way, and by administering simple surveys we can discover them. (Cullen 2010 p. 277)*

As a result

> *Experimental philosophers can effectively equate survey responses with intuitions only by ignoring the established social and cognitive science literature on survey methodology..... The conclusion to be drawn is that, despite their pretensions, many experimental philosophers have given no serious thought to methodology. This not only undermines their claim to be doing science, as we shall see, it often leaves the philosophical significance of their findings unclear. (Op. Cit. p. 278)*

Cullen himself undertook a number of studies to explore how question framing and question format affects the ways questions are understood. The words that are used and the relative openness of the answer form required were both found to shape the responses made. Here is the conclusion he draws from his investigations.

> *Research has repeatedly shown that subjects rely on pragmatic cues and conversational norms to generate intelligent responses to survey questionnaires. It is only by effectively identifying intuitions with survey responses that experimental philosophers have been able to conclude that "intuitions . . . vary according to whether, and which, other thought experiments are considered first", or that it is a "fact that epistemic intuitions vary systematically with culture and [socioeconomic status]". These assertions are not made on the basis of "straightforward data*

> *about people's intuitions concerning specific cases"; rather, they are made on the basis of how people are inclined to use certain English words like "really knows" and "only believes" within the unusual conversational context of an experimental philosophy survey. (Op Cit. p.294-95)*

Other researchers such as Couper (2000) and Gosling et. al.(2004) have pointed out that while construct validity may be a problem for standard survey and experimental methods, the uncontrollable factors introduced by use of web-based surveys make them almost useless as an investigative technique. Both the lack of face to face opportunity for explanation and feedback and the likelihood of technology differences introducing artefacts into the presentation of the stimulus material lead to the conclusion that we must we very wary about using measures from web surveys as the basis of generalisation. Dillman & Bowker (2000) cite a number of examples of how such simple things as screen resolution, html parsers and the like can affect the way a question form is presented and hence shape the experience and understanding of the participant. Such variations could quite easily have uncontrollable effects on the interpretation which participants make of questions.

What conclusion are we to draw from all this? We have seen that construct validity requires that if we are trying to measure something, that thing must be measurable. Second, it requires that the ways we set about measuring do not themselves affect the measures we obtain, or if they do they do so in ways we can discount. On both counts there are grounds for serious reservations with regard to EP's studies. In common with much of social and human science, EP operates on the principle usually attributed to Stanley Smith Stevens, namely that measuring psychological attributes is just a question of allocating numbers according to a rule. There is no prior necessity to determine if the phenomenon/phenomena are actually quantifiable. This is not to say that any such attributes are not measurable, only that we cannot be certain that the attributes we are interested in actually are. The consequence is we cannot be certain that our measures are meaningful. Second, we have seen that the design of the instrument we use to make these measures, namely survey based question and answer tasks, can have significant effects on the measures taken. Recognizing the influence of the instrument on the responses does not mean trying to neutralize the pragmatic element in question and answer frames. What it does mean is that such considerations must be borne in mind when we interpret the answers and code them in measurement schemes. What we cannot do is treat them as what Michell (2004) called "transparent windows" on psychological attributes.

External Validity

What about the samples themselves? Are they representative of the population(s) about whom the generalisations are made? There are two issues here: the first is about the reliance on undergraduates; the second is about the use of samples drawn from volunteers. Both concern the inferences we can draw from either type of sample.

As long ago as 1986, David O Sears (1986) warned of the danger of relying on undergraduates as the sole resource for psychological experiments. His worry was that psychology's theories and conclusions might tell us more about the social and cognitive characteristics of this particular segment of society than about society as a whole. In particular, such reliance might result in an over or under estimation of the value of some attribute (or of the range within which such an attribute might fall) should the attribute vary with the demographic and other characteristics of this sub-population. The obvious parameters to reflect on are age and socio-economic status since undergraduates (even today) are generally relatively young and come from relatively well-off groups in society. Being young, they have the normal psychological characteristics of the young. They tend to be more volatile in their own sense of self-worth; their social and political views are less crystallized than those of older people; they are more egocentric and require stronger signs of approval from peers. Finally, their social relationships tend to be less fixed. All these are *established findings* of psychological research. But, undergraduates are different not just to the rest of the population but other young people as well. They have been selected because of their cognitive capacities and tend to be more compliant to authority, more motivated by deferred gratification and so on. In his discussion, Sears traces how psychological theories and models of human behaviour have changed during the period in which the discipline has become so dependent on undergraduate experimental subjects. Such theories and models emphasis just the characteristics shown by this particular population.

