Contemporary Debates in Epistemology, Volume 3

Chapter Question: How to effectively use thought experiments in epistemology?

A Guide to Thought Experiments in Epistemology

Wesley Buckwalter

Department of Philosophy; Institute for Philosophy and Public Policy, George Mason University

wesleybuckwalter@gmail.com

Abstract: The purpose of this chapter is to provide a guide for conducting thought experiments in epistemology effectively. The guide raises several considerations for best practices when using this research method. Several weaknesses in the way thought experiments are conducted are also identified and several suggestions are reviewed for how to improve them. Training in these research techniques promotes more productive scholarship in epistemology, saves time and resources wasted on less efficient approaches, and reduces the risk that researchers are fooling themselves when they use thought experiments in philosophy.

Keywords: methods, intuition, judgement, social cognition, knowledge

> *The first principle is that you must not fool yourself and you are the easiest person to fool.* - Richard Feynman

In her column on NPR's Cosmos & Culture, Tania Lombrozo observes that there is something deeply ironic about the success of experimental philosophy (Lombrozo 2013). Though experimental philosophers use many methods, one popular way to conduct experimental philosophy is through vignette-based research. This involves presenting narrative thought experiments to participants and collecting their judgments about them, rather than evaluating those thought experiments from the armchair alone. Quite often, the thought experiments used in behavioural experiments are ones adapted from the traditional philosophical literature. The irony of this, Lombrozo writes, is that many of the results of these behavioral experiments probably could have been discovered from the armchair. That is, many studies use thought experiments that "involve compelling empirical phenomena that people can readily appreciate" and that the reactions people give to them "can be elicited by simply presenting people with the relevant vignettes in a short article" (ibid). If this is true, then it suggests that thought experimentation could sometimes be one effective approach to philosophical inquiry. The key words there being "sometimes" and "could". So, when are thought experiments used effectively in epistemology?

While an enormous amount of scholarship in philosophy is dedicated to almost all the theoretical aspects of thought experiments one could think of, relatively little of it addresses the practical and applied questions of how to design, conduct, evaluate, or interpret thought experiments when conducting philosophical research. Neither do philosophy students typically receive much direct training in the method of conducting thought experiments. Often students learn the method through osmosis. But we should not rest content without a more critical examination of research practices. Without dedicated scholarship and training in philosophical research methods, the use of thought experiments in epistemology is bound to remain only partially effective. And whatever insights or progress philosophers hope to achieve with the use of thought experiments, all can agree that if thought experimentation is worth doing, then it is at least worth doing well (Turri 2016b: 55).

The purpose of this chapter is to provide a guide for conducting thought experiments in epistemology effectively. The guide raises several considerations for best practices when using this research method. Several weaknesses in the way that thought experiments are conducted are also identified and several suggestions are reviewed for how to improve upon them. Applying these research techniques promotes more productive scholarship in epistemology, saves time and resources wasted on less efficient approaches, and reduces the risk that researchers are fooling themselves when they use thought experiments in philosophy.

***

Thought experimentation is a distinctive method in philosophy and plays a substantive role in several foundational debates in epistemology (Machery 2017; Stich and Tobia 2016). It is sometimes also colloquially referred to in philosophy as the "case method" or "philosophical intuition pumping". In a typical episode, the method involves three parts. In the first part, or the design phase, researchers conceive of and construct narrative vignettes about real or imagined scenarios of antecedent philosophical interest. In the second part, or the result phase, researchers report their judgments about the scenarios that they or others have designed. In the third part, or the discussion phase, researchers claim that these judgments reveal something new about the philosophical categories or phenomena that they are researching.

There is perhaps no better illustration of this method in epistemology over the last several decades than the broad body of research investigating the connection between knowledge and luck. This connection has been studied using countless thought experiments (for discussion, see Blouw, Buckwalter, and Turri 2018; Gendler and Hawthorne 2005; Turri 2016a). In “fake barn” cases, for instance, philosophers ask us to imagine that a protagonist sees a real barn while driving though a strange area otherwise filled with fake barn façades (Goldman 1976). When imagining such a context, many philosophers report the judgment that the protagonist in this situation does not know that they saw a real barn (Pritchard 2005; Sosa 1991). From this result, several philosophers have argued that this judgment shows various things must be true about the nature of knowledge, perception, or possibility (Chisholm 1989; Goldman 1976).

