The construction of philosophical intuitions

Do intuitions about thought experimental scenarios depend on epistemically irrelevant factors? Proponents of the 'negative program' in experimental philosophy say, "yes"; their critics say, "no". In this paper, I argue that examining the psychological mechanisms we rely on to construct counterfactual representations can strengthen the negative program. I sketch a plausible approach to cognitive architecture, which would support the negative program; and I argue that research on the neuroscience of counterfactual thinking provides at least initial support for this approach.

Some experimental philosophers claim that intuitions about thought experimental situations often depend on epistemically irrelevant factors. This 'negative program' in experimental philosophy has revived discussions about the value of these intuitions and about the intuition pumps that are deployed across philosophy.[1] But critics have argued that survey-based methods are insufficient to fund a plausible critique of philosophical intuition. I believe there is more to say. Specifically, I contend that the critical project of experimental philosophy can be strengthened and extended by thinking carefully about the psychological mechanisms we are likely to rely on in constructing representations of thought experimental scenarios. My aim is to provide a sketch of the sort of cognitive architecture that would support the negative program. And I argue that research on the neuroscience of counterfactual thinking provides some initial support for this model. That said, I concede that more research is necessary to confirm the hypothesis that I offer; but even if the model I propose is inaccurate, I hold that it can help to clarify the hierarchy of models that are needed to link behavioral observations to the confirmation of a hypothesis in experimental philosophy.

1. Intuitions and answers

Advocates of the negative program frequently note that we don't have conscious, introspective access to the factors that guide the production of our judgments; and in light of this fact, they argue that it is impossible to discriminate between intuitions that are produced by legitimate as opposed to illegitimate means. Assuming that people's judgments about thought experimental scenarios are expressions of their intuitions, this position can be stated briefly as follows. Many experimental data reveal that judgments about thought experimental scenarios are affected by epistemically irrelevant factors. If intuitions tracked the truth, our judgments wouldn't be affected by these factors. So we have reason to think that intuitions don't track the truth, and to be skeptical of the judgments people provide in response to such scenarios. This argument yields a challenge for any philosophical rationalist who claims that their intuitions are immune to epistemically irrelevant factors. And many rationalists concede that the existing data provide at least a *prima facie* reason to think that intuitions are sometimes affected by factors like ethnic and cultural heritage, the order in which scenarios are read and interpreted, the wording of particular scenarios, and the current affective states that participants are experiencing.

There are many routes by which *a priori* methods could be defended against this challenge. It could be argued that 'genuine' intuitions are accompanied by a distinctive phenomenology, which somehow reveals their veracity. But establishing the truth of this claim would require showing that experimental philosophers have failed to distinguish genuine from ersatz intuitions, and that they have only examined ersatz intuitions. It could also be argued that genuine intuitions require a distinctive kind of philosophical competence, which distinguishes them from the answers experimental philosophers collect (cf., Bengson 2013; Kauppinen 2007; Ludwig 2007; though see

[1] This is a *new wave* of criticisms. Criticisms of intuition-based methodologies have long played a role in feminist epistemology and philosophy of science, noting that the evaluation of thought experiments depends on situational and demographic considerations that yield no epistemic warrant.

Weinberg & Alexander 2014 for critiques of this position). Or it could be argued that the methods employed by experimental philosophers are misguided in some other way, and that survey-based methodologies are—for some reason—unfit to discover the intuitions we really have. Each of these replies depends on showing that there is an epistemically significant difference between the answers provided by participants in experimental studies and the intuitions held by philosophers. And if such a difference could be established, then there would be reason to reject the challenge to intuition-based methods in philosophy that is commonly mounted by proponents of the negative program. But how could we establish the existence of such genuine intuitions, which are immune from distortion, and more epistemically secure than the answers people offer in experimental contexts? My contention is that we can't.

This is not to claim that we should accept the claims made by experimental philosophers without criticism. The collection of data often occurs in experimental situations where participants are asked to read unfamiliar scenarios, and consider philosophically important issues that they may not have encountered before. Such scenarios abstract away from many relevant details to describe a counterfactual possibility, and experimental philosophers are often forced to move further away from everyday experience to make their scenarios concise enough to be viable for survey-based research (Scholl 2008). Furthermore, when philosophers rely on thought experimental methods to test the boundaries of their theories, they do so within a broader practice of theory construction. So, philosophical texts that include such scenarios also tend to provide guidance for interpreting the situation that is described. But in an experimental context, the context must be removed; and this makes it hard to know if the participants are all interpreting the situation in the same way. Finally survey-based methods require converting immediate responses to thought experimental scenarios, which may be little more than diffuse inklings, into affirmations, rejections, or numbers along specified continua; and, there is no obvious way of treating these values as philosophical judgments. For example, it is rarely clear whether all participants use the numbers on a scale in the same way, and it is rarely clear what a small difference within a population (e.g., one-point difference on a seven-point scale) actually amounts to with respect to a philosophical claim. So even if survey-based methods do yield data that are statistically significant, we are left to wonder about their *philosophical importance* (Huebner in press).

Nonetheless, it would be a mistake to reject the experimental challenge out of hand. Experimental philosophers capitalize on the wide range of research-and-development that has been carried out in social psychology over the past 50 years. They employ tools and techniques that tend to elicit statistically significant differences in survey-based research; and their experiments are not isolated attempts at investigating philosophical intuitions, but a species of research embedded in a stable tradition of examining cultural and demographic variation, as well as local effects of word choice and affective valence. Moving beyond this empirical tradition, philosophers outside of the experimental movement have long noted the difficulties inherent in the evaluation of thought experimental scenarios (Dennett 1988, 1995; Gendler 2007). Minor differences in presentation can often modulate, attenuate, and even reverse the judgments people make about various scenarios. So the novelty of the negative program does not derive from its challenge to philosophical rationalism, it derives from the fact that experimental philosophy deploys statistical tools to demonstrate the robustness and generality of common criticisms of thought experimental methods.

