Thought Experiments and Intuition Pumps in Philosophy

Keith Dowding

Abstract

I make a distinction between two sorts of imaginary case: thought experiments and intuition pumps. The first are essentially deductive, do not involve counterfactuals, the story strictly irrelevant for the conclusion. We project once, from conclusion to other cases for thought experiments. Intuition pumps are essentially projective, counterfactual and whose narrative details are vital for the conclusion. As counterfactuals, intuition pumps are not straightforwardly truth-functional. Counterfactuals’ pragmatic ambiguity depends upon context with the first projection vital for the second projection to other cases. In like-minded communities intuition pumps can be an important conceptual strategy. However, prior beliefs determine the nature of the projection, sometimes the first projection, the ‘intuition’, is all. Thus the epistemic status of the conclusions drawn from the two types is very different. The distinction drawn is sharp but there are many cases sharing features of both types and some can be usefully treated as either.
Introduction

The methods of analysis in philosophy are coming under increasing scrutiny. One long-established method is the use of imaginary cases, sometimes called thought experiments or intuition pumps. I will distinguish between several different types of imaginary case, arguing that the differences in their logical form mean we need to differentiate the epistemological significance of their evidence. Given those differences, we need to be careful in how we assess their conclusions and how, methodologically, philosophers should use them.

I distinguish between two major types of imaginary case (my term for the general class): the first I will call ‘thought experiments’ and the second ‘intuition pumps’. As persuasive or explanatory techniques, the two types are very different. Table 1 below summarizes the important differences between the two paradigm forms of imaginary case. However, I will also give examples of cases that share features of both intuition pumps and thought experiments, and suggest that some imaginary cases can be treated as either a thought experiment or an intuition pump. I will argue that the different epistemological status of each form and as they are used is vital for how we should treat and analyse their conclusions.

Thought Experiments

What I call ‘thought experiments’ is a smaller class than that which usually goes under that name in the literature. By my stipulation, thought experiments have four characteristics that distinguish them from intuition pumps. First, they are essentially deductive. Second, the details of the narrative that goes with them are strictly irrelevant for their conclusion. Third they are not counterfactual. Finally, their conclusions tend to be counterintuitive.
Table 1
Comparison of Thought Experiments and Intuition Pumps

<table>
<thead>
<tr>
<th>Thought Experiments</th>
<th>Intuition Pumps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essentially deductive</td>
<td>Essentially projective</td>
</tr>
<tr>
<td>Details of the story strictly irrelevant for the conclusion</td>
<td>Some narrative details constitute the essential elements of the argument</td>
</tr>
<tr>
<td>Not counterfactual</td>
<td>Counterfactual</td>
</tr>
<tr>
<td>Conclusion counterintuitive</td>
<td>Conclusion an intuitive projection</td>
</tr>
</tbody>
</table>

Let me begin by giving two well-known examples of thought experiments, both drawn from natural science, where thought experiments have an accepted and important home. I will use them as paradigm examples to illustrate the four features I have given.

Galileo Galilei’s thought experiment, that I call *Joining the Objects,* involves showing that two objects passing through a medium of the same resistance will fall at approximately the same speed (dependent on size), and hence at the same speed in a vacuum, contrary to received wisdom from Aristotle, who suggested that bodies of different weights had different natural velocities. Galileo demonstrates this through a dialogue that involves Salvador leading Simplicio into a contradiction.

*Joining the Objects*

Salvador asks Simplicio whether he admits that falling bodies acquire a definite speed fixed by nature that cannot be increased except by force or resistance. Salvador then asks what happens when two bodies whose natural speeds are different are united: will the more rapid be retarded by the slower and the slower hastened by the more rapid? Simplicio agrees that is indeed what will happen. Salvador then points out that
the two objects united weigh more than the heavy one alone, hence the new object should fall more rapidly than either of the initial two, a direct contradiction. Simplicio admits, ‘I am all at sea because it appears to me that the smaller stone when added to the larger increases its weight and by adding weight I do not see how it can fail to increase its speed or, at least, not to diminish it.’ (Source: Galilei 2012/1638, First Dialogue)

Galileo’s thought experiment is used to demonstrate a contradiction in a previously held view and to set up an argument for a different theory. But thought experiments may also be used to demonstrate an important positive result. Virtually all of Einstein’s positive results emanated from thought experiments. His work in both special and general relativity used thought experiments both prior to and as part of his mathematical demonstrations. In *Train*, Einstein shows that, given that relativity says that objects can only be said to be in motion relative to each other, two events can only be said to be simultaneous in relation to some location (Einstein 1961/1916). That is, there is no absolute sense of simultaneity. As well as demonstrating an implication from relativity, *Train* also knocks down a common-sense expectation.

*Train*

Suppose a very long train is travelling along a railway line with constant velocity \( v \). Two events at points A and B on the embankment will appear to occur at particular points (A and B) along this very long train. So simultaneity can be referenced relative to the train in terms of the embankment. Are two events that are simultaneous with reference to the embankment also simultaneous with reference to the train?
If lightning bolts strike the embankment at A and B and observer M₁ halfway between them sees them as simultaneous, the bolts will also be events at A and B on the train. Will person M₂ on the train also see those flashes as occurring simultaneously?