> *In all these respects, the idiosyncracies of social psychology's rather narrow data base parallel the portrait of human nature with which it emerges. To caricature the point, contemporary social psychology, on the basis of young students preselected for special cognitive skills and tested in isolation in an academic setting on academic tasks, presents the human race as composed of lone, bland, compliant wimps who specialize in paper -and- pencil tests. The human being of strong and irrational passions, of intractable prejudices, who is solidly embedded in tightly knit family and ethnic groups, who develops and matures with age, is not that of contemporary social psychology; it does not provide much room for such as Palestinian guerrillas, southern Italian peasants, Winston Churchill, Idi Amin, Florence Nightingale, Archie Bunker, Ma Joad, Clarence Darrow, or Martin Luther King. (Sears 1986, p 527)*

Sears' discussion gathered what might thought of as circumstantial evidence. Robert A. Peterson undertook a meta-analysis study of a huge range of behavioural and psychological studies. His results emphatically corroborated Sears' speculations. His conclusion was

> *...it is important to emphatically point out that the present findings are not a per se indictment of research employing college student subjects. Rather, they simply demonstrate that research results produced using college student subjects may differ from research results produced using nonstudent subjects, just as research results based on seven-year*

> *old subjects may differ from research results based on 70-year-old subjects. (Peterson 2001 p459)*

Even though some more recent investigators such as Druckman and Kam (2009) have a more optimistic outlook, they too warn of the dangers of misestimating when studying certain topics. They suggest investigators should be attuned to the difference the characteristics of the cohort might make to the findings reached and where such effects are to be expected, dual samples from students and non-students should be taken.

Statistical Inference

It is axiomatic that the tools of statistical inference can only be used with random (or probability) samples. Such 'fair' samples are defined as those where any member of the target population has an equal, non-zero probability of being included. For obvious reasons, self-selected samples such as those used by EP are non-random. As a consequence, the standard battery of inferential tests for sample distributions and hypothesis testing cannot be used. Once again, it is not that any self-selected sample is automatically non-representative of the population at large, simply that we have no way of testing whether it is. As a result, we cannot regard generalisations from non-random samples as robust.

Basic descriptive statistics (frequency distributions, point estimates, measures of variance and correlation) can be used with these samples, albeit with care. Self-selected samples are not wholly useless. However, most descriptive statistics require interval and ratio forms of measurement and not the nominal and ordinal forms common in the research we have been examining. Using the mean to calculate the 'average' for a test of agreement which uses a 7 point Likert scale makes no sense if we cannot be certain that the 'psychological distance' between 'tends to disagree' and 'neither agrees or disagrees' (scaled at 4 and 5 respectively) is the same as the psychological distance between 'tends to agree' and 'strongly agrees' scaled at 6 and 7 respectively. Without such certainty, Likert scales are not measures in any strict sense. The 'average' we should use in these cases is the median not the mean. Since all analyses of variance and correlation utilise the mean, they cannot be used either.

Summary

We have not argued that the studies carried out by EP are entirely invalid. We have made a much more nuanced case. Given the requirements for ecological, construct, external and statistical validity, it is hard to see how the study set-ups and the samples used provide a robust basis for generalization. The reasons we have offered for this conclusion draw upon considerations debates in the social and psychological literatures about studies of the kind which EP proposes. Given that this is such contested territory, we are surprised that those proposing the extension of these methods into philosophy have not sought to explain why they think those with reservations are wrong and that this move will not add its own form of confusion to the resulting philosophical discussions.

4 AN EXERCISE IN TROLLEYOLOGY

Apart from our discussion of Knobe and Nichol's study of the compatibilism/incompatibilism 'paradox', we have couched our discussion in general terms. In this section, we will apply these general considerations to one specific example, the study carried out by Liao, Weigmann, Alexander and Vong (LWAV) (2012) to test just how far the stories Judith Jarvis Thompson tells about a runaway trolley and the moral choices which she derives from them, jibe with what a group of ordinary people would say. There is nothing about this example which makes it special other than the detail LWAV provide about how they conducted their study. Working through this detail will allow us to raise an issue we have not brought out before, namely the mapping of the story told in the study onto the story told in the philosophy. This will allow us to say something about the implications of the experiment for philosophical discussions. But we begin by summarising the trolley example.

The Trolley Example (version 1)[6]

As Judith Jarvis Thompson (JJT) (1985) has been happy to acknowledge on several occasions, the trolley example is not original to her. It was first put together by Philippa Foot (1967) in a discussion of the morality of abortion and what is called the doctrine of 'double effect'. For Foot, the question was the basis we might have for appearing to condone (i.e. accept the rightness of) terminating a pregnancy (i.e. killing a foetus) to save the life of the mother. This choice is an example of 'the double effect'. To explicate this philosophical issue, Foot gives an example which went something like this:

> A tram/trolley is running out of control downhill. Ahead on the main track ahead 5 men are working. There is a spur off the main line on which 1 man is working. The driver has no other choices but to plunge ahead and run into the 5 men thereby

[6] We give two versions. The first is the stripped down version used in LWAV's experiment. The second is a much fuller version derived from the various versions JJT has given. We use it to provide the context in which she worked though the issues she was mulling over.