This is why philosophical rationalists often attempt to show that judgments offered in experimental contexts are not expressions of genuine intuition—if there is something special about the way in which philosophical intuitions are produced, then it will remain possible that only immediate and unreflective responses to thought experimental probes are called into question by the experimental results of the negative program. Importantly, there are at least two reasons to think that experimental judgments are particularly problematic. First, there may be numerous cases where participants provide an answer in response to an experimental probe, "even though they do not have any intuition one way or another about it" (Bengson 2013, 506). In these cases, their responses are constructed on-the-fly, yielding something more like a guess than a report of an intuition.[2] The existence of such guesses would not be sufficient to license worries about the epistemic value of philosophical intuitions; their prevalence would only provide evidence that something about the experimental situation had led most participants to come up with roughly the same guess. And while it may seem odd to suppose that such constructed responses would generate robust patterns of experimental data, research in social psychology suggests that participants routinely construct similar responses to survey questions, even where they don't have antecedent commitments (see Schwartz 2007 for a review). Whether they do so in a particular case is difficult to establish; and to the best of my knowledge, experimental philosophers have not yet shown that their data reveal intuitions rather than guesses of this sort. Second, and relatedly, there may be cases where thought experimental probes lead participants to offer answers that don't correspond to what they actually believe. Specifically, their responses may be driven by external theoretical considerations, or by attempts to please the experimenter, yielding '*stray answers'* that diverge from their antecedent intuitions (cf. Bengson 2013, 9; Huang & Bargh 2014). Unless there is some way to rule out the production of guesses and stray answers, the patterns of response revealed by experimental philosophers should not be treated as evidence about the intuitions that people possess. Since there is no obvious way to rule out such possibilities on the basis of the experimental data alone, it is unclear whether intuitions are ever recovered by asking participants to respond to thought experimental probes.

Many critics of experimental philosophy have argued that survey-based methods can only warrant conclusions about the patterns of answers people give in experimental situations (Carmel 2011; Scholl 2008). I too have argued that the data collected by experimental philosophers are typically insufficient to license an inference about the presence or lack of philosophical intuitions. The statistical models employed in experimental philosophy can reveal differences between populations, and they can show that these differences are unlikely to arise by chance; but this doesn't rule out the possibility that these differences arise as a result of guesses or stray answers. Consequently, these models don't provide direct support for any positive hypothesis regarding the source of the statistically reliable patterns in the data. If answers do not need to express intuitions, the existing experimental data cannot provide unambiguous support for the negative program, and they cannot provide clear evidence about the presence or lack of philosophical intuitions.

But the even we suppose that the results in experimental studies are the result of similar guesses and stray answers, the philosophical rationalist is not out of the woods. They need to show that there is something special about the ways in which philosophical intuitions are produced, and that this distinguishes them from the answers offered by experimental participants. But why should we suppose that this is the case. Context dependency and situational malleability may provide evidence that participants are not expressing antecedently held attitudes, but are instead constructing their answers on-the-fly to satisfy the demands of experimental situation (Schwarz 2007). And survey-based methods may be too coarse grained to reveal the operative principles for the cognitive mechanisms responsible for producing these answers. But these facts can only motivate criticism of experimental philosophy if there is reason to believe that there are genuine intuitions. So rather than funding a criticism of the negative program, we should see these facts as clarifying where the burden lies in safeguarding philosophical intuition against the encroachment of empirical data. If the distinction between genuine and ersatz intuitions is to do any explanatory work, it must be shown that there are genuine intuitions, and that they can be uncovered in some way that precludes the intrusion of epistemically illegitimate factors.

2. Experimental methods and intuitions.

One way to establish the existence of genuine intuitions would be to appeal to cases where answers converge across experimental populations and experimental manipulations. John

[2] The argument that I develop here extends an insight developed by John Bengson (2013). However, I adopt the term 'guesses' to avoid the ableism of Bengson's original terminology.

Bengson (2013) develops one version of this claim, arguing that people are immediately struck in particular ways by ‘clear cases’, and he contends that people have epistemically reliable intuitions about these cases. He suggests that such intuitions can be used to calibrate claims about more problematic intuitions, constituting an epistemic baseline. But the presence of such patterns is consistent with the possibility that the larger task demands of doing philosophy in an academic setting, or other kinds of situational constraints, evoke relatively stable patterns in responses to even these cases. So the evidence that such patterns are evidence of genuine intuitions must come from another source. Experimentally, the rationalist philosopher may attempt to rely on claims about proper experimental controls or similarities across replications; or they might claim that subsequent argumentation can provide evidence of epistemically secure intuitions (Bengson 2013, 513). However, I contend that if guesses and stray answers are at play in the production of statistically reliable patterns of data, these attempts to safeguard genuine intuitions against the criticisms of the negative program are strangely inapposite. .

From an experimental perspective, it is clear that proper controls block the intrusion of some extraneous factors in survey-based research. And only properly controlled experiments can allow for statistical analyses that make the effects of experimental manipulations intelligible. But since statistical regularities emerge in patterns of *guesses* and *stray* answers (Schwartz 2007), the presence of such regularities alone cannot justify claims about the presence or absence of genuine intuitions. Doing so would require providing independent reason to think that some experimental manipulations evoke intuitions rather than leading participants to construct answers on-the-fly. But laboratory experiments and survey-based research have a peculiar ecology, which may lead to the construction of answers no matter which controls are employed; controlled laboratory research is designed to prevent participants from relying on the heterogeneous array of heuristics, assumptions, and cognitive strategies that they employ in everyday life. By leading each participant to take up the same, or a similar cognitive strategy in responding to the experimental probe, experimentalists may be inadvertently triggering cognitive strategies that lead participants to construct similar answers instead of leading them to retrieve antecedently held intuitions (cf., Huebner 2010, 2011, 2012).

Similar worries arise in attempts to rely on subsequent experimentation. Good psychological methodology requires replication, and the presence of similar statistical patterns in responses to different thought experimental probes, or with different populations, tells us that participants are using *similar* evaluative strategies in formulating their answers. But again, without some strategy for distinguishing epistemically reliable intuitions from guesses and stray answers, the presence of *similar patterns* alone cannot guarantee that *the same intuition* is evoked across experimental situations. The same answer may be evoked by similarities in the experimental situation or by similarities in the thought experimental probe. This worry cuts deep. Perfect replications of previous results are rare in survey-based research. And to the extent that experimental probes are seen as similar enough to yield replications or extensions of previous results, this is because researchers *treat them* as expressing the same philosophically significant principle. So when similar scenarios yield similar results, it may be because they share numerous structural features that could be taken up (consciously or non-consciously) in the construction of an answer. Consequently, stable patterns that are preserved across nominally ‘different’ probes may be the result of structural similarities between these probes, and not the result of similarities between the intuitions that are evoked.