Whilst person M₂ coincides with M₁ just when, from M₁’s viewpoint, the lightning strikes simultaneously, M₂ is moving towards B and away from A with velocity \( v \), and so he will see the flash from B before he sees the flash from A. From the point of view of the embankment, flashes A and B are simultaneous; from the point of view of someone on the train, they are not. So events that are simultaneous with respect to the embankment are not simultaneous with respect to the train. Relativity says there is no way to decree that the embankment is at rest and the train in motion, only that they are in motion relative to each other. Thus there is no means by which to say that two events are
absolutely simultaneous, only that they are simultaneous relative to some point. (Source: Einstein 1961/1916, pp. 25–7)

**Essentially Deductive**

In my account, thought experiments are essentially deductive. By ‘essentially deductive’ I mean it is the deduction that provides the conclusion. In *Joining the Objects* Galileo proceeds by demonstrating that the claim that objects have a natural speed of descent cannot be true by use of *reductio ad absurdum*. For if two bodies are joined whose natural speeds are different, then the combination of speeds should retard one and speed the other: the speed of the joined object should be some average of the two. However, once joined, the two objects weigh more than the heavier alone, so according to Aristotle should fall faster than either of the two objects. From our initial assumptions, we have deduced a contradiction.

In *Train* we are invited to see that, given a fast-moving train is racing away from one lightening bolt to another, and light has a speed, one flash has to catch up with the observer on the train, whilst that observer is racing towards the second flash. The observer on the bank is stationary with regard to the flashes, so their equidistance means she sees them simultaneously whilst the observer on the train does not. We are reminded that relativity theory tells us that the embankment and train are in motion relative to each other, but neither can be claimed to be at absolute rest. So we deduce that the two flashes of lightning are simultaneous only relative to some point; and hence simultaneity of time is a relative not an absolute concept.

Of course, whilst the argument in a thought experiment is essentially deductive, we do generalize, induce or project from the conclusion. In *Joining the Objects* we generalize from the case of two objects to all objects. That is an
uncontroversial projection, but sometimes the generalization or projection to others case is problematic. However, it is the deduction that is essential element of the thought experiment any problems that follow with projection occurs with their application.

Details of the Story Are Strictly Irrelevant

In thought experiments the details of the story can be thought to be strictly irrelevant to the veracity of the conclusion. This seems to be clear in the case of Train. Einstein does not need a train or bolts of lightning, or even two observers who are moving relative to each other. The story could have involved, say, two planets moving relative to each other, and two supernovae, and rather than observers we could have imagined a stream of photons landing on the surface of the planets. Or we can simply represent it as a set of propositions: (1) two events, x and y, seem instantaneous from point A; (2) relative to point A, point B is moving towards x and away from y; (3) therefore light from x reaches B prior to when it reaches point A, and light from y reaches B after it reaches point A. (4) Thus what seems instantaneous at point A does not seem instantaneous at point B.

Similarly, we can represent Galileo’s thought experiment as a set of propositions (Gendler 1998). (1) Natural speed is mediative: that is, if A has a natural speed $S_1$ and B has a natural speed $S_2$, then, combined, the speed will be between $S_1$ and $S_2$. (2) Weight is additive: if A has weight $W_1$, and B has weight $W_2$, then the combined weight will be the sum of $W_1$ and $W_2$. (3) So natural speed cannot be proportional to weight, since a mediative property cannot be directly proportional to an additive one.
It is through such representations that John Norton (1991, 1996) claims that thought experiments can always be represented as simple arguments – that is, the narrative structure can always be removed and the same conclusion reached. Norton does not deny that the narrative in thought experiments performs a rhetorical or psychological role: it helps us understand the point being made. In other words, giving an argument in thought experiment form makes it easier for us to grasp the lesson. But he maintains that, as a matter of logic, they do not contain anything more than the argument.

Norton’s account has been criticized on three grounds. The first is that not every thought experiment is deductive and it is harder to represent projective examples as simple arguments (Cooper 2005). My category of thought experiments is narrower than Norton’s, so avoids that charge. A second is that different arguments have been associated with the same thought experiment. However, Bishop (1999), who makes this case, illustrates it with Einstein’s Clock in Box thought experiment, used to challenge Heisenberg’s uncertainty principle. In Bishop’s telling, Einstein and Bohr held different arguments associated with it, but that is not the case. At first, Bohr had to agree with Einstein, but the following day he showed that when one considers practical measurement issues (the basis of the uncertainty principle), the thought experiment has another conclusion. Einstein was forced to agree with Bohr.

The third is a more elusive claim and it applies directly to examples such as Train and Joining the Objects. As Gendler (1998) represents it, the thought experiment contains background assumptions that seem acceptable to us, because they are actual world conditions that are missing from the pared-down argument. Particulars such as the objects or the flashes of lightning, which are familiar and therefore manifestly understandable, are used to lead us to the conclusion. We are able to separate the
essential features of the narrative from the inessential. That is, once we see the point in *Train*, we can then see that it does not matter whether it is a train or a bus, or flashes of lighting or some other events simultaneously seen from one point.

Gendler might be correct in the sense that, psychologically, the narrative leads us to grasp the point more swiftly, and might do so because of unspecified background features. However, I take it that these conditions play a rhetorical or psychological role in thought experiments’ persuasive success. Nevertheless, the characteristics of thought experiments that demarcate them from other types of imaginary case is their deductive character; and that deductive character is important in how we assess their conclusions. I take the truth in Norton’s claim that thought experiments can be represented as more formal arguments to show that the thought experiment is deductive, but we can recognize that all natural deductions generally require some accepted background conditions that are not explicitly expressed in the argument. Indeed the breakdown of some deductive arguments, such as *Newton’s Bucket* (which was an actual experiment not one in thought alone), comes from hidden assumptions within the deduction. Or at least that is one way of representing Mach’s critique (Norton 1996). However, to the extent that the narrative of a thought experiment is not vital to the conclusion, the projections are not projections of the thought experiment itself.