> killing them or to turn off and run into the other man thereby killing him. These are the only choices he has.

It is important to note just how simple and restricted this story is. The driver has no other available actions. There are no means of giving warning or any other ways of saving the day. The simplicity and restrictiveness of course are designed to make the choices stark and mutually exclusive.

For Foot, it seemed natural to prefer to sacrifice (albeit reluctantly) one life in order to save many. But this preference could only be justified in very specific circumstances. She interpreted this as indicating that in making the decision, we were applying a set of governing principles which marked the difference between letting someone die and intentionally killing them. Letting someone die was permissible (in certain circumstances and under the principles); choosing to kill them was not.[7]

In her recent discussions (e.g. THOMSON 2008) , JJT has developed an alternative version of the trolley story. She called this the "Bystander Example". The driver has fainted and the only person who can intervene is a bystander who happens to be wandering past. The bystander can pull a switch and divert the trolley, thereby saving the five. JJT tells us she thinks that the re-directing of the trolley with its fatal implications for the single man is permissible. She then adds a further variant — the loop case. The track does not bifurcate but rather loops back on itself. She proposes that if the trolley plunges on it will kill all 5 but eventually will stop and not kill the one. If it takes the loop, the single man is fat enough to stop the trolley without killing the others. Yet again, JJT feels it is permissible to kill the one to save the five.

The last version of the trolley case is even more baroque. Is it permissible for the bystander to throw another bystander (this time a fat one) off a bridge and so stop the trolley before it reaches the five but at a cost of killing him? But now, aren't we killing the fat man to save the five? If so, JJT proposes that our intuitions seem to say we shouldn't do it.

Although there are actual cases where individuals have been forced to make the kinds of choices Foot and JJT describe, the trolley examples do not purport to be descriptive of any of them. They are used simply and solely to set up the choices as part of an argument about the principles governing certain kinds of action.

LWAV's Experiment

In discussing her examples, JJT constantly refers to the ways 'people she has asked' have made judgments about the cases and says things like "It seems to me…." and "I hope you too would agree …. ". The progression of her argument rests upon her taking us through a set of positions with regard to the examples. As with other EP studies, what interests LWAV is that numerous studies in cognitive science and psychology have demonstrated ordinary people's judgments

[7] Foot's example has generated a whole literature of its own as has JTT with various versions of the cases. We are not going to take part in any of these debates.

about the rationality, plausibility and conviction of sets of propositions is as much affected by contextual features as by the purely logical properties of the propositions themselves. One important feature is the order in which the propositions are considered. They ask if the order in which JJT presents her examples has an effect on the intuitions we might have about those examples? This is the question they set out to test. Their null hypothesis is that order has no effect and the intuitions we have about the cases will be the same no matter in what order the propositions are presented.

To test this, LWAV use a standard psychological set up with a randomised, controlled between-subject design. The scenarios, although broadly similar to JJT's, do differ in ways that may turn out to be significant. The agent making the decision is identified as 'Abigail' rather than simply being an anonymous bystander. To maintain surface consistency between the 'stories', Abigail pushes buttons to enact her decisions. In both loop cases, the potential 'victim' is defined as an innocent bystander and not a line worker or a fat man. In the surrogate for the Fat Man case, the bystander is moved in front of the trolley consequent upon Abigail activating the platform on which he is standing. The protocol for the experiment is given in Appendix 1. The last two scenarios are not analogues of any JJT examples and are for control purposes only. There are three conditions expressed as changes in the order of the cases with Push being presented before Loop in Condition 1 and Standard being presented first in Condition 2. (See Appendix 2). The survey was run as an online questionnaire with 145 subjects selected through an online crowdsourcing service and randomly allocated to each of the conditions. As each scenario was presented, the subjects were asked to score their agreement on a six point scale with 6 indicating strong agreement. The sample had the following demographics:

> *The majority of the subjects were between18 and 30 years old (58%), and the sample had a female bias (70%). 84% listed English as their primary language, and 31% had some background in philosophy. (Liao et al. p.664)*

Appendix 2 provides the mean scores for the responses under each condition. Apart from Condition 2 where Standard score is 4.34, the answers to all questions were highly concentrated in the 'mildly disagree' category. As they take this result to be statistically significant, LWAV conclude order appears to effect judgement. Testing for order differences between those who agree and disagree with an action found that a small majority agreed with the permissibility of Loop when presented after Push whereas a slightly bigger majority disagreed with Loop when presented after Standard. Again, this difference is statistically significant.