Finally, it is worth remembering that thought experimental probes evoke different answers within populations as well as between populations, and such differences often remain relatively stable across experiments.[3] So we need an explanation of both the differences that arise as a

[3] Demographic differences often interact with the structure of *similar* thought experimental probes to yield statistical regularities with wide standard deviations and a high degree of variance within a population. Weinberg et al (2001) report a shift from 74% affirmative responses to 50-60% on Gettier intuitions as a result of cultural factors; and, Swain et al (2008) report a shift from 60% affirmative responses to 40% as a result of order of presentation on thought experiments intended to support reliabilism. This shows that 50%

result of using different probes and different populations, and an account of these similarities. As I argue below, a plausible explanation of these patterns is likely to appeal to the strategies participants employ in integrating the most salient features of a thought experimental probe with their own representations of previously encountered or imagined situations. And if I am approximately right, an account of these constructive process will explain why some things feel like intuitions, why individual differences and differences in cultural background sometimes play a significant role in the production of answers, and why similar thought experimental probes tend to evoke similar answers across experimental situations even though there are no similarities in antecedently held intuitions. Consequently, the kinds of arguments that have been developed by philosophical rationalists are likely to gut quite deep in showing that experimental tasks are unlikely to fund claims about genuine intuitions. But again, this doesn’t demonstrate that there are genuine intuitions, as opposed to ersatz intuitions that are constructed to satisfy the demands imposed by a particular philosophical community.

Of course, many philosophical rationalists may be happy to concede that experimental investigations are unlikely to reveal the existence of genuine intuitions. And this may be true because subsequent argumentation and philosophical discussion are required to distinguish genuine from ersatz intuitions. But does recalcitrance in the face of subsequent argumentation provide evidence that there are no problematic constructive processes at play in the production things that *feel like* intuitions on subsequent examination. I think not. People often engage in post-hoc rationalization when questioned about the answers they provide in experimental contexts (e.g., Bargh & Chartrand 1999; Festinger, Riecken, & Schachter 1956; Haidt 2001; Wegner 2002; Wilson 2002). Indeed situations that evoke answers that conflict with previously held beliefs or attitudes, or that lead people to offer an answer where they did not have a previously have a belief or attitude, are most likely to lead people to search for post-hoc rationalizations that can justify their answers. This is not just a matter of adopting strategies to reduce cognitive dissonance (Festinger 1956); in such situations people engage in motivated reasoning, yielding a process of memory search and belief construction that is biased toward justifications that are consonant with the answers they have given (Kunda 1990). To the extent that they succeed in finding reasons that can justify their answers, they may come to believe that these were their reasons all along. It is not clear to me that we have any reason to rule out a similar possibility in relation to the arguments we develop in philosophical discussions, where initial responses to thought experimental probes trigger a targeted search for reasons that will justify our initial hunches. Unless there is reason to believe that there is something special about philosophical conversation, then recalcitrance alone should not confer any additional epistemic value. Philosophical rationalists therefore need a way to rule out the potential contributions of confabulation, dissonance reduction, and motivated reasoning in philosophical argumentation.

To summarize, the philosophical rationalist needs some reason to believe that the experience of an intuition is something more than a thin veneer that is cast over a potentially problematic, and non-conscious constructive process. This requires identifying genuine intuitions in a way that distinguishes them from ersatz intuitions, guesses, and stray answers. Indeed, this is common knowledge among philosophical rationalists. Bengson (2013) argues that intuitions are conscious mental states, which immediately present the world as being a particular way; and he claims that they are not consciously *constructed* nor voluntarily arrived upon by reasoning or deliberation. He seems to hold that intuitions are genuine so long as they feel like intuitions, and so long as we aren’t actively and consciously engaged in a process of voluntary belief construction. Karl Ludwig (2007) denies that intuitions have any distinctive presentational or phenomenological character, and he argues that they are judgments formed solely on the basis of one’s competence with a

of participants across cultures offered an affirmative response to Gettier-type cases, and that 40% of participants had reliablist intuitions regardless of the order of presentation. While culture and order of presentation affect *some* participants, it is not clear that every participant is so affected, nor that cultural differences have an affect on every participant. As I argue below, this is a good reason to adopt a constructive approach to answers and intuitions.

concept—precisely how we identify them, I'm not sure. And Antti Kauppinen (2007) argues that philosophical intuitions withstand careful reflection and variations in factors that might influence the production of a one-time judgment; unlike answers, they are robust. Each of these suggestions is an attempt to distinguish intuitions from epistemically distorted states; and if one of them succeeds, it may be possible to treat things like initial hunches, guesses and stray answers as experimental artifacts, which arise because people are forced to *construct* novel answers about unfamiliar situations. This would allow the philosophical rationalist to adopt a familiar sort of strategy from the cognitive sciences, relegating epistemically unreliable factors to the status of performance variables (cf., Chomsky 1965, 3). But how likely is it to succeed in the case of philosophical intuitions? When we look closer at the arguments philosophical rationalists offer, it seems clear to me that we should not be particularly optimistic about the success of such a strategy.

Many philosophical rationalists have attempted to specify the conditions under which successful intuiting is likely. Bengson (2013, 526) argues that we might be able to prune back the environmental factors that inhibit the production of genuine intuitions by figuring out which kinds of extraneous factors lead us to experience something as an intuition when it is not. And Kauppinen (2007) argues that philosophically competent judges evaluate thought experimental scenarios in ideal philosophical situations, and rely only on semantically and epistemically relevant factors. Kauppinen notes the difficulties inherent in specifying precise boundaries around the 'ideal conditions' for intuiting. And Bengson acknowledges that demonstrating the existence of philosophical intuitions requires specifying the 'ideal conditions' under which feelings of intuition are properly produced. Unlike Kauppinen, he provides a heroic attempt to specify the conditions under which a feeling of intuition is a reliable guide to how we are in fact struck—and his claims make it clear just how much is at stake in this defense of philosophical intuition. He argues that we must be sensitive to the strength and clarity of a purported intuition, check it for consistency and coherence with other purported intuitions, and ensure that we have made all of the philosophically relevant distinctions. We must also be attentive to the wide variety of contextual factors and ambiguities that might affect our judgments, seek out corroboration about our intuitions from our epistemic peers, and consider nearby thought experiments in ways that help us to avoid the all too familiar focus on 'weird' cases. And we must avoid alcohol and other drugs, if our intuitions are to be trustworthy.