*Not Counterfactual*

Norton (1991) says that thought experiments are necessarily counterfactual, since they are not actual experiments. But they are actual thought experiments and the thoughts are not counterfactual ones. We might say thought experiments are actual experiments, but ones conducted in thought. Indeed many believed Galileo had
carried out a physical experiment. Of course, the story that defines a particular thought experiment as that thought experiment is counterfactual but by the account above the detail of the story is not important to the derived conclusion. It is for this reason that the conclusions of thought experiments are non-problematically truth functional. If the assumptions as contained in the thought experiment are factual, the conclusion is true. There might be pragmatic ambiguity in how that conclusion is projected on to other cases, but the conclusion itself is non-problematically truth functional.

*The Conclusion Is Counterintuitive*

The point of a thought experiment is to bring new knowledge. We could provide thought experiments that demonstrate what we already know and accept. But there is no point to such an exercise. Thought experiments have their charm in the fact that they demonstrate a new truth, that seems unassailable once explained, but that was denied, unknown or not thought of previously. Indeed, many thought experiments provided conclusions that were surprising when first elucidated but which to later generations seem rather obvious.

Galileo’s thought experiment might be an example of that; and in my own life the simple Prisoners’ Dilemma (discussed below) has that quality. When I explained it to undergraduates 25 years ago, it always surprised them and often initiated some pushback. Now it is usually accepted without comment. So, unlike my first three conditions, it is not a necessary feature that the conclusion of a thought experiment is counter-intuitive, however, this fact about them generally contrasts them with intuition pumps that rely upon their intuitive nature.
**Intuition Pumps**

Daniel Dennett (1980) introduced the phrase ‘intuition pump’ and initially, perhaps, the term was somewhat pejorative; though Dennett (2013) later celebrates intuition pumps as a vital part of the philosopher’s armoury. The name was designed to bring out a feature of imaginary cases where the conclusion of a narrative immediately strikes us as obviously true: ‘little stories designed to provoke a heartfelt, table-thumping intuition – “Yes, of course, it has to be so!”’ (Dennett 2013, pp. 5–6), however, that does not distinguish intuition pumps from thought experiments in my account. Such table-thumping insights can equally come from deductions. I take the name because in my account intuition pumps work by projecting from past experience onto the narrative to reach their conclusion. Intuitions (as discussed briefly below) work through such projections.

It is this initial projection that makes intuition pumps essentially projective. Intuition pumps are generally used either to challenge the definition of some concept, or to challenge the principles of a theory, often a normative theory. In the former case the intuition pump is developed in order to suggest that the term as defined seems inappropriate in this case. In the latter the intuition pump is used to suggest that theory must be wrong since its recommendations seem inappropriate in this case. In both the procedures intuition pump is used like an empirical crucial dis-confirmatory case study. Indeed John Rawls (2001) explicitly draws an analogy between using our moral intuitions and the experimental methods of natural science. Sometimes intuition pumps are used more as confirmatory evidence, to illustrate a distinction that the author is making.

The important difference between intuition pumps and natural experiments however is the nature of the evidence. With thought experiments the conclusion
follows strictly from the premises. With intuition pumps the conclusion follows only given what seems to us to be the case. Our intuitions are used as a data point to challenge the definition of the concept or challenge the theory. But our intuitions are based upon our past experiences. We are trying to fit the case study into our current beliefs. We project from our past experience onto the current imaginary case and then use that to assess the concept or theory that the case is set to challenge or illustrate. Sometimes such projection is relatively uncontroversial, at other times it is. Indeed, to the extent that such projection simply reinforces what each warring party already thinks the intuition pump is only useful as a rhetorical aid to persuade the uncommitted. The better the story, the better its persuasive force.

Since the projection relies upon the story some details of the story are vital for the intuition pump to successfully convince. (That is not to say that there could not be an equally good intuition pump with different vital details making the same point; but it would be a different pump. It is an empirical question whether two such pumps are equally effective in inducing the same conclusion in all audiences.) Indeed, it is those very details that make the imaginary case an intuition pump. Importantly, intuition pumps are counterfactual. As such, they entail pragmatic ambiguity: that is, the interpretation of an intuition pump depends on its precise context.

I shall develop this point below, but note here that the pragmatic ambiguity of counterfactuals is closely related to their projective feature. Intuition pumps require us to project previous familiar patterns on to an unfamiliar, indeed unusual, case to reach a conclusion, one that is, moreover, then generalized to reach a further conclusion. Thought experiments require projection only from the conclusion to other cases, intuition pumps require projection to reach their conclusion and then further projection to generalize to other cases.
Given those four features, however, intuition pumps are diverse. A full discussion of that diversity I leave to another occasion. I discuss several examples. Let us begin with a favourite example in undergraduate moral philosophy classes, Surgeon.