Thought experimentation is incredibly versatile and allows for large researcher degrees of freedom. Philosophers use the method to explore many variables across many different contexts to meet their needs in accordance with their research goals. With thought experiments, it is sometimes thought, researchers are limited only by their own imaginations. Because of this flexibility, it is extremely important to begin research projects by specifying these goals concretely. The first step in applying the method should be to identify a testable research question and to reflect on how the use of thought experimentation will test that question. In fake barn cases, for example, there are several plausible research questions that researchers could have otherwise developed such a thought experiment to test:

Does knowledge include an anti-luck provision?

Is perceptual knowledge sensitive to a certain number of defeaters?

Is seeing something sufficient for knowing it?

Are relevant alternatives important for when we are inclined to attribute knowledge?

Is the truth of "knowledge" sentences sensitive to relevant alternatives?

What does common sense say about knowledge and luck?

Though there is overlap between many of these research questions, there are also several subtle and potentially important differences. For example, a research project on knowledge attribution might benefit from different materials and analysis than research designed to study the nature of knowledge or perception. If one is interested in studying anti-luck previsions, a researcher might ask themselves, then would it be wise to consider only cases that so heavily feature perception? The more pointedly and concretely that a researcher can frame their central research question, the better they will design a case to test it, and the more suitable that thought experiment will be to help answer their research question.

As a corollary to this, researchers must also specify the kind of thing they intend thought experimentation to ultimately show and align this with their central research question. Though prior research has tended to focus on whether judgments in thought experiments constitute evidence or not, there are many goals that philosophers can have when conducting thought experiments. For example, they might use them to:

Motivate a theory

Explain how something works

Engineer a new concept

Call attention to an area of promising inquiry

Isolate a theoretically useful factor

Evaluate a necessary or sufficient condition

Propose a new or overlooked distinction

Test a theory by giving a case in which it could yield the wrong predictions

Identify a central tendency in the natural or social world

Reveal a position as common sense or a behavior as typical

Classify something as ungrammatical, abnormal, or otherwise normatively deficient

While there is surely overlap between many of these goals, registering them at the outset directs and clarifies the design of the resulting thought experiment. For example, this might lead a researcher to develop materials that prioritize certain narrative contexts over others or to isolate key variables of interest and minimize incidental details given their topic of interest. Distinguishing between these things also helps researchers to ask better questions and give judgments about them with greater specificity and precision. And the clarity that results from this exercise helps researchers to ensure that the conclusions that they draw from such judgments are warranted and well supported. Revealing a position as common sensical has different success conditions than proposing a theoretical distinction. Identifying a central tendency in social cognition requires different evidence than evaluating a necessary condition does. And we might think that calling attention to new areas of inquiry requires much less than falsifying a theory.

With central research questions and goals specified, it is now time to begin designing cases that answer them. Though case construction will vary with the details of the research project, all researchers should approach this process with extreme care, skepticism, and caution. The reasons for this are due to both the details of the method and the social structure of the activity. With respect to the method, researchers must bear in mind that they are ultimately the ones who both design the thought experiment and evaluate their own success at conducting it, with very little external oversight throughout this process. This increases the risk that results are due to the task the researcher has devised rather than a true test of the central research question (Rosenthal 1976). With respect to the activity, as is also well known in science, career and publication incentives can subtly influence research findings and behavior (Smaldino and McElreath 2016). In short, researchers have a stake in the outcome of thought experiments and are highly incentivised to create thought experiments to get the answers they want. Together, these factors create ample opportunity for researchers to inadvertently construct thought experiments and evaluate them in ways that do not accurately track the philosophical phenomena they are investigating.

Because of this risk, every aspect of the case that is introduced by the researcher should be examined with extreme prejudice. It is well known that the presentation effects regarding the way that thought experiments are framed, ordered, and worded can affect the way they are evaluated (Machery 2017; Schwitzgebel and Cushman 2015; Tobia, Buckwalter, and Stich 2013; Tobia, Chapman, and Stich 2013). But more generally speaking, vigilance is important when it comes to many features of thought experiments, such as:

Length

Complexity

Familiarity

Naming

Order of details presented

Subject matter and narrative framing

Naturalness and plausibility

Outlandish or esoteric situations

Emotional or moral content

Greater cultural context

Owing to the central research question under investigation, one or more of these things are often utilized in the construction of philosophical thought experiments. That is not necessarily a bad thing. But there are trade-offs to consider. The more control that researchers exert over thought experiment design, the more closely they may be able to test variables of interest. But the more control that is exercised over thought experimental conditions, the greater the threats to its validity. The more things researchers shove into thought experiments, the less likely that their results are trustworthy, generalize to other contexts, or to real-life settings. As these threats to the validity of thought experiments increase, we should decrease our confidence in the results. Philosophers must evaluate these trade-offs to determine when they allow for the full testing of their research question while minimizing the risk that the components listed above are not driving their judgments over and above the philosophical content of interest.