Should we defend intuition at the cost of relegating philosophy to the status of sober thinking? Should we stick with intuitions through thick and thin (Weinberg & Alexander 2014)? I think that doing so would be a mistake. And experimental philosophers have done a masterful job of showing that such arguments are unlikely to succeed. But negative and critical arguments can only take us so far. And in the remainder of this paper, I adopt another tack, offering positive arguments designed to show that a constructive process is likely to be operative in the production of 'intuitions', guesses, and stray answers. Specifically, I argue that we should not appeal to the experience of a difference between *intuitions* and *constructions* as evidence of a distinction between genuine and ersatz intuitions; put differently, I argue that it's a mistake to assume that we can distinguish epistemically reliable intuitions from ersatz intuitions that rely on epistemically problematic strategies for constructing representations.

3. The construction and production of mental spaces

I take it as well-established that people can be led to create plausible stories about the reasoning processes they have employed, even where it is unlikely that they have employed them (Nisbett & Wilson 1977; Wegner & Wheatley 1999). This is not to deny that everyone has access to their own thoughts, but to acknowledge that most acts of introspection are at least partially retrospective and constructive. Examining how things seem to us is a temporally extended process, and it requires identifying and categorizing our introspected states. In this process, meta-cognitive mechanisms that obsessively and mechanically convert thought into linguistic form, and vice versa, impose structure on our mental states, as they are categorized and identified as contentful (Carruthers 2009; Huebner & Dennett 2009; Jackendoff 1996). In many cases, this imposition of linguistic and categorical structure masks the constructive processes that have produced an answer (Akins 1996, 353). This is why responses to thought experimental scenarios can be the results of guesses and *stray answers*, even where they do not seem to be so.

Whether the answers that people provide to thought experimental are guesses or stray answers is an empirical question. And discriminating cases where epistemically irrelevant factors are incorporated into a response, from cases when they are not, requires a plausible account of the introspectively opaque causal factors that are operative in producing such responses. This requires examining a range of evidence that philosophical rationalists rarely consider. But, examining the cognitive and computational mechanisms that facilitate counterfactual reasoning offers an important insight into the constructive processes employed as people interpret thought experimental scenarios. Many of the evaluations carried out by a human brain rely on 'skeletal' representations, which can be fleshed out only when doing so is necessary. Moreover, a wide range of empirical data (derived from behavioral experiments, imaging tasks, and neuropsychological studies) suggest that thinking about the past and the future, as well as considering counterfactual possibilities, depends on a common network linked to planning, imagining, and episodic memory (Addis et al 2007; Buckner & Carroll 2006; Schacter et al 2015). *These facts* have important implications for understanding how previously imagined and encountered situations are used to construct cognitive models of counterfactual situations, or so I shall argue.

We live in an information-rich environment, but biological limitations on processing speed and constraints on working memory make it risky for us to waste time on encoding and recall where we can get around the world by other means (cf., Bartlet 1932). So we typically remember only the *gist* of a situation, rather than encoding a fully elaborated memory that records every detail. Many things are irrelevant to our future actions, or too common to bother remembering. But by encoding a skeletal representation of a situation, we can remember those things that are likely to be important for future actions without cluttering memory with irrelevant information. These claims may seem obvious; but they have important implications for thinking about the encoding and retrieval of information. While there is debate over the structure of the skeletal representations that must be fleshed out in remembering and imagining, it has become relatively clear that mechanisms dedicated to counterfactual simulation are employed in reconstructing episodic memories and imagining alternative ways that the world could be (Schacter & Addis 2008; Schacter et al 2015).

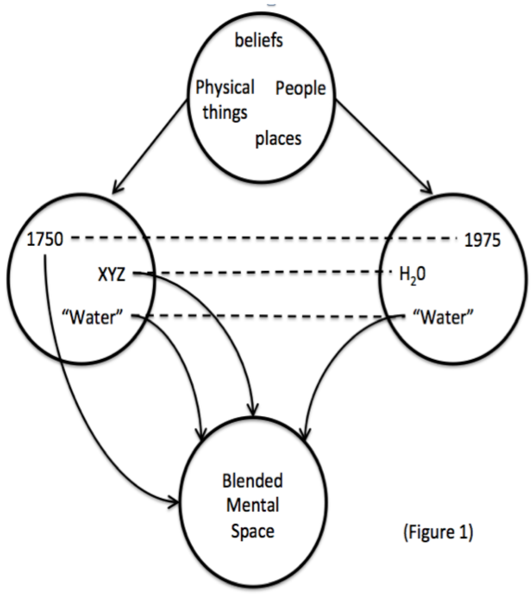
Imagining possibilities and possible futures depends on a constructive system that can "draw on the elements and gist of the past, and extract, recombine and reassemble them into imaginary events that never occurred in that exact form" (Schacter & Addis 2007). In most cases, this process takes representations of previously experienced and encountered situations, and populates a field of possibilities by fleshing out a mental model that allows for inferences about things that are not explicitly encoded in the structure of the representation. This capacity emerges early, and five-year-old children possess a clear sense of the features that are shared, as well as the features that differ across different imaginable possibilities; and they also have a strong sense of what is possible in different fictional worlds (Skolnick & Bloom 2006a, 2006b). This capacity begins from previously encoded assumptions about what our world is like, but even children realized that it is reasonable to introduce modifications on the basis of ''what the story tells us explicitly, what we can directly deduce from specific conventions of the fictional genre, and, most importantly, how similar to the real world the fictional world is described as being'' (Skolnick & Bloom, 2006b, p. 77). There is reason to believe that adults rely on these same capacities in enjoying well-directed films and cleverly written literature. And I believe that carefully crafted intuition pumps evoke a similar constructive process, leading us to draw on initial assumptions to create a 'blended mental space' within which we can evaluate counterfactual possibilities (Fauconnier & Turner 2002). I can't establish this suggestion decisively, but I hope to show that it provides a plausible way to think about existing data in experimental philosophy, as well as existing theorizing by philosophers.