Surgeon

You are a surgeon with five patients who will soon die if you cannot provide transplants. Two each need a lung, two a kidney, and one needs a heart. Luckily another patient has just arrived for a yearly check-up, who is a perfect genetic match for all your patients; you could whip out his lungs, kidneys and heart to save the five, though of course he would die as a result. Should you kill the healthy patient to save the five others? (Source: Thomson 1985, suggested by comments in Foot 1968)

This intuition pump is usually presented in the interrogative form as I have expressed it here. The reader is asked a question to which the answer seems obvious: the surgeon should not operate. (Though not so obvious that everyone gives that answer.) Surgeon is often used as a critique of consequentialist or utilitarian political philosophy though was first used as part of a discussion of the doctrine of double effect. In both uses it clearly seems to operate as a crucial case study that shows either utilitarianism gets the wrong answer, or as part of an argument about the limits of double effect. The narrative details are clearly important. In the structurally similar intuition pump Trolley (see below) people are much more likely to agree that one person can be killed to save five. Indeed the point of Foot’s original use of the cases

---

1. This is another feature of intuition pumps, they can project different intuitions and can be used for different purposes, but I do not discuss this further here.
was to interrogate differences in the doctrine of double effect. When it is acceptable to kill someone as a side effect of saving others. The point of the example comes from the story.

The story is counterfactual. We are being asked what we would do, or what should be done, in these cases. When we are asked this we are taking our prior beliefs based on our previous experiences and projecting onto the current case. We then reach a conclusion on that case. This has the hallmarks on an intuition as generally understood by psychologists.

There is no real agreement in philosophy, or indeed in psychology, over the precise nature of intuitions. That in itself might lead us to suspect that what we call ‘intuitions’ might be a basket of similar experiences. However, some consensus exists, among both philosophers and psychologists, that intuitions are non-inferential moves that we make to reach beliefs or dispositions to believe. Psychologists tend to view intuitions as part of our ‘fast thinking’ within the dual-process mind, which is what makes them non-inferential, or at least not consciously inferential. Studies of expert intuitions – where experts (for example in chess and fire fighting) seem to make better judgements then non-experts, but are unable to precisely explain or justify those judgements – suggest that experience allows people to identify patterns in previous data which they now project on to new data. This is how we can view our initial reactions to cases such as Surgeon or Trolley.

However, we can also reflect on those initial reactions and modify them. The projections do not have to be ‘fast thinking’ or even necessarily intuitive. For example, we might at first think that the surgeon should not operate, but then deduce from our utilitarian moral beliefs that he should. And then perhaps use some further utilitarian thoughts that suggest on rule-utilitarian grounds he should not. Deduction is not
ruled out in intuition pumps and they can be made inferential. However, the essential quality of *Surgeon* is the story, and how we project from that story to the conclusion using our other background beliefs. Someone who concludes on utilitarian grounds that the surgeon should operate will not convince others to be utilitarian. She will simply have used a different projection. Nor will she necessarily convince others that surgeon is not a counterexample to utilitarianism because of some rule-utilitarian get-out, because those who immediately intuit that the surgeon should not operate will not feel that the conclusion is something that should be inferred. The conclusion relies upon being intuitive.

*Trolley*

A trolley (or tram) is running out of control down a track. It will hit and kill five people on the track unless you pull a lever beside you that will send it down a spur. However, there is another person on the spur track and by diverting the trolley you will kill that person. What should you do?

(Source: initially Foot 1978; Thomson 1985)

Intuition pumps in moral philosophy might seem to have a different character than in other branches of philosophy. The intuitions they pump are decidedly normative. I think this makes a difference to their assessment though do not pursue that claim here (see Jackson 1998 for the grounds of an opposing view). My third example draws from the philosophy of mind, where intuition pumps also have a natural home (possibly because they too have a normative character, tracking issues of free will and the nature of consciousness). This imaginary case is the one that Dennett was discussing when he introduced the term ‘intuition pump’ into philosophical vocabulary. John Searle’s *Chinese Room* does not provide such a strong intuition as *Surgeon*. 

15
A person in a locked room who knows no Chinese has a set of commands written in English allowing him to correlate a set of formal symbols with another set of formal symbols. These rules allow him to respond in written Chinese to a set of questions written in Chinese so that the Chinese question posers are convinced the man understands Chinese. Thus any computer program that enables a computer to have a conversation with someone in Chinese would not understand the conversation. Without understanding the machine cannot be thinking and thus does not have a mind. (Source: Searle 1980)

This intuition pump might look like a thought experiment. Do we not deduce that a computer program cannot have a mind from the fact that a person following a set of rules does not have a mind? As far as we can deduce that something as described in the Chinese room does not have a mind, perhaps it is deduction. But Searle concludes from the example that no mind can simply be a computing device. That is the physicalist claim he is disputing. And that projection relies upon the projection that is induced in details of his story.

Searle’s conclusion depends upon the idea that complex computing devices that constitute the human brain are analogous to the set-up in his Chinese room. We have to project from the room with its human occupant to the human brain. It is only from that projection that the conclusion is plausible. And here the projection from his specific set-up is important. Were Searle simply to describe a machine or a brain where the neurons are not conscious and conclude therefore the brain cannot be conscious his imaginary case would not work. In that case he would simply be describing what physicalists describe when they make their claim. Thus the specifics of
the story are vital for the success of the intuition pump. Of course, the language does not have to be Chinese, and we might change the setting, but it does depend upon the operation being conducted by a person who does not understand the output of the operation. It is the contrast with a person inside the machine, who has a mind, but whose mind is not important to the output, that brings on the idea that the machine itself cannot have a mind. That is the detail that pumps the conclusion. The projection requires us to contrast the conscious mind of one of the cogs, to the lack of understanding of that cog to the output, and thence to the lack of understanding of the cogs as a system. We cannot deduce the conclusion from the premises alone.