There are several ways to optimize these trade-offs. Some pertain to even incredibly basic aspects of thought experiments. Consider, for example, the naming of stimuli persons. Research in experimental cognitive science shows that naming can have complex effects on many perceptions and social judgments (Bertrand and Mullainathan 2004; Kasof 1993; Laham, Koval, and Alter 2012; McKelvie and Waterhouse 2005; Newman, Tan, Caldwell, Duff, and Winer 2018). As researchers have wisely observed, "experimental design, at the most abstract level, is an exercise in variance control" and names are information that can create variance (Newman, Tan, Caldwell, Duff, and Winer 2018: 1445). For example, names can impact perceptions of age, race, socioeconomic status, warmth, intelligence, likeability, or ethicalness. Such associations are important to consider, especially when they could be relevant to evaluating the central research question. This is also important when designing thought experiments but is often not taken very seriously in philosophy. For example, consider classic cases in epistemology that literally name characters "Mr. Nogot" and "Mr. Havit" in thought experiments designed to evaluate whether someone else has knowledge or not (Lehrer 1965). As a general rule, it is probably not a very good idea to name characters in ways that relate so closely to judgements or concepts that cases are meant to reveal. Maybe this doesn't ultimately make a difference in this instance, but we can do better. Moreover, as researchers have also observed, common names used in anglophone philosophy (e.g. "Smith", "Jones", "Bob") tend to signal ethnic expectations and may be perceived differently across cultures (Schwitzgebel 2015). The implications of this also warrant serious reflection.

To do better, the overarching principles of thought experimentation should be to minimize unnecessary complexity and length, avoid needlessly esoteric or stilted situations, and to maximize believability and naturalness. When this is not possible, researchers should consider designing thought experiments using multiple cover stories to ensure that their judgments are targeting underlying philosophical phenomena and not inadvertently driven by incidental details of the task they have created. And when complex thought experiments involve complex steps or multiple stipulations, researchers should ensure that they are processing these things and holding them fixed as they make their judgments. Doing better may also involve attending to small details and apprising oneself of any relevant psychological literature as the need arises. For example, it is worth giving more consideration to naming choices given possible associations or interference with them when processing names (see Newman, Tan, Caldwell, Duff, and Winer 2018 for perceptions of several common names). Without attending to such details, the risk of interference with thought experimentation increases.

A concrete illustration of this risk has been demonstrated in recent research on free will judgments. Thought experiments in the free will literature often solicit judgments about concepts familiar to us from ordinary life, such as freedom and moral responsibility. But the thought experiments often involve very complex, imaginative situations that depart significantly from the contexts in which these concepts are regularly used. For example, many thought experiments stipulate that determinism is true to test what judgments are compatible or incompatible with it in various circumstances. But researchers have demonstrated that this stipulation is often rejected when cases are processed. Specifically, researchers have demonstrated that everyday indeterministic metaphysics intrudes on the way we process deterministic thought experiments (Nadelhoffer, Rose, Buckwalter, and Nichols 2020; Rose, Buckwalter, and Nichols 2017). Despite what the thought experiment stipulates about an all-knowing supercomputer predicting that Jeremy will rob the bank, we think it’s still possible that he won’t do it. So, despite what the thought experiment stipulates, judgments from such thought experiments simply cannot tell us whether something is being judged compatible or incompatible with determinism. Failing to attend to these facts leads to confusion and misrepresentation. In this case, it also obscures something potentially interesting about possibility and responsibility.

The example above suggests several lessons about case construction. Stipulating core features of cases can sometimes aid thought experimentation. But if stipulations of thought experiments are not actually processed or accepted, then they may ultimately be useless at answering the central research question they were designed to address. It is reasonable to suppose that stipulations are more difficult to process as cases increase in complexity and novelty, or when there is a significant mismatch between imaginative contexts and the circumstances in which we normally apply related concepts. In this circumstance, researchers might profitably draw inspiration from cases from real life situations more likely to mimic the context in which the relevant concepts are typically used (Furman 2021). Subsequently, free will researchers have worked to create narrative materials featuring simpler and more straightforward situations from everyday life that are less likely to trigger our indeterminist tendencies, and have developed questions attempting to detect and guard against this when it occurs (Turri 2017a). Applying this lesson to epistemology, researchers should consider this risk careful when developing cases that depart significantly from everyday situations in which ordinary concepts of like knowledge or belief hold sway. And just because something is stipulated doesn't entail that it's being tested.