The clearest way to think about blended mental spaces is by considering an example, which shows how blended mental spaces could provide the cognitive resources necessary to understand claims about ‘possible worlds’. Consider Putnam’s Twin Earth, leaving aside its plausibility and its purported implications for semantic theory. I am concerned with the way that Putnam leads different readers to interpret the situation he is concerned with in similar ways, by triggering the construction of a blended mental space that integrates representations drawn from ‘ everyday activity’ with representations provided by his philosophical narrative. Here is Putnam’s (1975, 139-140) original description of the case:

> For the purpose of the following science-fiction examples, we shall suppose that somewhere in the galaxy there is a planet we shall call Twin Earth. Twin Earth is very much like Earth; in fact, people on Twin Earth even speak English. In fact, apart from the differences we shall specify in our science-fiction examples, the reader may suppose that Twin Earth is exactly like Earth. He may even suppose that he has a Doppelgänger—an identical copy—on Twin Earth, if he wishes, although my story will not depend on this...One of the peculiarities of Twin Earth is that the liquid called “water” is not H_2O but a different liquid whose chemical formula is very long and complicated. I shall abbreviate this chemical formula simply as XYZ. I shall suppose that XYZ is indistinguishable from water at normal temperatures and pressures. In particular, it tastes like water and it quenches thirst like water. Also I shall suppose that the oceans and lakes of Twin Earth contain XYZ and not water, that it rains XYZ on Twin Earth and not water, etc.

Putnam closes his discussion by noting that we should imagine the Twin Earthlings as having knowledge of chemistry approximately equivalent to the knowledge people on Earth had in approximately 1750.

This scenario skillfully leads readers to activate a conceptual integration network consisting of four distinct kinds of representations, some of which rely on the scenario that is described, some of which rely on stored generic information about the world, and some of which are constructed on the fly.. In the scenario, Putnam specifies two input spaces. The first that leads readers to represent the features of our world that will be important to his thought experimental paper; these representations would include the current year (1975, at the time that the original paper was published), the currently known facts about H_2O, and facts about how ordinary English speakers tend to use the word “water”. The second specifies a narrow range of counterfactual states, which facilitate the mapping of states of Twin Earth onto important features of the world that his readers will readily understand. This includes a year that corresponds to an earlier year on earth (“this situation is sort of like what it would have been on earth in 1750”), a newly identified substance, XYZ, which corresponds to a well-known substance, H_2O, and parallel uses of the term “water” on Twin Earth and Earth). Putnam also assumes, whether explicitly or tacitly, that his readers will share a common, generic embedding space, which represents structures and features that can be common to both Twin Earth and Earth, including things like people, mental states, places, and things. Finally, in reading the scenario and attempting to understand what possibilities it affords, reader construct a blended mental space as the narrative unfolds, selectively projecting salient features from the three input spaces, and integrating them to complete patterns that are left under-described, and to elaborate upon features of the thought experimental situation that are left unspecified (this process is depicted in Figure 1). Put more simply, I propose that a blended mental space is constructed in working memory by integrating Putnam’s narrative with representations of previously experienced and imagined situations. This process draws on existing cognitive resources that are shared to differing degrees by his readers, and allows readers to construct a simplified model of the counterfactual situation. In this way, skeletally represented conceptual packets are able to be fleshed out in ways that provide us with a local understanding of the situation; this process of fleshing out the relevant representation yields an idealized mental model, which allows us to draw inferences about things we haven't considered previously and things that aren't included in the description of the scenario (Fauconnier & Turner 2002, 102).



(Figure 1)

Conceptual integration networks have their original home in our richly textured social world, and they often depend on the fact that many aspects of the world are stable enough that they don't need to be encoded. But no matter where they are produced, constraints on processing speed and limitations on working memory ensure that the elements and features in a mental space are only represented skeletally (they yield a *gist-based* representation of a counterfactual situation, rather than a florid and detailed representation of a possible world). This keeps the structure of mental spaces minimal, partial, and abstract. As we move about our world, this is a good thing. It allows us to rapidly revise, modify, and interpret novel possibilities as more elements or features are integrated into the representation that we are working with. When we are presented with a narrative, whether in the context of a philosophical thought experiment or a science fiction novel, we dynamically construct an abstract representation of the possibilities afforded by the situation described. These representations may sometimes appear to be simple, as we entertain them consciously, but they are constructed from features that are stored in long-term memory, leaving significant room for the mental space to "be modified dynamically as thought and discourse unfold" (Fauconnier & Turner 2002, 102).

As we move about our world, we are frequently prompted to construct mental spaces to understand narratives and other language-based phenomena. The Mental spaces we use most frequently become entrenched in long-term memory (Fauconnier & Turner 2002, 103). These models contain partial and incomplete representations of a situation, and they focus our attention on the specific elements and features that are most salient; but they also provide resources for carrying out additional inferences, and constructing more fleshed out representations. When we construct a mental space, we reflexively construct conceptual integration networks that situate novel mental spaces against memories of previously encountered or imagined situations. This process often occurs completely unnoticed; though further revisions and modifications of a mental space often become noticeable "when the emergent meaning to which they apply in the blend seems remarkably distant from the domain of the input from which they came" (Fauconnier & Turner 2002, 143). Building on a suggestion by Nick Epley & Thomas Gilovich (2006), I contend that conflicts between a novel mental space and mental spaces that is entrenched in long term memory can trigger a deliberate and effortful search for features that ‘should have been included’ in an input space; it can also trigger a more deliberate search of the possibilities that are afforded by the blend itself, to see if there are ways of making it seem more plausible, or to justify it in terms of our other goals and values (cf., Fauconnier & Turner 2002, 44).

These processes allow us to adjust the structure of an initially constructed mental space, and we stop adjusting only once we have reach something that reflectively feels like a plausible representation of a counterfactual situation. Unfortunately, it’s not always easy to see whether the situation we have represented in a blend is plausible, or plausible enough. So the constructive process is often repeated multiple times, even in cases where our first intuition would have been good enough to capture the salient aspect of the situation we were trying to understand. So we make adjustments that seem like they will be sufficient, and we them test again to see whether that adjustment was sufficient. If things feel ok, we stop. But, if they don’t, we adjust again and test for plausibility—repeating, over and over, if doing so is necessary (Epley & Gilovich 2006, 312-13). This process of adjusting the structure of a blended space is always effortful, often conscious, and frequently deliberate. So we frequently become aware of having used effortful, conscious, and deliberative processes to adjust our representation of a space of possibilities. But many blends do not evoke this process of revision and recalibration. Sometimes this is because the original blend was good enough, and sometimes it is because we don’t have much to calibrate it against. But in each case, the original constructive process remains inaccessible to introspection, even if later revisions are introspectively available (Epley & Gilovich, 2001).