From the above results, LWAV conclude that since a random set of anonymous individuals seem to have their judgment of the cases affected by the order in which they were presented, JJT's arguments for the intuitiveness of the selections she describes and therefore for the principles underpinning them must be suspect.

> *Either a convincing case must be made that context is relevant to the moral permissibility of redirecting the trolley in Loop or Loop intuitions cannot legitimately play the evidentiary role that they were supposed to play in Thomson's argument against DDE. (Op. Cit. p 667)*

They speculate why order seems to be important in forming judgements of moral permissibility. This might be because subjects were making comparisons between the cases or because different features became salient in them depending on the order. Their final conclusion is:

> *In either case, the challenge is not to explain why our intuitive judgments track such things, but why we should think that ethical truth tracks such things. Without that kind of explanation, whatever other reasons we might have for wanting to reject DDE, it would be problematic to base the case against DDE on Loop intuitions. Such a situation would also call into question the philosophical justification of theories, like Kamin's DTE, that were constructed to accommodate Loop intuitions. Any theory proposed specifically to accommodate some set of evidence would have to be re-evaluated in light of identified problems with that evidence. (Op. Cit. p. 667)*

Discussion

Although LWAV's experiment is interesting in its own way, in that it replicates the results of countless other studies it is rather run of the mill. However, the conclusions they draw from it are not and, importantly, turn on the acceptability of their definition of JJT's assertions about what are and what are not intuitive as empirical generalisations. A related second issue is this. No matter whether JJT's assertions are empirical generalisations about people's intuitions, is the LWAV experiment a robust method of accessing those intuitions? The experiment assumes that the responses are data of the intuitions the subjects hold. How secure is that assumption? Third, irrespective of the outcome of discussion of the first and second question, does any conclusion about ordinary intuitions actually affect the logic of the arguments JJT makes? In other words, is discussion about the generality or not of the intuitions a philosophical irrelevance? These are the issues we will explore.

1 Varieties of context

We start with the conclusion that LWAV draw from their experiment and the notion of decontextualised 'ethical truth' they are arguing against. They say the order in which cases are presented makes a difference to how we interpret them and hence to what they mean and what we infer about them. What they are trying to refute is the view that our intuitions about cases are not in some way related to the context in which they are presented, and by implication attribute this belief to JJT. If she didn't think this (or so the implication seems to be) she wouldn't be so carefree about the ordering of the cases. We will see in a moment that JJT is far from carefree (or careless) about this ordering of the cases and lots of other contextual detail.

For the moment, though, let us pursue this chimera of decontextualised interpretation and the allegation that JJT holds to it. What on earth could this possibly be? Even in the most highly formalised of languages (maths and logic), it is accepted the precise meaning of a symbol or concatenation of symbols is fixed in part by their placement. The indexicality of meaning is a feature of any language, including formalised ones. No-one argues against the indexicality of meaning. However, unless LWAV can smuggle into the way they set up their experiment a shadow argument to the effect that meaning in language might be context free, their whole exercise makes no sense. Behind their set up, then, is the implication that meanings could be fixed and could be read off cases or examples irrespective of their context. This is the straw man to be knocked over by their experiment. Moreover, this position has to be attributed to JJT for the set up to work. Without both implications their experiment is pointless. All it could provide is yet another demonstration of what everyone accepts, namely that meanings are determined in context.

The second thing to say has to do with the suggestion that JJT's talk about what she feels, what others would say and so on is 'evidence' or is used as 'evidence' of some "ethical truth". But is this what is going on? When she says "Everybody to whom I have put this hypothetical case..." is she summarising a carefully constructed study with systematically gathered data on a par with LWAV's? JJT has not systematically or unsystematically (as Gilbert Ryle might say) gathered evidence. Her purpose is not to report findings and it is a complete misunderstanding of what she is up to to suppose that she is. In her argument, such expressions are more rhetorical handrails guiding us along and moving us forward than lists of 'facts' she has uncovered. LWAV's interpretation of JJT's expressions as 'evidence' is itself derived from their view that what is at stake in her paper is some version of ethical truth; that is, valid general propositions about what we should or should not do akin to valid general propositions about the physical world around us. But is this what JJT is trying to establish? Is she assembling data which she can test to determine what 'the facts' or 'the truth' are? Only an extremely strained reading can come to this conclusion. Time and time again, she reiterates her central purpose, namely the examination of the conditions under which we might find intentionally killing someone permissible and the extent to which that conclusion jibes with the principles set out by Philippa Foot. This is not about 'ethical truth' eternal, universal or otherwise. This is about the coherence of an argument in the face of different conditions to the ones under which it was set up. Foot's principles are not putative general laws or anything like them. They are guides for action not propositions, true or untrue.