The features above also heighten the need for the use of controls and tight comparisons in the construction and evaluation of thought experiments. Take for example, the growing body of research in epistemology on deference and testimonial asymmetry. Deference is a belief formation process that occurs when one believes something in virtue of the fact that someone else believes it. Many philosophers have argued that something "feels fishy" about deference to moral testimony as opposed to non-moral testimony (Enoch 2014: 237; McGrath 2011; Williams 1995). One of the ways this feeling is brought to bear is through comparisons between cases. For instance, researchers have claimed that "there is something off-putting about the idea of arriving at one's moral views by simply deferring to an expert" there is "no problem with deferring to a tax specialist about one's taxes" (McGrath 2011: 111), and have concluded that "moral deference seems more problematic than deference in many other domains" (ibid, 115). Similarly, other researchers compare cases involving deference to non-moral testimony, such as asking a friend who won the 1994 world cup, to those involving deference to moral testimony, such as asking a friend whether you ought to support a military intervention in Syria (Hazlett 2017).

Comparing judgments about deference cases between domains seems like a sensible starting point for diagnosing reactions to deference and the domain specific features that make it seem problematic. But the cases above also differ in lots of other ways too that should also be considered and scrutinized. For instance, they differ not just in whether they involve a moral proposition, but also in terms of how serious they are, what is at stake, their political and legal content, whether they involve future or past events, what country they take place in, and so on. Maybe these things matter and maybe they don't. But the fewer the differences between cases, the greater our confidence can be that our judgments about them track the intended features of cases and that these features test our central research question. This reveals what is often meant by the need for "matched pair comparisons" in thought experimentation. The basic idea is that thought experiments should manipulate only the independent variables of interest when making judgments about multiple versions or conditions. If moral content is the variable of interest above, for example, then ideally, the cases devised to reveal its role in our judgments should be similar in all other ways apart from their moral content.

Sharpening manipulations in this way also helps us to develop more pointed tests. We might also think that research on this question would benefit from specifying a more fine-grained dependent variable between matched pair conditions, or the questions we ask ourselves about the thought experiments. For instance, the testimonial asymmetry cases above have profitably alerted philosophers that "something fishy" could be going on when comparing deference in different domains. But moving beyond this initial reaction that something seems fishy, it would be useful to understand the underlying judgment a little better before theorizing about what that judgment means for the relevant philosophical categories. To do this, researchers should ask themselves several questions about the judgments that are being elicited. For example, is the judgment that deference to a friend about military intervention in Syria is imprudent, immoral, unconventional, unoriginal…or something else? How strong is that judgment? Does it come in degrees? Does it depend on how the question is phrased?

Putting this all together, the suggestion is that researchers should articulate the question they are asking about cases and their answers in as much detail as possible when they present thought experiments. Underspecified cases and vague reporting detract from effective thought experimentation. Instead, researchers might subject cases to a barrage of well-formed and detailed questions, using different phrasing and framing, to pinpoint the judgment of interest. Utilizing some of these techniques, progress has since been made in understanding our reactions to deference (Andow 2018). This progress illustrates how a lot of time can be saved at the outset by constructing and evaluating cases using well controlled comparisons and explicitly communicating variables to test them.

Without controlled comparisons, confounds are likely to arise in thought experimentation that undermine philosophical research. One illustration of this has been documented in bank cases, which have been instrumental in motivating epistemic contextualism. The bank cases are pairs of cases used to elicit the judgment that "knowledge" sentences can sometimes have different truth values depending on contextual features of the situation such as on what is at stake, thereby motivating the theory that "knows" is a contextually sensitive expression (DeRose 1992, 2009). In order to show this, researchers manipulate how much is at stake in pairs of cases and then claim that this difference impacts what seems true about "knowledge" sentences. Importantly, however, this was not the only difference between classic bank cases. The cases also differ in the self-regarding knowledge statement made by their protagonists, which is not of central theoretical interest. Specifically, in the low stakes bank case the protagonist claims, "I know," while the protagonist of the high stakes case claimed, "I don't know". So, judgments about the case could be due to contextual features like stakes, but they could also be due to this difference in what the protagonist said about themselves independently of stakes. When researchers put these cases to the test, they found that this confound was likely responsible for the predicted pattern of results, and not the contextually relevant features identified epistemic contextualists (Turri 2017c). In these thought experiments, at least, it seems that people are responding to what a protagonist happens to claim about their own mental states, and not what is at stake for the protagonist in low or high stakes contexts.