In the simplest integration networks, the process of mapping previously encountered possibilities onto novel situations is rapid and nearly automatic. Where the organizing frame is rich enough to make the cross-frame mappings clear, the production of the blended space does not require any kind of additional search. Because the brain does this “instantly and unconsciously, we take the construction of meaning for granted. Or, rather, we tend to take the meaning as emanating from its formal representation, the picture, when in fact it is being actively constructed by staggeringly complex mental representations in the brain of the viewer” (Fauconnier & Turner 2002, 5). In many cases, we have no awareness—nor even cognitive access—to the imaginative and constructive work that we do as we construct a blended mental space. Indeed, in many cases, we have a hard time seeing that there was room for construction to have taken place—our minds are indeed elegant kludges!

4. Constructing intuitions

We are now in a position to see why attempts at distinguishing epistemically secure intuitions from epistemically flawed answers are unlikely to succeed. When someone is presented with a thought experimental scenario, whether in the context of reading a philosophy article, or in the context of an experimental situation, subpersonal conceptual integration mechanisms are automatically brought on-line to construct fleshed-out representations of the skeletal content presented in the scenario. As I noted above, this process draws on multiple sources of information, including the representations presented in a thought experimental scenario, stored representations of previously encountered facts about the world, and representations constructed on-the-fly, to bridge the gap between these other kinds of representations. The representations presented in the scenario will be shared among all participants; and most participants are likely to have similar (although only partially overlapping) representations of the features of our world that are relevant to interpreting the situations; however, attempts to bridge the gap between previously encountered situations and counterfactually described situations will require constructing a novel, blended mental space, which selectively projects salient features from these input spaces, and integrates them in a way that completes the patterns that have been left unspecified by the thought experimental probe. As a result, these thought experimental scenarios should tend to generate stable patterns of answers, in part because participants are being led to construct similar—though importantly, not identical—representations of counterfactual possibilities.

In interpreting thought experimental scenarios, we are led to dynamically construct abstract representations of the space of possibilities in which a scenario takes place. As people read a thought experimental scenario they reflexively construct mental spaces in working memory that integrate the details of the narrative with previously imagined or encountered situations. Very little of this process is accessible to first person cognition; and the construction of a mental space in response to a philosophical thought experiment will often draw from a wide range of representational resources, including many that are epistemically irrelevant—such as culturally local assumptions as well as idiosyncrasies of an individual's experience. That said, the concepts expressed in a thought experimental scenario will constrain the elements that are included in a blended mental space, and they can help to recruit representational resources that would have otherwise remained inactive. In many cases, this will lead to the representation of some aspects of the narrative in light of previously held beliefs about how possible worlds must be structured. Sometimes this takes place consciously and reflectively, other times it takes place subpersonally and reflexively; but the process always relies on the same constructive mechanisms.

If this account is roughly right, thought experimental probes will often function as anchors for conceptual blends, leading to convergence in the blends that are constructed by different people (cf., Hutchins 2005). However, as these simplified and idealized representations are fleshed out to draw inferences, people will be likely to recruit mental spaces stored in long-term memory; this would help to explain why shifts in context, or shifts in the description of a thought experimental scenario tend to evoke different responses. Participants would be constructing different conceptual blends as a result of difference in the scenario or differences in their contribution to the conceptual blend (Huebner 2010). To my mind, this is the most plausible way to make sense of the pervasive and unexplained within-group differences, as well as the wide standard deviations that commonly arise in experimental philosophy and moral psychology. The experimental situation is designed to lead every person who engages with the stimuli to see it in the same way; and, the situation of 'being a participant' in a psychology study is shared among all of the participants; but, because of the narrative character of thought experimental probes, they lead people to construct blended mental spaces that reflexively draw on the representational resources afforded by previously experienced and imagined situations.

As mental spaces are constructed, additional systems are often brought on-line to evaluate (often automatically, and subpersonally) the cognitive and behavioral significance of the representation, as it relates to other goal-directed behavior that might be engaged in. In general, these mechanisms will produce subpersonal representations that are sent to endogenous monitoring systems where they are converted into 'conscious propositional thoughts' and then further examined to see if they should be offered as answers, triangulated against previously held beliefs, updated, revised, or even rejected (cf., Huebner & Dennett 2009). As I noted above, various aspects of this constructive and evaluative process can take place consciously. And when they do, this process is often accompanied by overt behavior. As people examine thought experimental prompts, they sometimes vocalize the reasoning process they are going through as they read and evaluate scenarios. Having collected data for moral psychology experiments, I have often experienced this overt processing first-hand; by carefully attending to the overt behavior of my participants, it became clear that thought experimental scenarios rarely just strike participants in any way. They frequently appeared to feel uncomfortable with their answers, and they frequently revised their answer multiple times on-the-fly.

Unfortunately similar worries will arise even where there are relevant controls or subsequent investigations. If my hypothesis is roughly correct, coming to an understanding of a narrative is always a constructive and interpretive process. Even when people think carefully about the structure of their mental spaces they construct, the best they can do is to provide a careful analysis of the additional structures that were imposed on a skeletal representation of the thought experimental scenario. Convergence between earlier and later responses—even where it arises through careful philosophical reflection—provides little evidence that a person's response to a situation depends on an epistemically valuable intuition; after all, such convergence can equally well arise because philosophers (or non-philosophers) have attempted to construct new mental spaces, which will reduce the cognitive dissonance evoked by further questioning. Put differently, asking someone to examine the process by which they have come to their initial answer may trigger the construction of new conceptual blends, which draw from both the initial blend and new details that have come to mind as they have carried out a motivated search for information they failed to include initially. While further answers may be produced, we have little reason to think that this subsequent questioning is evidence of the veracity of a previously held intuition. Given the constructive nature of memory, asking someone to examine the process that led to her offer her initial answer may lead them to incorporate the reasons they reported as reasons that they employed in making her initial judgment. This, I take it, is bad news for philosophical rationalists.