What about ordering, meaning and interpretation? There are two things to say here. First, because the cases (standard, push and loop) are presented in such different ways, they are bound to be interpreted differently. There is consensus on this. The context in which LWAV present them is as an isolated set of choices about which participants are asked to give opinions. Because presenting them in this way would make the task extremely difficult (or at any rate scarcely controlled) they provide a set of instructions and explanations. This is their description of how

they set the context.

> *After being redirected to the test page, participants began by reading a general description of the study. They were instructed to read the stories carefully and to imagine the situations as best as they could. Participants were also informed about the estimated length of the survey (ten minutes), about the possibility of leaving the survey at any point, and about the fact that their data would be treated anonymously. Furthermore, they were told that they are not allowed to go back and change their answers. To control for this we recorded for each participant whether the back-button was pushed.*
>
> *According to the condition they had been assigned, participants were then presented with three (control condition) or five scenarios (test conditions). The scenarios were described using a short piece of text. To supplement the text, the scenarios were also accompanied by a diagram illustrating the situation (see figure A). These scenarios were specifically designed to be clear and straightforward and to contain no extraneous information.5 After each scenario participants were asked to indicate the degree to which they agreed or disagreed with a corresponding claim ..... Responses were made on a rating scale ranging from 1 (strongly disagree) to 6 (strongly agree). Agreement is read as agreement with the claim that it is permissible to do the action in question, which means that the higher the number the more inclined the participants are to hold that it is permissible to do the action in question. Additionally, participants had the opportunity to provide comments on each scenario. (Op. Cit. pp 664-65*

LWAV's scenarios differ in several quite important ways to JJT's. We have a personal name 'Abigail' to identify the agent pushing the button. What difference does this female personal identifier make to how her actions are construed? What general assumptions about how females make choices might participants draw on when judging these cases? Second, the person on the loop is an 'innocent bystander' not a worker. What difference to our intuitions about what is permissible is made when one has (we assume) as part one's working life an acceptance that one works in potentially dangerous situations (they know the dangers)? What difference does it make if they do not work in such an environment? Third, what does the mode of killing make a difference? Is throwing someone off a bridge 'the same' as pushing a button and moving a platform? Do we feel about it in the same way? What difference does the former's "up close and personal" character make to how we make judgements about it? The point is not we know they do make a difference, simply that we can plausibly imagine they might. And if we don't know, just how controlled is the context which the set-up is testing?

JJT sits her versions of the cases within a flow of arguments, each one being introduced to bring out or add a new point, an additional subtlety or some possible grounds for objection. We will track through the work she does to build her argument in a moment. The point we want to make here is that if LWAV think context makes a difference to meaning (presumably since they provide explanations and instructions, they will accept that there is more to context than simple ordering) then surely they will have to accept that any difference in the contexts in which the cases are presented might make a difference to the interpretations and conclusions we might come to over them. Putting it somewhat differently, given the very different ways in which they are situated, won't LWAV have to accept that what the standard, push and loop cases 'are' for the reader of JJT's paper is not the same for LWAV's subjects? And if they do accept this, how do they suggest we calibrate the different inferences and conclusions we might make? That (unsurprisingly) their participants' conclusions are different to JJT's is not testimony to the impossibility of establishing some ethical principles but to the differences in the ways in which the cases were set up. LWAV are not comparing like with like.

In contrast to the stark and stripped down versions of the cases which LWAV present, JJT's account is intricately and carefully designed, with telling detail specifically located at particular points to ensure the flow of the case she is making. The argument is designed for fellow philosophers and the cases shaped to fit the argument being unfolded. JJT's paper is a construction not a set of notes made while she was thinking through the issues. It required work, the work of carefully designing and building an argument to end with the outcomes it does. Of course, this work can be well or badly done. Like all constructions, it could be robust or weak. The point is that competence in building arguments like this is part of the practice of philosophy. Philosophers know it is done and they know how it is done. The argument is an artifice and its artefactual character readily recognised. One of central skills in building a successful philosophical argument is achieving argumentative momentum. The energy, or force, of the argument builds up as the pieces are put together. The ordering is precise and not random. To see what we mean by this, look at the following first level analytic content analysis of JJT's argument as she develops it as far as the Fat Man on the Bridge example.

