What does philosophy contribute to the study of the mind?

by Susanna Siegel

to appear in *The Philosophers' Magazine*, Winter 2020

Susanna Siegel is Edgar Pierce Professor of Philosophy at Harvard University. She is author of *The Contents of Visual Experience* (Oxford 2010), *The Rationality of Perception* Oxford 2017), and numerous articles about perception and its roles in knowledge and society.

Often when I meet someone outside my usual circles our conversation goes like this. I tell them I'm a philosopher and I study the mind. They say, Oh, so you have a lab? And I say: No, I've never done an experiment in my life. Still puzzled, they ask politely, So how do you get your results without experiments?

It's a reasonable question.

They probably wouldn't become any less puzzled on hearing that decades of productive debate and disagreement over the mind-body problem have relied on highly unrealistic fictional examples. Especially if they have the impression that any inquiry into the mind that doesn't deal primarily with experimental evidence is doomed to be an inferior cousin of science. From this point of view, it can seem obscure what philosophy can contribute to the study of the mind.

I'd like to make some types of inquiry in the philosophy of mind seem less obscure, especially for anyone who has this impression. The quickest way to dispel it would be to describe the kinds of philosophical inquiry into the mind that rely directly on experimental results at various points in their inquiries to draw broader conclusions about the how the mind works – which is quite a bit of philosophy.

A more challenging and rewarding route justifies the use of fictional examples in philosophical inquiry – even highly unrealistic ones. How anything fictional could bear on anything real is an important and challenging question in its own right, and much discussed in connection with idealizations in science. Experimental settings can be more or less ecologically valid, so analogous methodological questions arise there, too. In philosophy, when do unrealistic examples impede or mislead inquiry, and when are they innocuous or even productive?

Even philosophical inquiries that orient their questions around such examples instead of around questions that experiments can answer still have a lot in common with the experimental sciences.

We can find some of these continuities by focusing on three different roles for fictional examples in the philosophy of mind. Such examples can (i) pose loaded questions, (ii) illustrate a philosophical problem, and (iii) test hypotheses about the way things (including our minds) should be – normative hypotheses about what ought to be, and modal hypothesis about the way things could be or necessarily have to be, rather than descriptive hypotheses about the way things actually are. Discussing these roles bring into focus what makes an example realistic or not, when being realistic matters, when it doesn't, and why.

Relying heavily on fictional examples is not the only mode of inquiry in the philosophy of mind, and often the jobs done by such examples could be done by other things. But the role of invented situations brings out both the continuities and discontinuities with the experimental sciences. It also shows us continuities between philosophy of mind and the literary humanities. So discussing these roles for fiction in inquiry is a good start, I hope, to defusing the puzzlement about how both philosophy and the humanities more generally could contribute anything to the study of the mind.

Role 1: Pose a loaded question

One of the most influential thought-experiments in philosophy involves the philosopher Frank Jackson's color scientist named Mary. Mary knows all the facts about color science but has never seen any colors. Really -- never? Yes, never, because she has always lived in a black and white room. Yet she seems to learn something new when she finally sees red. What she knew before she saw any colors wasn't enough to reveal what seeing red is like.

The story ends there, with Mary emerging from her highly unrealistic situation.

Needless to say, someone couldn't live alone their whole life, and couldn't learn much of anything even from books. (Who taught Mary to read?) It also seems unlikely that a human being could avoid seeing anything colorful on her body, by pressing a fingernail or seeing her own blood or a bruise or an afterimage. The story barely makes sense!

But with all its unrealisticness, this story has stimulated much productive discussion about the mind-body problem. Mary is just a placeholder for a body of knowledge. If you focus on Mary purely as a knowing subject, ignoring the rest of her human features, her epistemic situation is easy to grasp. And that's the beauty of the example. It can seem plain that Mary learns something. The role of the story is to ask what, if anything, someone could learn by seeing red, if they started out knowing all that Mary knows – which is all the physical facts about color. As soon as it is granted that Mary learns something from seeing red, anyone who assumes (as physicalists do) that the physical facts are *all* the facts has some explaining to do. Since the rest of Mary's humanity is incidental, that's why the story can be unrealistic.

In studying the mind or anything else, when should we try to answer questions about unrealistic situations that as very much unlike our own? This question applies equally to idealizations in experimentation, ecologically invalid experimental circumstances, and philosophical thought-experiments. A plausible answer is: when and only when the answer tells us something about situations we are in. If the correct response to Mary's situation is that despite any initial reactions you might have to the scenario, she learns nothing, then physicalism is true. If the correct response is that she learns a new fact, then physicalism is false. If the correct response is more nuanced than either of these options allows, for instance if Mary somehow re-learns an old fact but in a new way, then the status of physicalism needs more discussion.

Sometimes, unrealistic scenarios won't tell us what we want to know about the situation we're actually in, and they'll fail to do that because they abstract away too much from factors that matter to the inquiry. I'll discuss an example like that in connection with the third role for fictional examples - testing normative hypotheses.

If Mary's humanity is incidental in Jackson's example, why bother making her human at all? Why not leave specific invented creatures out of the picture, and just ask in general whether if someone knew all the physical facts, they could know what it's like to see red?