Perhaps the proponent of philosophical rationalism can resort to a fallback position, claiming that only people who are properly trained in academic philosophy have epistemically reliable intuitions about philosophically significant claims. Indeed, it's commonly suggested that philosophical training makes judgments about thought experimental scenarios more reliable (Williamson 2007, 2011), and that there are distinctive philosophical skills that must be deployed in evaluating thought experimental scenarios (Kauppinen 2007; Ludwig 2007). Perhaps a responsible rationalism must commit to finding ways to improve the use of intuitions by way working hard to uncover sources of potential errors in the production of philosophical intuitions. I remain skeptical of this response. While 'intuition pumps' sometimes function as useful cognitive calisthenics, and while they sometimes lead us to better ways of theorizing, they just as frequently solidify philosophical prejudices, serve as justifications for theoretical commitments acquired in graduate school, and further entrench the dominant trends in our discipline.

Consider the purported implausibility of thinking that the system constituted by John Searle and his Chinese room is capable of understanding Chinese (Searle 1982). Searle won many converts to his neuro-chauvinist view of subjective experience with the thought experiment constructed around this situation; and this thought experiment spawned numerous commonsense and philosophical objections to functionalism. But some philosophers and cognitive scientists (e.g., Dennett, Minsky, Papert, and Simon) immediately began to argue that it *was plausible* to treat such a system—on the assumption that it could be constructed—as a clear locus of understanding. The fact that these people were deep in the midst of research projects in artificial intelligence and computational modeling played an important role in producing their judgments. They had a wide array of previously encountered and previously imagined situations upon which to draw in constructing the cognitive blends they used to evaluate Searle's thought experiment. People without this background couldn't construct the same kinds of mental spaces. But it's not at all obvious how to evaluate the intuitions that result from taking artificial intelligence seriously. Who is more accurate, the philosophical expert, or the expert in artificial intelligence? I have no idea how we could answer such a question a priori.

When a thought experiment is deployed in a philosophical argument, it can function as an intuition pump because it's embedded within a theoretical background that constrains the range of plausible strategies for fleshing-out its simplified and idealized representation of a possible world. But this doesn't always happen in 'epistemically benevolent' ways. Sometimes, previously defended theoretical commitments are taken up in the construction of a blended mental space. Sometimes a directed search for a blended mental space will confirm antecedently held principles, leading to the construction of answers grounded on little more than 'wishful thinking'; and where this happens, the expression of an answer can conceal a tacitly and unconsciously circular argument.[4] But there are no clear ways to discover which background conditions are relevant to interpreting a thought experiment. Consider the suggestion that the people of China could implement the functional architecture of a human mind (a situation independently proposed by two philosophers with very different views on the scenario). When DHM Brooks (1986) developed this scenario, he offered it as support for his claim that a properly organized group could experience any psychological state whatsoever—including drunkenness; Ned Block (1978), by contrast, used this scenario to support his claim that consciousness cannot be functionally

[4] Thanks to Jonathan Weinberg for pushing me to clarify this point.

realized since there is nothing that it is like to be the nation of China. Regardless of how this scenario might affect people when presented on its own, the existence of such divergent interpretations makes it clear that a person's theoretical background is playing a critical role in the interpretation of a thought experimental situation.

These facts bring us to the critical insight about how typical human minds are likely to function outside of the laboratory, outside of the philosophy room, in ecologically valid contexts. In the world we all inhabit, the socially significant judgments we make everyday are likely to be embedded in representationally rich social environments. Such environments provide a rich array of corrective interpersonal feedback, and they are populated with material anchors and background assumptions about the social norms that are at play in various highly structured communities. The absence of corrective feedback and interpersonal engagement is likely to generate problems with the evaluation of philosophical thought experiments in one-off cases. But it is important to note that our attempts to answer philosophical questions often depend on our ability to entertain imagined corrective feedback. Even where there is no overt social engagement, the attitudes of lone introspectors are likely to be attuned to the kinds of responses that would be offered by their colleagues, friends, mentors, church leaders, and critics. As they evaluate their intuitions for plausibility, they are likely to reflect upon, and perhaps even re-evaluate their initial judgments in light of assumptions about the form that corrective feedback would take if their answers were submitted to scrutiny. In some cases, people will see their initial responses as open to critical scrutiny because they have been bombarded with counterexamples, or because political or religious leaders who call dominant social institutions into question have influenced their engagements with the world. It would be truly surprising if this did not lead people to recognize that they had reason to mistrust, or at least to double-check, their initial responses. But, where this occurs, people may be led to offer responses to thought experimental scenarios that they are unsure whether they can avow as being their own. These stray answers are not merely failures to express initial intuitions, but revisions that people are likely to make in light of predicted corrective feedback.

I contend that many of the responses people give when they are presented with thought experimental scenarios are likely to be 'guesses', 'stray answers', and non-consciously constructed 'intuitions', no matter how they show up introspectively. Appreciating the value of answers and intuitions requires a shift in our theoretical perspective. Over the years, many social psychologists have come to realize that their methods are inapt for uncovering the mechanisms responsible for producing responses to survey studies. What such experiments uncover are gross patterns in socially significant behavior (Wegner & Gilbert 2000). Like these social psychologists, we must recognize that neither answers nor intuitions can serve as evidence about the nature of the world in which we live. When we examine the answers that people give in response to philosophical thought experiments, we should attend to the wide variety of ways in which extraneous variables affect our coarse-grained strategies for incorporating information *at some point* as we evaluate unfamiliar situations. My suggestion—a point which I shall have to develop elsewhere—is that the analysis of intuition pumps can be employed in the service of a kind of heterophenomenology, the interpretive method of cataloging overt speech acts, systematizing them as far as is possible, and then generating an account of how things are likely to hang together from the perspective of commonsense psychology. But that is a project for another day.