To address this worry, philosophers should consider susceptibility to confounders when they design and evaluate thought experiments (for discussion, see Dafoe, Zhang, and Caughey 2015, 2018; Skelly, Dettori, and Brodt 2012; Steiner, Atzmüller, and Su 2016). This might involve listing and reporting all potential confounders at the outset of the design phase, attempting to adjust from them when evaluating thought experiments, or adjusting judgments in light of them. Philosophers may also lessen the risk of confounding variables through design choices such as shortening and tightening the materials to minimize extraneous details or unmatched comparisons. And philosophers might also appreciate the wisdom in failure. If, in the process of designing thought experiments, several drafts of the case do not seem to work as well as others do, researchers should ask themselves why. What did you have to do to a case to get the thought experiment to yield the judgment that it did and what is the significance of that change?

Thought experimentation should also utilize control conditions. In the case of fake barns, for instance, philosophers might pair judgments about the classic fake barn case in which knowledge is in question against the closest adaptations possible in which knowledge is unambiguously present or absent. And if the research question involves defeaters, researchers could consider manipulating the nature or number of fake barns in comparisons to cases without fake barns. Doing these things has led to significant progress in understanding these judgments (Colaco, Buckwalter, Stich, and Machery 2014; Turri 2017b). As it happens, researchers have discovered that mentioning nearby fakes in fake barn cases actually seems to prevent iterated error from affecting knowledge attribution!

Even when devising optimal thought experiments free of confounds, we do not always know where our judgments about them come from. The resulting debate following bank cases and the role of stakes in knowledge attribution illustrates this basic observation (see also Buckwalter 2021 on error salience). Assume for a moment that stakes really do impact knowledge judgments in bank cases. One interpretation of this finding is that practical factors are a component of knowledge and that this should be included in a new definition of knowledge (Stanley 2005). But this interpretation assumes something about the psychology of the judgment made in the thought experiment. Specifically, it assumes that the judgment cannot be explained by judgments about other factors already included in the definition of knowledge. For example, suppose that the presence of stakes impacts our evaluation of how strong the evidence is, what the protagonist happens to believe, or what the evaluator of the thought experiment herself regards as true about it (Bach 2005; Weatherson 2005). If it is widely agreed that knowledge depends on these things, and knowledge is affected only as a result of belief, evidence, or truth judgments shifting, then knowledge judgments in these thought experiments do not motivate the new definition of knowledge. Thus, it is extremely important for researchers not only to give a clear judgment about a case free of confounds, but to also tell the difference between the mechanisms underlying their judgments if they are to understand their philosophical significance. In practice it is very difficult to tell this.

Understating the mechanisms underlying reactions to thought experiments is complicated by several factors. In the case above, for example, understanding the mechanism involves isolating and evaluating the individual and combined impact of simply too many variables on knowledge judgments. For example, a researcher would need to individually evaluate the impact of stakes, importance, truth, evidence, justification, and belief judgments on knowledge judgments, then consider all the possible interactions that are possible between these six things (Turri and Buckwalter 2017). And of course, these are just some of the variables that have been identified as theoretically relevant to the philosophical debate at hand. They clearly do not exhaust the possible mechanisms that underlie knowledge judgments. It is unclear that even the most astute introspection and evaluation of thought experiments could isolate the causal factors that influence so many judgments in this case with much confidence.

These observations suggest extreme caution and humility in the interpretation of thought experiments. Researchers should avoid strong interpretations of findings from thought experiments in situations where multiple competing factors of interest arise. This concern is compounded by the fact that philosophers often study closely related concepts and that questions being asked about them are often underspecified. Because of this, researchers should rigorously assess the possibility that their judgments are caused by factors that would undermine their favored interpretation of them. When the possibility of this is great enough, researchers should consider suspending judgment or re-evaluating the role that this judgment plays in their philosophical research with open eyes. In such a situation, research might also turn to established tools of cognitive science that have been developed specifically for this purpose, such as mediation and causal analysis. Such tools have been profitably applied to this and other research questions and have significantly advanced our understanding of the role of stakes in knowledge attribution (Turri and Buckwalter 2017; Turri, Buckwalter, and Rose 2016).