1	Foot's Trolley Case	The standard version	'Everybody's' response is it is permissible to turn trolley
2	Foot's Surgeon Case	5 people need donated organs. A young fit man can be used as a source. Is it permissible to kill him to save the others?	'Everybody's' response is it is not permissible to kill young man
3	Foot's Question	Why is the first permissible and the second not?	
4	Foot's solution	I Killing 1 is worse than letting 5 die II Killing 5 is worse than killing 1	
5	JJT's rejection of Foot	Failing to operate is letting die not killing.	

6	JTT's Bystander case	Bystander can throw switch to re-direct trolley and kill 1 to save 5.	JTT's 'feeling' is that this is permissible
7	JTT's rationalisation	(a) Bystander has no official role with regard to the trolley and its consequences (b) Driver positively acts to run into the 5 or the 1. By refraining from acting the bystander does nothing	JTT 'feels' that Bystander *may* intervene (but is not required to do so)
8	JTT's analysis	Foot's principle I is inadequate to the trolley (and hence to the surgeon?) case	
9	JTT's elaboration of context of Surgeon Case	Surgeon has previously misprescribed the 5 and has caused their conditions. His prescription is going to kill them	JTT thinks it is 'plainly' not permissible to kill 1 to save 5
10	JTT's inference	Foot's I is true (killing 1 is worse than killing 5) but it does not tell us what to do in the Surgeon Case	
11	JTT's revision of Foot's II	II' When the choice is doing something *here and now* to the 5 or the 1, then II is permissible	
12	JTT's Summary	Simple reference to the difference between 'killing' and 'letting die' is too blunt. Need to look at the contexts in which decisions are being made	
13	JTT's introduction of a new principle	Observe the Kantian principle of always treating people as ends and not as means	
14	JTT's Loop Case	Introduced to show what she intends by treating people as means not ends. Workman on loop is Fat Man	Some people do not think this case is clear cut. Nonetheless, killing Fat Man seems right.
15	JTT's conclusion	Not clear that the distinction between treating people as 'means' and 'ends' would allow Surgeon Case and would not allow Fat Man case.	The key distinctions seem to be (a) Bystander moves the threat from 5 people to 1 (b) but does so by doing something that does not infringe that one person's rights
16	JTT's development	Discussion of rights and their application to the cases	Seems to come down to the management of harm and its distribution from 5 to 1 without infringing the latter's rights. But JTT is not sure why.
17	JTT's Fat Man on Bridge Case	Agent pushes Fat Man off bridge	Pushing Fat Man off bridge uses him as mean and infringes his rights and so is not permissible. As an act, turning Trolley by Bystander does not itself infringe rights and so is permissible

The core of JTT's argument is a puzzle over the difference between the Surgeon case and the Trolley case. Why does it seem permissible to turn the trolley but not to take the organs? Her view is that the intuitions we might have about Foot's original cases cannot be 'rationally' based on the distinction between 'killing' and 'letting die'. She introduces each of her variants in order to

move away step by step from Foot's encapsulation of principles based on this distinction towards an account based on the interweaving of actions in context and the possession of rights. The responses, characterised as her own or other people's, are set in the argumentational context of a gradual elaboration of the simple ('blunt') original distinction

It is important to recognise that this is not a matter of deciding that JTT's account of the cases is 'better' than LWAV's. It is, rather, that given the different ways the accounts are set up, they are *different cases.* The points being made about them are almost entirely different. LWAV treats the cases as self-standing stories to stimulate intuitions about the permissibility of killing 5 or 1. JTT treats them as indicating the importance of context and the inadequacy of simple distinctions. What they are 'about' for both is entirely different and, as a consequences, the intuitions felt about them are entirely different and non-comparable. The mapping between LWAV's summarised results and JTT's informal descriptions simply will not do.

The importance of context in the presentation of the two sets of cases is not restricted to the direct comparability of the 'intuitions' stimulated by them. Which intuitions are to be stimulated is set by the context too. The most obvious example of this is JTT's setting up of the first Fat Man case. This is not 'about' killing or letting die but about treating people only as 'means' rather than ends. The second Fat Man case (pushing off the bridge) carries both the connotations of the first *and* the connotations of the discussion of rights and their infringement which is placed immediately after the first case and immediately before this one. Pushing the Fat Man off the bridge is about using people as means and infringing their rights, not simply about choosing to kill 1 to save 5. What we find in the story (what JTT says she finds in the story) is precisely what she has put there to be found when constructing it. The story is not a plain case of a set of choices, it is a carefully considered exemplification of how the means/ends issue and the issues concerning rights might affect the choices under discussion. This is not manipulation; it is philosophical competence. It is the work of reasoning in philosophy (at least in this form of philosophising).