Acknowledgments: These ideas have been percolating for a long time, and I hope that I have not forgotten anyone who offered helpful guidance on this paper. My initial thoughts regarding the possibility of using research on blended mental spaces to think about the judgments that people make about thought experimental scenarios were inspired by conversations with Dan Dennett and Ed Hutchins. James Mattingly and John Bengson provided helpful feedback on an early version of this paper; Jonathan Weinberg and Liam Kofi Bright read more recent drafts; and Ruth Kramer listened to the arguments that I develop here in conversation. All of these people offered insightful criticisms, ideas that strengthened my arguments, and helped me to clarify my claims about the nature of philosophical judgments. As always I would also like to thank all of the

anonymous reviewers of this paper; I appreciate the time they have taken to read it and to offer comments, even where I have disagreed with them.

Works cited:

Addis, D., Wong, A. & Schacter, D. (2007). “Remembering the past and imagining the future: common and distinct neural substrates during event construction and elaboration. *Neuropsychologia* 45, 1363–1377.

Akins, K. (1996) Of sensory systems and the ‘aboutness’ of mental states. Journal of philosophy, (7): 337-372.

Bargh, J., & Chartrand, T. (1999). The unbearable automaticity of being. American Psychologist, 54, 462-479.

Bartlett, F. (1932). Remembering: A study in experimental and social psychology. Cambridge: Cambridge University Press.

Bengson, J. (2013). Experimental attacks on intuitions and answers. Philosophy and Phenomenological Research.

Block, N. (1978). Troubles with Functionalism. Minnesota Studies in the Philosophy of Science 9: 261-325.

Brooks, DHM. (1986). Group Minds. Australasian Journal of Philosophy 64: 456-70.

Buckner, R. & Carroll, D. (2007). “Self-projection and the brain,” Trends in Cognitive Science 11, 49–57.

Carmel, D. (2011). Experimental Philosophy: Surveys alone won't fly. Science, 332, 1262.

Carruthers, P. (2009). How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition. Behavioral and Brain Sciences 32 (2): 121-138.

Chomsky N. (1965). Aspects of the Theory of Syntax. Cambridge: MIT Press.

Dennett, D. (1995). "The Unimagined Preposterousness of Zombies," Journal of Consciousness Studies, vol. 2, no. 4, 322-326.

Dennett, D. (1988). "Quining Qualia," in A. Marcel & E. Bisiach (eds) Consciousness in Modern Science, Oxford: Oxford University Press.

Epley, N., & Gilovich, T. (2006). The anchoring and adjustment heuristic: Why the adjustments are insufficient. Psychological Science, 17, 311-318.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: An examination of self-generated and experimeter-provided anchors. Psychological Science, 12, 391-396

Fauconnier, G., & Turner, M. (2002). The Way We Think. New York: Basic Books.

Festinger, L. (1956). A theory of cognitive dissonance. Evanston: Row, Peterson.

Festinger, L., H. Riecken, & S. Schachter (1956). When Prophecy Fails: A Social and Psychological Study of a Modern Group that Predicted the Destruction of the World. University of Minnesota Press

Gendler, T. (2007). Philosophical Thought Experiments, Intuitions, and Cognitive Equilibrium. Midwest studies in philosophy, 31, 68-89.

Haidt, J. (2001). The emotional dog and its rational tail. Psychological Review, 108, 814-834.

Huang, J. & J. Bargh (2014). The selfish goal: autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences*, *37,* 02, 121-135.

Huebner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the cognitive sciences,* 9 (1): 133-155.

Huebner, B. (2011). Critiquing moral psychology from the inside. Philosophy of the social sciences, 41, 50–83.

Huebner, B. (2012). Reflection, reflex, and folk intuitions. Consciousness and Cognition.

Huebner, B. & D. Dennett (2009). “Banishing ‘I’ and ‘we’ from accounts of metacognition,” *Behavioral and Brain Sciences*, 32: 148-149.

Huebner, B. (in press). What is a philosophical effect? Models of data in experimental philosophy. *Philosophical studies*. DOI: 10.1007/s11098-015-0469-2

Hutchins, E. (2005). Material anchors for conceptual blends. *Journal of pragmatics*, *37*(10), 1555-1577.

Jackendoff, R. (1996). “How language helps us think,” *Pragmatics and cognition*, 4, 1-24.

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

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108, 480-498

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

Nisbett, R. & T. Wilson (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.

Putnam, H. (1975). The meaning of meaning. Minnesota Studies in the Philosophy of Science, 7: 131-193.

Schacter, D. & Addis, D. (2007). The ghosts of past and future. *Nature* 445, 27.

Schacter, D. & Addis, D. (2008). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. in J. Driver, P. Haggard, & T. Shallice (eds.), *Mental Processes in the Human Brain* (pp. 27-47). Oxford: Oxford University Press.

Schacter, D., R. Benoit, F. DeBrigard, & K. Szpunar (2015). Episodic future thinking and counterfactual thinking. Neurobiology of Learning and Memory, 117, 14-21.

Scholl, B. (2008). Two kinds of experimental philosophy, and their methodological dangers. Paper presented at the SPP Workshop on Experimental Philosophy.

Schwarz, N. (2007). Attitude construction: evaluation in context. Social Cognition, 25, 638-656.

Searle, J. (1982). Minds, Brains, and Programs. Behavioral and Brain Sciences.

Swain, S., Alexander, J. & Weinberg, J. (2008). The instability of philosophical intuitions: Running hot and cold on Truetemp. Philosophy and Phenomenological Research, 76: 138-155.

Wegner, D. (2002). The illusion of conscious will. Cambridge, MA: MIT press.

Wegner, D., & D. Gilbert. (2000). Social psychology—The science of human experience. In Subjective experience in social cognition and behavior, edited by H. Bless and J. Forgas. Philadelphia: Psychology Press.

Wegner, D. & T. Wheatley (1999). Apparent mental causation: Sources of the experience of will. American Psychologist, 54, 480-492.

Weinberg, J. & J. Alexander (2014). The challenge of sticking with intuitions through thick and thin. In *Intuitions*. A. Booth & D. Rowbottom (eds). Oxford: Oxford University Press.

Weinberg, J., S. Nichols, & S. Stich. (2001). Normativity and epistemic intuitions. Philosophical Topics, 29: 429-460.

Williamson, T. (2007). The philosophy of philosophy. Oxford: Blackwell.

Williamson, T. (2011). Philosophical expertise and the burden of proof. Metaphilosophy, 42, 3, 215-229.

Wilson, T. (2002). Strangers to Ourselves: Discovering the Adaptive Unconscious. Cambridge: Belknap Press.