Judgments in thought experiments may also be impacted by features of the researchers themselves, such as culture, native language, gender, age, personality, and so on. This research is still in its infancy, and we currently do not fully understand the full effects of demographic variables on thought experimentation. According to some researchers, for example, philosophical intuitions are surprisingly robust across demographic differences (Knobe 2019). According to analyses done by other researchers, however, more than 90 studies to date involving over 75,000 participants in experimental cognitive science have reported demographic variation in judgments to philosophical cases (Stich and Machery 2022). This debate about the frequency and degree to which philosophical intuitions vary notwithstanding, the possibility that demography can moderate judgments when using thought experiments cannot be ignored. Thankfully, this possibility presents armchair philosophy with an opportunity to embrace several sensible reforms that extend to the use of thought experiments. For example, it is widely agreed that increasing demographic variation in training, publishing, and hiring is essential for the equitable future of the field. Doing so would also increase representation in thought experimentation by individual researchers. Additionally, thought experimentation can be conducted using group inquiry among teams of researchers as opposed to solitary inquiry, with greater diversity among members of research teams. This would decrease the risk that judgments are merely tracking the opinions of any one person or social group, and itself may increase diversity in the field (Buckwalter and Turri 2016). And of course, philosophers could always turn to the methods of experimental cognitive science to continue to evaluate the degree to which reactions to thought experiments are universal or are impacted by who we are or where we come from.

Lastly, foundational thought experimenters may be impacted in interesting and unforeseen ways by the training and experiences of researchers. For example, several studies suggest that philosophers have unique ideas about the connection between knowledge and luck. Several researchers have found that ordinary people attribute knowledge differently than many philosophers say they do in fake barn cases (Colaco, Buckwalter, Stich, and Machery 2014; Turri 2017b). This has also been found in other sorts of Gettier cases characterized by researchers as "authentic evidence" cases. For example, when researchers presented such Gettier cases to professional academics across seven fields in the humanities, social, and natural sciences, they found evidence that the non-philosophers attribute knowledge very differently than philosophers do (Starmans and Friedman 2020). Beyond knowledge and luck, further questions about the generalizability of philosophical judgments arise in bank cases (Rose et al. 2019), as well as other principles many philosophers have found intuitive, such as reliablism (Turri 2016d), the truth insensitivity of epistemic justification (Turri 2016c), and the truth condition on knowledge (Buckwalter and Turri 2020).

There could be several explanations for these differences relevant to effective thought experimentation. It is possible that these differences in reactions to thought experiments like Gettier, fake barn, or bank cases reflect philosophical expertise with the relevant philosophical phenomenon. If this is true, then the natural next step is to try to better understand these reactions to improve thought experimentation in the future. As of present writing however, empirical efforts to demonstrate which factors explain superior performance in thought experiments have largely been unsuccessful (e.g. Kneer, Colaço, Alexander, and Machery 2022). It is also possible that differences reflect field-wide biases that are insular to the field of professional philosophy, reflect theoretical biases, or other esoteric aspects of academic training in philosophy. However, it is unclear at present writing what these insular aspects could be. In either case, philosophers will benefit from studying these differences to better understand and improve their research practices. Until then, a sensible approach to thought experimentation includes discussion or collaboration with those both within and outside the field.

***

Summarizing the considerations above, the following checklist promotes effective thought experimentation in epistemology:

- ☐ Identify a central research question and determine how the thought experiment will test it.
- ☐ Devise cases using best practices that include controls and matched-pair comparisons, while minimizing extraneous length or details.
- ☐ Consider trade-offs in thought experiment design to improve testing and validity.
- ☐ Articulate questions fully and with as much specificity as needed.
- ☐ Assess likely confounds, incentives, potential sources of demographic variation, or field-specific biases.
- ☐ Exercise caution in the interpretation of case judgments and scrutinize judgments for other possible explanations than those preferred by the researcher.
- ☐ Question generalizability through discussion, collaboration, or experimentation.

Doing these things improves the quality of research in epistemology, decreases the time wasted on false starts or spurious debates, and increases the likelihood that the judgments made about thought experiments are philosophically significant.

Acknowledgements – For helpful feedback, discussion, and comments on previous drafts, I thank Carolyn Buckwalter, Elijah Chudnoff, David Colaco, and John Turri.

References

Andow, James (2018), 'Aesthetic Testimony and Experimental Philosophy', in Florian Cova and Sébastien Réhault (eds.), *Advances in Experimental Philosophy of Aesthetics* (Advances in Experimental Philosophy: Bloomsbury Academic).

Bach, Kent (2005), 'The Emperor's New 'Knows'', in Gerhard Preyer and Georg Peter (eds.), *Contextualism in Philosophy: Knowledge, Meaning, and Truth* (Oxford University Press), 51--89.

Bertrand, Marianne and Mullainathan, Sendhil (2004), 'Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination', *American Economic Review,* 94 (4), 991-1013.

Blouw, Peter, Buckwalter, Wesley, and Turri, John (2018), 'Gettier Cases: A Taxonomy', in R. Borges, C. de Almeida, and P. Klein (eds.), *Explaining Knowledge: New Essays on the Gettier Problem* (Oxford: Oxford University Press), 242–52.