Another famous intuition pump also directed against physicalism in the theory of the mind is Frank Jackson’s Mary.

Mary

Mary is confined to a world with only black and white experiences. She becomes a brilliant scientist who learns all the physical facts about perception. One day she leaves her world and experiences colours for the first time. She sees red for the first time. Clearly she has learned something new. Since she already knew all the physical facts about colour, she must have learned something non-physical when she first experienced colour. Therefore, her knowledge of what other minds experience when they see red is expanded from her knowledge of physical descriptions. Physicalism must be wrong. (Source: Jackson 1982)

Mary is an intuition pump because it relies upon us agreeing that indeed when Mary sees red for the first time she has learned something new. It relies on projecting the thought that her phenomenal experience of colour must be qualitatively different from
all other knowledge she could have from physical facts about perception. Our experience of colour seems different in quality from facts we might know about perception and the nature of colour. And that is what this intuition pump relies upon.

The intuition pumps works on the ground that it simply seems correct that the phenomenal experience of colour is not a physical fact on a par with what a scientist who studied colour could possibly know. Again, we can see that the conclusion cannot be deduced from the story, it just seems (at least to most people) that the experience of colour cannot be the same as all the physical facts. The details of the story are important. Of course, she does not have to be a ‘she’, nor called Mary, but we have to imagine a person who has never seen colour and who has all the physical facts at her disposal for the intuition Jackson desires to be pumped in his readers. We project from our experiences that when we see something that we have not seen before, of course it is a new experience. What its critics maintain is that we cannot project from what it means to have all the physical facts at our disposal.

Yet another intuition pump used against physicalism with regard to the mind is Zombie.

**Zombie**

Zombies are exact physical and functional replicas of conscious human beings apart from the fact that they have no inner mental life. So they are not phenomenally conscious beings. Since we can conceive of such zombies they are possible, physicalism is false since it maintains that the physical and functional properties of a person are logically sufficient to ensure consciousness. (Source: Block 1990)

At first this might seem like a deduction but as a deduction it relies on the assumption that what is conceivable is possible, and if zombies as described are possible, then it
cannot be logically, or conceptually, the case that consciousness is the same as physical description. Both the claim that what is conceivable is possible, and, that zombies are possible (in a relevant sense of conceivability given the former claim) are projections. They are counterfactuals and thus not straightforwardly truth functional. So problematic is the inference that Frank Jackson (as reported by McGeer 2003) describes the case as an intuition rather than an argument based on an intuition.

In all of these examples the conclusions are intuitive. They are projections from what we already think. Not all share the intuitions; physicalists certainly do not find these anti-physicalist intuition pumps convincing. That might be because they simply do not project the intuition or, if they do, they point out dis-analogies between the story as told and the application to which the argument is addressed. They deny the projection that is being made from the narrative. Whilst actual experiments show people do not all share intuitions in moral intuition pumps. Surgeon is the example so far given where we might find the greatest percentage of people who agree that the surgeon should not operate. That is not surprising: few of us can imagine our doctors operating on us without our permission. It is an easy projection to make.

**Mixed Cases**

I have offered up paradigm examples of imaginary cases that are relatively neatly classed as either thought experiments or intuition pumps. I will now present some cases that share features of both. Victoria McGeer combines the two intuition pumps Zombie and Mary to provide a problem for anti-physicalists in the philosophy of mind. She trades on a claim made by Chalmers (1996) that Zombie and Mary together provide a greater problem for physicalists than they do as independent cases. Chalmers suggests that physicalists admit Mary learns something about the world
when she sees red for the first time, but deny that that fact entails that physicalism is suspect. Physicalists also find it hard to deny that the conceivability of zombies leads to a denial of physicalism but then deny the possibility of zombies. However, he suggests, Zombie can support the inference from Mary to the denial of physicalism and Mary can explain what is missing in the lives of zombies. McGeer (2003) rebuts this argument with her Zombie Mary.

Zombie Mary

Zombie Mary is confined to a black and white room and there learns all the fact about the neurophysiology of colour experience. Then she leaves the black and white room and sees colour for the first time. Her visual system is affected in just the same way as human Mary. Does she learn any new non-physical fact about the world? As human Mary’s zombie counterpart, she reports a new visual experience. Given that, she, like human Mary, can judge that she did not understand other visual experiences in the way she claimed. However, if she is a zombie she must be wrong, since she cannot have such an experience herself. So she does not learn anything new about the world. So the fact that zombie Mary seems to learn something new about the world is not a reliable indicator that physicalism is wrong.

McGeer here uses the combination of two intuition pumps to produce a contradiction. Either the anti-physicalist must claim that zombie Mary has no new experiences, which does not licence us to claim that human Mary does, or the story does not licence the claim that zombies are truly conceivable, since they cannot be functionally and physically identical to humans. Zombie Mary utilizes intuitions from two intuition pumps in a reductio thought experiment.
**Zombie Mary** is a thought experiment that combines two intuition pumps. *Schrödinger’s Cat* is a thought experiment that relies upon a really strong intuition. Before setting it out, we ought to recall what it was designed to do. Quantum mechanics provides a set of equations that provide the underpinning of modern science. The equations describe the behaviour of very tiny objects and they provide the only description of these objects. The problem is that these equations suggest that nothing can be said about these objects until they are observed, which some people interpret as meaning that nothing is real about them. In the world of quanta everything is governed by probabilities. Famously, this claim disturbed Einstein. Edwin Schrödinger was equally disturbed about these implications. His imaginary case was designed to demonstrate its absurdity by showing that such uncertainty at the quantum level could spill over to the macro level.