2 The Reasoning Context of the Cases

One argument proponents of EP often make is that the intuitions of ordinary people are not the same as the intuitions of philosophers. Ordinary people are not 'philosophical experts' and hence do not have the finely honed or developed sensibilities which such experts have. Although this is obviously true, it is somewhat misplaced. It is not that philosophers are experts on moral or similar questions and ordinary people are not, it is that they have different kinds of expertise and so have different kinds of intuition. This being the case, in contrast to LWAV's our questions would be, first, what are these different intuitions about? And second, if we wanted to isolate them, would the procedures of psychological experiments be the right way to do it?

To answer the first question, we would have to understand a great deal more about how common sense and philosophical reasoning are carried out and the different bodies of 'taken for

granted knowledge' each relies on — that is, the conventions and practices, routines and rules of thumb that are used in each. Their different interests, purposes and related practicalities shape each type of reasoning. Common sense reasoning is reasoning in the ordinary world of daily life. Philosophical reasoning is reasoning within and for a highly attuned and specialised community. The intuitions relied upon in undertaking either reflect these differences. This is not to say that they are totally unrelated (the one thing that is right about EP) simply that their relationship is complex. Philosophical ethical reasoning has its roots in common sense reasoning but is framed in distinctive ways and pursues different objectives. Common sense ethical reasoning takes place in the hurly burly of daily life and concerns matters such as 'Should I tell the sales assistant, he has undercharged me for that item?', 'How do we balance the responsibilities of an older and younger child when they both misbehave?', 'Does my use of a hosepipe during a water shortage make enough difference to count?' and so on. Only rarely (thankfully) is it about such weighty matters as choosing to kill one or five persons. As we have already said, the point is that these decisions are taken *in media* res and not in reflective mode. They are about what to do *now*. Philosophers consider their problems in an abstract and reflective mode. They are not about how they themselves should act (or, perhaps, only occasionally). Their concerns are not how would they and others feel if such and such action were to be undertaken, but rather can such an action be rationally justified and if so, how? Rational justification here means be defended by a course of argument. The expertise we (and this includes philosophers) have in making every day ethical judgments is of a different kind to the competences philosophers display in assembling their arguments to justify their conclusions. The intuitions relied upon might well be equivalent but they are not identical. If we talk about both as ethical reasoning, what does common sense ethical reasoning rely on and what professional philosophy? What are the taken for granted assumptions of both? Clearly LWAV's experiment takes us no closer to understanding this question and the set-up it uses makes the identification of each set of intuitions impossible.

To get a robust answer to our second question, we need to be fairly confident of the mapping between moral considerations as stimulated the experimental set up and the moral considerations that are relevant in daily life. Such confidence will allow us to deploy an operationalisation where both the situation(s) under examination and the intuitions invoked to make judgments on them are similar or similar enough. We have already explored the differences in context between the presentation of the cases and, on LWAV's own arguments, these differences seem great enough for the cases to be interpreted in markedly different ways. With regards to the similarity of intuitions, unless LWAV are operating with some form of cognitive innatism (we have a complete or exhaustive bundle of intuitions from which we select depending on the situation) which, from their commitment to contextualism we assume they aren't, then exactly what our intuitions might be about individual cases will vary (and sometimes wildly) from case to case (and perhaps person to person, as JTT suggests). Intuitions are not fixed. Rather what we think, and what we think we ought to think, are a matter of continuous interpretation in the context both of the situation we are in and of the one we are examining. This

is not to say that experiments can yield no insights. But what they can be insightful about are people's reasoning when they are in a highly constrained 'reflective mode' (to wit, the constraints in place on not changing their minds and altering responses etc.).

Operationalisation is one set of considerations. The commonality of meanings about 'experimental objects' such as cases, instructions and, most importantly, responses is another. Not only does there have to be reasonable mapping between the experiment and ordinary situations, there has to be isomorphism between the interpretations of the subject and the interpretations by the experimenter of those interpretations. Here of course all sorts of personal and biographical detail intrudes (which is why 'Abigail' and 'innocent bystander' might be important). This is not just about the meaning of words, but the significance of differences in the cases and, of course, any one's relative familiarity with this kind of problem. When counting, summarising, aggregating responses, we are assuming there is this commonality of interpretation both across the participants and between the participants and the experimenter. Given LWAV's contextualising position, how is this to be defended?

Practical reasoning in philosophy might be described as the construction of a trajectory based upon a set of presuppositions from an initial puzzle to an apparently acceptable solution, where each step appears reasonable and consistent with the presuppositions and the previous steps taken. The rationality of the trajectory is given by the coherence of the steps and presuppositions. Professional competence in forming and following such trajectories consists in the skills of designing and interpreting them. Professional practical reasoning of whatever type and ordinary practical reasoning are not identical although they do have much in common. Both depend upon the availability of 'intuitions' and 'common sense understandings' but these intuitions and understandings are not identical. When JJT or any other philosopher talks about the intuitions she or we might have, they are those of the professional practical reasoner not of the lay person reasoning about the same questions, though of course some of the considerations in play are in common. Professional practical reasoning uses common sense as a resource but not just common sense.