Buckwalter, W. and Turri, J. (2020), 'Knowledge, Adequacy, and Approximate Truth', *Conscious Cogn,* 83, 102950.

Buckwalter, Wesley (2021), 'Error Possibility, Contextualism, and Bias', *Synthese,* 198, 2413–26.

Buckwalter, Wesley and Turri, John (2016), 'Perceived Weaknesses of Philosophical Inquiry: A Comparison to Psychology', *Philosophia,* 44 (1), 33-52.

Chisholm, Roderick (1989), *Theory of Knowledge* (Englewood Cliffs, New Jersey: Prentice Hall).

Colaco, David, Buckwalter, Wesley, Stich, Stephen, and Machery, Edouard (2014), 'Epistemic Intuitions in Fake-Barn Thought Experiments', *Episteme,* 11 (2), 199-212.

Dafoe, Allan, Zhang, Baobao, and Caughey, Devin (2015), 'Confounding in Survey Experiments', *Annual Meeting of The Society for Political Methodology* (University of Rochester, Rochester, NY).

Dafoe, Allan, Zhang, Baobao, and Caughey, Devin (2018), 'Information Equivalence in Survey Experiments', *Political Analysis,* 26 (4), 399-416.

DeRose, Keith (1992), 'Contextualism and Knowledge Attributions', *Philosophy and Phenomenological Research,* 52 (4), 913-29.

DeRose, Keith (2009), *The Case for Contextualism* (Oxford: Oxford University Press).

Enoch, David (2014), 'A Defense of Moral Deference', *Journal of Philosophy,* 111 (5), 229-58.

Furman, Katherine (2021), 'What Use Are Real-World Cases for Philosophers?', *Ergo,* 7.

Gendler, Tamar Szabó and Hawthorne, John (2005), 'The Real Guide to Fake Barns: A Catalogue of Gifts for Your Epistemic Enemies', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* 124 (3), 331-52.

Goldman, Alvin I (1976), 'Discrimination and Perceptual Knowledge', *Journal of Philosophy,* 73 (20), 771-91.

Hazlett, Allan (2017), 'Towards Social Accounts of Testimonial Asymmetries', *Nous,* 51 (1), 49-73.

Kasof, J. (1993), 'Sex Bias in the Naming of Stimulus Persons', *Psychol Bull,* 113 (1), 140-63.

Kneer, Markus, Colaço, David, Alexander, Joshua, and Machery, Edouard (2022), 'On Second Thought: Reflections on the Reflection Defense', in T. Lombrozo, J. Knobe, and S. Nichols (eds.), *Oxford Studies of Experimental Philosophy* (Oxford: Oxford University Press), 257-96.

Knobe, Joshua (2019), 'Philosophical Intuitions Are Surprisingly Robust across Demographic Differences', *Epistemology & Philosophy of Science,* 56 (2), 29-36.

Laham, Simon M., Koval, Peter, and Alter, Adam L. (2012), 'The Name-Pronunciation Effect: Why People Like Mr. Smith More Than Mr. Colquhoun', *Journal of Experimental Social Psychology,* 48 (3), 752-56.

Lehrer, K (1965), 'Knowledge, Truth and Evidence', *Analysis,* 25 (5), 168-75.

Lombrozo, Tania (2021), 'The Ironic Success of Experimental Philosophy', *Cosmos & Culture* <https://www.npr.org/sections/13.7/2013/03/23/175145568/the-ironic-success-of-experimental-philosophy>, accessed August 3, 2021.

Machery, Edouard (2017), *Philosophy within Its Proper Bounds* (Oxford: Oxford University Press).

McGrath, Sarah (2011), 'Skepticism About Moral Expertise as a Puzzle for Moral Realism', *Journal of Philosophy,* 108 (3), 111-37.

McKelvie, S. J. and Waterhouse, K. (2005), 'Impressions of People with Gender-Ambiguous Male or Female First Names', *Percept Mot Skills,* 101 (2), 339-44.

Nadelhoffer, Thomas, Rose, David, Buckwalter, Wesley, and Nichols, Shaun (2020), 'Natural Compatibilism, Indeterminism, and Intrusive Metaphysics', *Cognitive Science,* 44 (8), e12873.

Newman, L. S., Tan, M., Caldwell, T. L., Duff, K. J., and Winer, E. S. (2018), 'Name Norms: A Guide to Casting Your Next Experiment', *Pers Soc Psychol Bull,* 44 (10), 1435-48.