*Schroedinger’s Cat*

A cat is penned in a sealed chamber. Inside the chamber there is a Geiger counter and a small piece of radioactive material. Over a time period \( t \), the probability that one atom will decay is 50 per cent. If an atom decays the Geiger counter will click triggering a hammer that will release a poison gas instantly killing the cat. At time \( t \), the psi-function of the system is expressed as the cat being equally alive and dead.

(Source: Trimmer 1980, which is an English translation of the original Schrödinger paper)

Since it seems obvious that the cat cannot be alive and dead simultaneously, Schrödinger hoped to persuade that there must be an underlying reality prior to measurement, disproving the Copenhagen interpretation.
Schrödinger’s Cat is an interesting imaginary case. Is it, in my terms, a thought experiment or an intuition pump? From known experimental effects of quantum theory and the Copenhagen interpretation, the conclusion that the cat is neither dead nor alive until an observer looks is deduced. So it is a thought experiment. However, the conclusion is so counter-intuitive that it seems to challenge the assumptions. However, unlike Joining the Objects, it does not provide a formal contradiction, it just seems weird. The conclusion is so counterintuitive it can be used to challenge the Copenhagen interpretation. Defenders of the Copenhagen interpretation simply take the conclusion on board admitting it is completely counterintuitive, but then, they point out so are the accepted experimental results of quantum effects. Science is not out to confirm our intuitions, but to challenge them with a deeper reality than our perceptual reality. And that perhaps demonstrates an important methodological or perspectival difference between scientific and philosophical method. Of course, physicists have not left the matter there. Some have challenged the narrative itself: after all, a cat is an observer, so it would know if it were dead or alive, whilst the Multiverse interpretation at least allows the cat to be both dead and alive in different worlds. (Tegmark (1997) suggests a Quantum Suicide imaginary case that demarcates the Copenhagen and later Multiverse interpretations of quantum theory. And there are many other such suggestions: see, for example, Annals 1995).

We also have imaginary cases that have been used as both a thought experiment and an intuition pump. There are many examples from game theory: the Prisoners’ Dilemma provides the most familiar example.

Prisoners’ Dilemma

Two criminals are captured and interrogated. The police have evidence to convict them of a minor crime, but will ignore that crime if they can
get a conviction on the major crime. Each prisoner is offered a deal: if you confess, ensuring your accomplice is convicted, you go free, and your accomplice will get the maximum tariff of eight years; if you both confess, you both get remission, so each will serve five years; if neither confesses, you will both be convicted of the minor crime and serve two years. In the normal form matrix below, the numbers stand for years in jail, but can be considered to be associated with an ordinal preference ranking (0>2>-5>8).

<table>
<thead>
<tr>
<th></th>
<th>Confess</th>
<th>Not Confess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confess</td>
<td>5,5</td>
<td>0,8</td>
</tr>
<tr>
<td>Not Confess</td>
<td>8,0</td>
<td>2, 2</td>
</tr>
</tbody>
</table>

Confessing is the best strategy, no matter what the accomplice does.
The dominant strategy (confess/confess) leads to a worse outcome for each player than the dominated strategy (not confess/not confess).

The Prisoners’ Dilemma derives from an experiment conducted by Merrill Flood and Melvin Dresher at the RAND Corporation in the early 1950s (Flood 1952), whilst the story associated with the payoff structure is attributed to Albert Tucker (Poundstone 1992, p. 8). The formal deduction that rational players will choose the suboptimal outcome was known prior to the narration. Treated as a thought experiment, it demonstrates a counterintuitive conclusion that has been used to illustrate collective action problems in various theoretical and empirical forms.

The deductive conclusion of the one-shot prisoners’ dilemma is now universally understood, but when it first came to the attention of social scientists and philosophers from the mid-1960s onwards, its conclusions (at least in application) were challenged. Here it was being used as intuition pump. The thought was that surely
prisoners in that situation could work their way out of the suboptimal conclusion. When used as an intuition pump, it is the story that is given precedence, not the structure of the game. When used as a thought experiment, it is the structure (as established by the utilities and the one-off nature of the one-shot simple game) that is given priority. When we give precedence to the story, what we do is to challenge the structure, in order to challenge the assumptions.

Generally speaking, our intuitions about the game is given by the fact that we rarely, if ever, face single-shot games and, once iterated, the folk theorem shows that any finite combination of outcomes is rational for infinitely repeated games (Binmore 2007, pp. 333–40). Thus the Prisoners’ Dilemma in repeated form can encompass any intuition we might have about a single outcome. But treated as an intuition pump, the story might also challenge the assumptions about the utilities assigned to people in such situations, or indeed about how rational people are in such situations. Some formal writers are frustrated with those who treat the Prisoners’ Dilemma as an intuition pump, and where that challenge is to the deductive structure of the game itself they are justified in being frustrated. But to think about whether the game is correctly applied to any given situation – that is, the applicability of the assumptions to a situation – can be a highly pertinent consideration. Here our intuitions about the story that illustrates the game might provide a better guide to projection than the game itself.