3 Philosophical Implications

So, what does all this mean for the LWAV instance of EP and its critique of JTT and her philosophical arguments? EP asserts that because our ordinary intuitions differ from those of philosophers, their arguments must be suspect or, at least, lack firm grounding. As we hope we have shown this is an unsubstantiated claim. In the LWAV case, we are faced with an unargued idealisation of the possibility of a decontextualised moral and ethical truth (as Nagel put it, ' the view from nowhere'). This idealisation is spurious. Instead of proposing that JTT and common sense have different 'viewpoints' on ethical truth depending on their contexts, we want to say that the 'truth', 'conviction', "certainty' that each might secure is worked through in context *and the contexts are worlds apart*. The one is centred on the need to make real decisions in real situations that matter. The other is centred on the detailed construction of a flow of coherent argument that

will satisfy and convince a professional community. The intuitions of the former do not bear upon the latter in any ways that make a philosophical difference.

5 SUMMARY

EP is nothing if not self-confident. It believes it has identified a fundamental weakness in the practise of contemporary philosophy and has strenuously promoted a favoured remedy, namely the adoption of experimental and survey investigative techniques used in the psychological and social sciences. We have not been concerned to assess what difference such adoption might make to the nature of philosophical discussions and whether it would bear at all on the philosophical issues being raised and discussed. We have simply confined ourselves to asking if the methods proposed are likely to do the work envisaged. Our conclusion is that things look muddled. The invocation of experimental or survey methods do not, of themselves, clarify much, if anything. They seem replete with problems of formulation and framing as well as relying on a somewhat underpowered inferential strategy. We are not saying that it is not possible to formulate a robust EP, only that what is on offer at the moment won't do. Whether when revised and strengthened, the resulting EP will throw any light on the perplexing problems taken up by philosophy is not something which has concerned us in this discussion. All we can say from the arguments we have made is that as currently promoted, it has no hope of doing so.

References

Bond, T & Fox, C.	2001	Applying the Rasch Model: Fundamental measurement in the Human Sciences. Mahwah. Lawrence Erlbaum.
Couper, M.	2000	Web surveys. Public Opinion Quarterly, vol 64, pp 464 - 494
Cullen, S.	2010	Survey driven Romanticism. Review of Philosophical Psychology, vol. 1, pp 275-296
Dilman, D & Bowker, D.	2002	The web questionnaire challenge to survey methodologies, in Batinic, B, Reips, U-D, Bosnjak, M (eds) Online Social Sciences. Ashlan, Hogrefe & Huber pp 53 - 74.
Druckman,J & Kam, C.	2009	Students as experimental participants. Institute of Policy Research Working Paper WP-09-05
Foot, P.	1967	The problem of abortion and the doctrine of double effect. Oxford Review, Number 5, 1967
Gosling, S, Vazire, S, Srivastava,S & John, O.	2004	Should we trust web-based studies? American Psychologist vol 9 no 2, pp93 - 104.
Kauppinen, A.	2007	The rise and fall of Experimental Philosophy. Philosophical Explorations. Vol 10, issue 2, pp 95-118
Liao, S, Weigman, A., Alexander, J. & Vong, G.	2012	Putting the trolley in order: Experimental philosophy and the loop case. Philosophical Psychology, vol 25, issue 5 pp 661 - 671
Machery, E. & Stich, S.	2012	The role of experiment in the philosophy of Language in Russell, G. & Della Graff F. Routledge Companion to Philosophy of Language. London. Routledge pp 495 - 513.
Michell, J.	2004	Measurement in Psychology. Cambridge. CUP
Nichols, S & Knobe, J.	2008	Moral responsibility and determinism in J. Knobe & S. Nichols, Experimental Philosophy. Oxford OUP pp105 - 128
Peterson, R.	2001	On the use of college students in social science research. Journal of Consumer Research, vol 28, no 3, pp 450 - 461
Sears, D	1986	College sophomores in the laboratory. Journal of Personality and Social Psychology, vol 51, no 3 pp 515 - 530
Thomson, J. J.	1985	The trolley problem. The Yale Law Journal, vol 94, pp 139-1415
Thomson, J. J.	2008	Turning the trolley. Philosophy and Public ffairs vol 36, no 6, pp 359 - 374
Weinberg, J, Nichols, S. & Stich, S.	2006	Normativity and epistemic intuitions, in Viale, R., Andler, D & Hirschfield, L. Biological and Cultural Bases of Human Inference, Mahwah. Lawrence Erlbaum