Pritchard, Duncan (2005), *Epistemic Luck* (New York : Oxford University Press).

Rose, David, Buckwalter, Wesley, and Nichols, Shaun (2017), 'Neuroscientific Prediction and the Intrusion of Intuitive Metaphysics', *Cognitive Science,* 41 (2), 482–502.

Rose, David, et al. (2019), 'Nothing at Stake in Knowledge', *Noûs,* 53 (1), 224-47.

Rosenthal, Robert (1976), *Experimenter Effects in Behavioural Research* (New York: Irvington Publishers, Inc.).

Schwitzgebel, Eric (2015), 'Names in Philosophical Examples'. <https://schwitzsplinters.blogspot.com/2015/11/names-in-philosophical-examples.html>, accessed June 10, 2022.

Schwitzgebel, Eric and Cushman, Fiery (2015), 'Philosophers' Biased Judgments Persist Despite Training, Expertise and Reflection', *Cognition,* 141, 127–37.

Skelly, Andrea C., Dettori, Joseph R., and Brodt, Erika D. (2012), 'Assessing Bias: The Importance of Considering Confounding', *Evidence-based spine-care journal,* 3 (1), 9-12.

Smaldino, Paul E. and McElreath, Richard (2016), 'The Natural Selection of Bad Science', *Royal Society Open Science,* 3 (160384).

Sosa, Ernest (1991), *Knowledge in Perspective* (Cambridge: Cambridge University Press).

Stanley, Jason (2005), *Knowledge and Practical Interests* (Oxford: Oxford University Press).

Starmans, Christina and Friedman, Ori (2020), 'Expert or Esoteric? Philosophers Attribute Knowledge Differently Than All Other Academics', 44 (7), e12850.

Steiner, Peter M., Atzmüller, Christiane, and Su, Dan (2016), 'Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap', *Journal of Methods and Measurement in the Social Sciences,* 7 (2), 52-94.

Stich, Stephen and Tobia, Kevin P. (2016), 'Experimental Philosophy and the Philosophical Tradition', *A Companion to Experimental Philosophy* (John Wiley & Sons, Ltd), 3-21.

Stich, Stephen P. and Machery, Edouard (2022), 'Demographic Differences in Philosophical Intuition: A Reply to Joshua Knobe', *Review of Philosophy and Psychology*.

Tobia, Kevin, Buckwalter, Wesley, and Stich, Stephen (2013), 'Moral Intuitions: Are Philosophers Experts?', *Philosophical Psychology,* 26 (5), 629-38.

Tobia, Kevin, Chapman, Gretchen B, and Stich, Stephen (2013), 'Cleanliness Is Next to Morality, Even for Philosophers', *Journal of Consciousness Studies,* 20 (11-12).

Turri, John (2016a), 'Knowledge Judgments in “Gettier” Cases', in Justin Sytsma and Wesley Buckwalter (eds.), *A Companion to Experimental Philosophy* (New York: Wiley-Blackwell).

Turri, John (2016b), *Knowledge and the Norm of Assertion: An Essay in Philosophical Science* (Cambridge, UK: Open Book Publishers).

Turri, John (2016c), 'The Radicalism of Truth-Insensitive Epistemology: Truth's Profound Effect on the Evaluation of Belief', *Philosophy and Phenomenological Research,* 93 (2), 348-67.

Turri, John (2016d), 'A New Paradigm for Epistemology from Reliabilism to Abilism', *Ergo, an Open Access Journal of Philosophy,* 3.

Turri, John (2017a), 'Compatibilism Can Be Natural', *Consciousness and Cognition,* 51, 68-81.

Turri, John (2017b), 'Knowledge Attributions in Iterated Fake Barn Cases', *Analysis,* 77 (1), 104-15.

Turri, John (2017c), 'Epistemic Contextualism: An Idle Hypothesis', *Australasian Journal of Philosophy* 95 (1), 141-56.

Turri, John and Buckwalter, Wesley (2017), 'Descartes's Schism, Locke's Reunion: Completing the Pragmatic Turn in Epistemology', *American Philosophical Quarterly,* 54 (1), 25-46.

Turri, John, Buckwalter, Wesley, and Rose, David (2016), 'Actionability Judgments Cause Knowledge Judgments', *Thought,* 5 (3), 212-22.

Weatherson, Brian (2005), 'Can We Do without Pragmatic Encroachment?', *Philosophical Perspectives,* 19 (1), 417-43.

Williams, Bernard (1995), 'Making Sense of Humanity: And Other Philosophical Papers 1982–1993', (Cambridge University Press).