Thus how we treat game theoretic imaginary cases is rather important. There are two dangers. The first lies in giving too much precedence to the story used to illustrate the game; if the maths truly represents a situation, then we must follow through the deduction. On the other hand, the story might lead us to believe that the mathematical result applies to this or other similar situations, because features of the
story independently suggest the plausibility of the formal deduction (Thoma 2015). In
that case the formal result might be misleading if generalized over sets of behaviours.

So we have one example, *Schrödinger’s Cat*, which is a thought experiment but
was designed because its conclusion is just too counter-intuitive, and another thought
experiment the *Prisoners’ Dilemma* that can usefully be treated as an intuition pump.
Thought experiments, once understood, should (given acceptance of their
assumptions and no errors in the deduction) receive universal acceptance – of their
conclusions, at least, if not their applicability. *Schrödinger’s Cat* shows sometimes those
conclusions are too hard to swallow easily. Many intuition pumps do not receive
universal acceptance of their conclusion, or their applicability to other cases, though
some do receive more acceptance than *Schrödinger’s Cat*. Often intuition pumps are
designed by one side in a dispute to bolster their claim, but rarely do they persuade
the other side. Experimental evidence shows that people within cultures often do not
share the same intuitions; and the percentages on one or another side in such
intuitional divides vary across social classes, ethnic groups and culture (see, for
example, Haidt 2012; Greene 2013 in moral psychology; Weinberg et al. 2001 for
examples in other fields.) So how, philosophically, should we handle intuition pumps?

**Intuitions and Projectability**

The key difference between thought experiments and intuition pumps is the
counterfactual nature of the latter in contrast to the former. We project a conclusion
from the narrative as told, trying to fit the story into a familiar pattern. That is what
makes them intuitive. Counterfactuals do not straightforwardly sustain truth-values.
The pragmatic ambiguity of counterfactuals depends upon context and how we
project a conclusion from the detailed narrative. Where the point of an intuition
pump is to challenge our concepts, then we are forced to plump for one projection rather than another, lest we countenance inconsistency. Two intuition pumps that most successfully demonstrate the utility of such projection are usually known as the Gettier counterexamples. I describe the best-known one (I call it *Job Applicant*), which challenges the traditional doctrine that knowledge is justified true belief.

*Job Applicant*

(1) Smith and Jones have both applied for the same job. (2) The company president has told Smith that Jones will get the job. (3) Smith knows Jones has ten coins in his pocket. Together (2) and (3) entail (4) ‘the man who will get the job has ten coins in his pocket’. Smith acknowledges entailment (4) and by (2) and (3) is justified in believing it. It is also true that the person who gets the job has ten coins in his pocket. Since Smith believes (4), is justified in believing it and (4) is true, then given that knowledge is justified true belief, Smith knows (4). However, unbeknownst to Smith, he too has ten coins in his pocket and, despite what the president told him, Smith gets the job. Does Smith really know (4)? (Source: Gettier 1963)

It is almost universally accepted by analytic philosophers that Gettier has produced a strong counterexample to the claim that knowledge is justified true belief. We can think of *Job Applicant* as a test for a sufficient condition for a definition. Deutch (2010) points out that it can be represented formally as:

(1) There is an F that is not a G

(2) Hence, not all Fs are Gs

We do not need to represent it as:

(0) It is intuitive that there is an F that is not a G
(1) So, there is an F that is not a G

(2) Hence, not all Fs are Gs

To do so would be to represent all definitions or any aspect of conceptual analysis as intuitions. What is important, it seems to me, is our web of belief about the nature of knowledge, which involves other concepts such as ‘evidence’, ‘belief’, ‘justification’, ‘truth’ and ‘propositions’. In order to explain what knowledge is using these other terms, they need to be fitted together conceptually. What *Job Applicant* does is to suggest that the previous way of fitting them together does not work for all cases. Responses to Gettier include adding a condition to the definition of knowledge, putting conditions on justification, or on to the extension of propositions and other such moves. But all involve a conceptual shift.

That the counterexample is not universally held, as suggested by some cross-cultural empirical studies (Weinberg et al. 2001), seems irrelevant. We might simply respond that those who do not see its force do not have an analytic philosopher’s understanding of ‘knowledge’. For example, if someone claims that we have to know something before we can believe it, the response is likely to be that the person has simply misunderstood what analytic philosophers mean by ‘belief’ and ‘knowledge’. So, whilst most philosophers believe that *Job Applicant* shows that other conditions are required to define knowledge, even those who want to maintain that knowledge is justified true belief argue that we need some further conditions on what constitutes justification or on the nature of the beliefs that constitute knowledge claims.

Deutch is right, I think, that we do not have to understand the case as involving an ‘intuition’ in terms of some special experiential phenomenon; but equally we can represent it as a non-inferential claim that the F that is not a G is not a G. We have projected from past cases to the new one. We can possibly do a Gettier on
almost any concept; indeed in some fields that seems to be the sole purpose of intuition pumps. If natural language, in for example moral theory, is not logically closed, then we cannot expect to provide necessary and sufficient conditions for any concept to make it forever immune to counterexamples. Given the pragmatic ambiguity of counterfactuals, intuition pumps can only provide plausible projections for concepts or theories. That seems to be their major role in moral philosophy for instance. (Although in moral philosophy the added aspect is that the projection might constitute, rather than merely represent, its subject matter.)

Psychologists have shown that expert intuitions work best in areas where patterns are frequently repeated; without such repeated patterns, experts in a discipline make judgements that are no better, or not much better, than non-experts. This suggests that the more fantastic the imaginary case, the less useful our intuitions are likely to be, for the harder it is for us to project.

In response, philosophers might claim that, whilst psychologists see intuitions as fast thinking and non-inferential, when used in philosophy those immediate apprehensions are then reflected upon them in a decidedly conscious and slow-thinking manner. And for some philosophers, our reflections on some of our core

---

2 We can note here that results from experimental psychology, such as the percentage of doctors and hospital administrators who think the surgeon should operate in Surgeon, are not counterfactual. These survey results are facts. What we make of those facts is a separate issue from the use of intuition pumps in analytic philosophy. Philosophers, particularly moral philosophers, should not ignore them, but the evidential logic of such experiments is separate from that of intuition pumps as an analytic tool. Which is why I always refer to such uses as experimental psychology rather than experimental philosophy, a term that seems a complete misnomer.
beliefs show that we cannot inferentially defend them: they are simply core beliefs or ‘rational intuitions’. Bealer (1998; 2002) says our acceptance of core principles of logic such as *modus ponens*, cannot further be justified, but in order to rationalize we need to accept them. They just ‘seem right’, although – as he also acknowledges – some core mathematical intuitions can be shown to be false. And of course such intuitions are required for the deductions in thought experiments; they form part of our background knowledge or acceptance.

*Schrödinger’s Cat* is a problematic case for my distinction as it is a thought experiment that relies upon non-acceptance of its conclusion because it is so counter-intuitive. However, like some core mathematical intuitions that turn out be false, perhaps even that intuition could be wrong, though it is hard to see how it could be demonstrated to be wrong. Our previous patterns of experience are such that we cannot entertain the possibility of a cat being dead and alive simultaneously.

So some intuitions seem very secure. Others less so brought out by differential responses to intuition pumps. Those that are secure might be so because they are part of the pattern of our thoughts. Deduction, informally at least, is part of the way in which we think. The pattern is so ingrained when we think of its basis we cannot but think it is correct. Our earlier examples of intuition pumps, *The Chinese Room*, *Zombie* and *Mary*, try to tap into conceptual intuitions. Their critics find different aspects of the projections problematic. *The Chinese Room* subtly uses a projection that, if we were a cog in a machine, we would not know what is going on; therefore none of the cogs can know, so all the cogs together cannot. *Mary* taps into the fact we experience colour, so experiencing it for the first time must be new experience. Critics point out that we have to pragmatically project what a complete description of the physical facts actually means. As a counterfactual, there is no determinate answer. *Zombie* relies
upon on the intuition that conceivability entails possibility, but also upon the projection that there can be exact physical replicas of humans but who lack consciousness. Of course, those who think that consciousness is reducible to physics do not so project, and those who think it is not reducible do so project. As a counterfactual there is no determinate answer. And then we might discuss the host of intuition pumps that exist in moral philosophy. Perhaps the intuitions that are least secure are our moral ones. And that should not be a surprise since not only do different peoples pattern the moral world differently each of us is also tends to accepts plural moral values that might conceptually be fitted together in different ways with various weightings within those general patterns.

**Conclusions**

There are two types of imaginary case: those I call thought experiments and those I label intuition pumps. The former are deductive, need not involve counterfactuals, demonstrate and sustain truth-functional conclusions. The narrative is strictly unnecessary for the conclusion, but provides rhetorical force. That fact can be dangerous, since thought experiments might convince not through the logic, but the narrative. If the narrative does not, in fact, precisely mirror the logic, we might have two arguments leading to the same conclusion that might be confused. Nevertheless, this simply means we have to handle thought experiments with care, especially in the social sciences.

Whilst, logically, thought experiments do not have to provide counterintuitive results, there is little point in providing thought experiments to demonstrate what we already believe. Their counterintuitive nature is in contrast to intuition pumps, which are inductive, counterfactual and where the narrative constitutes the argument. As
counterfactuals, intuition pumps cannot demonstrate conclusions. They can only plausibly project. Some intuition pumps, such as Job Applicant, manage this so convincingly because they tap into webs of belief that show we need to make some changes to our concepts to maintain consistency. And some thought experiments, such as Schrödinger’s Cat have fantastic conclusions that many find hard to swallow. And of course, even thought the conclusions of thought experiments are truth functional, they do not have to be true. The fact that some intuition pumps seem to lead to such secure conclusions whilst some thought experiments do not helps explain why the distinctions I have made between thought experiments and intuition pumps have not been clearly seen previously. But furthermore we also have sets of mixed cases, thought experiments that are composed of intuition pumps, imaginary cases that can be interpreted either as thought experiments or intuition pumps or where the narrative that accompanies the deduction carry separate explanatory force and we have to be careful in how we interpret their results. Nevertheless, making the distinctions and examining any imaginary case in the terms I suggest will allow us to assess the epistemic value of their conclusions in a more systematic manner.

References


Block, Ned (1990), 'Consciousness and Accessibility', *Behavioral and Brain Sciences*, 13 (4), 596-98.


Deutch, Max (2010), 'Intuitions, Counter-Examples, and Experimental Philosophy', *Review of Philosophical Psychology*, 1 (4), 447-60.


Gettier, Edmund L. (1963), 'Is Justified True Belief Knowledge?', *Analysis*, 23 (6), 121-23.