This version of the question is far less vivid than the example involving Mary. It can be easier to think one's way into the question using examples, compared with pure abstractions.

The example of Mary occupies a special intermediate region between the concrete and the abstract. If you try to grasp the fully human side of Mary in any detail, you'll run into sand. You might also miss the important fact that the question "What does Mary learn?" is loaded. By contrast, the fully abstract, example-free version of the question is not quite so obviously loaded against physicalism.

A loaded question can be a good thing when it helps us feel the force of possible answers to difficult questions. Problems often take the form of simple questions with no straightforward answers. One way to understand that kind of problem is to understand the possible different answers to a question and what there is to be said in favor or against them. Understanding comes when you can identify specific possible solutions to a problem, even if you find that none of those solutions is perfectly adequate, and even if you're not sure at the end of the day which solution is correct.

So one contribution from philosophy to the study of the mind is its collective, systematic inquiry into problems. This type of contribution discusses multiple, competing answers to the same question, with some parts that criticize the answers, and other parts that defend them. Discussions of Jackson's Mary example have given rise to exactly that kind of collective discussion of what if anything Mary learns upon seeing red. The example posed a loaded question, and the subsequent various answers to it have improved our understanding of the mind-body problem.

Role 2: Illustrate a problem

A different role for examples is that they can illustrate problems all on their own, even before they unfurl reams of discussion. This role for fictional examples places different demands on how realistic the example should be to play its role well.

Here's a simple general question: What should we believe?

The answer to this question might seem straightforward: our beliefs should respect our evidence. We often rightly criticize people for holding beliefs that blatantly disregard available evidence. When people think that climate change isn't caused by humans, or that in Boston they are more likely to get Ebola than they are to get hit by a car, these beliefs are unreasonable, and they're unreasonable because they glaringly ignore what's known about causation and contagion.

Fictional examples can illustrate a problem with this answer. Imagine that two people come to you for career advice. Like anyone else, they have their hopes and fears. They're equally talented, and they have the same ambitions. They have learned what it is possible to achieve with their talents. They've also learned, from studying extensive evidence, that in their field of choice, which let's suppose is surgery, talented, determined people in their social group systematically underachieve compared with equally talented people who aren't in social group X. This fact bothers them. It bothers them greatly. But they respond to this evidence in different ways.

One of the young people is so bothered by the discrepancy between what equal amounts of gumption and grit can achieve that he decides to choose a different career from the one he'd most like to have. He figures that if he tries for what he wants, he'll just end up disappointed. So he scales back his horizons and lowers his ambitions. He's resigned, though his beliefs fit the evidence.

The other young person is also deeply bothered by the same thing. He responds to his feelings differently: he talks himself out of believing the evidence. This young person who believes against the evidence is fresh and determined, undeterred by the disturbing thought that the same amount of work done by someone from a privileged group will yield a better outcome than he can expect. He's buoyed by his resolve.

Here we have a determined person who believes against the evidence, and a resigned one whose beliefs fit the evidence. Both young people have the problem that what's useful for them to believe is in conflict with what they've got evidence for believing. Often times, we advise people to bring their beliefs in line with the evidence they have. But would you advise the determined person to lower his expectations?

The problem of what to believe is sharpest in cases of conflicts like this one.

We actually know quite a bit about this problem. We understand its contours -- its many possible solutions and their drawbacks, pitfalls, appealing features. Some people think they know the correct solution. For example, proponents of evidentialism think you should always believe in accordance with the evidence.

A solution to this problem would give you a principled way to respond to the conflict between two kinds of seemingly rational pressures – believing against the evidence when it will facilitate something that is proper to want, and believing in accordance with the evidence. If the solution (contrary to what evidentialists say) is that sometimes it is okay to believe against the evidence and other times is in't, then this (anti-evidentialist) solution would tell you which cases are which.

Here there is arguably a discontinuity between the sciences of the mind and the humanities. Analyzing problems not proprietary to the humanities, but the subject-matter in this case has been proprietary to the humanities. The closest thing to the subject-matter in the sciences is a large body of research in psychology on how people actually come by their beliefs, how they will respond to evidence uncongenial to what they are invested in believing, and what patterns of inquiry we can expect from people given their motivations. These findings come from the science of motivated cognition. But none of these findings directly address the question of what we should believe. This part of cognitive science does not provide a solution to that problem – and it doesn't even try to. It isn't even clear how it could, since simply describing what the human mind actually does under various circumstances doesn't tell us what counts as an error. When such descriptions strongly suggest to us that we make errors in motivated cognition, we are drawing on prior assumptions about what count as an error. An important role for philosophy is to make those assumptions explicit in order to examine them.

Could the example I used to illustrate this problem have been as unrealistic as the example of Mary is, and still illustrate the problem? I doubt it. But that's not because it is a piece of fiction. A more elaborated fictional narrative could illustrate the problem even more directly than our pair of young people do. It could do this by fixing on the psychic disharmonies that could beset someone no matter what they end up believing. By describing these disharmonies, a narrative about them could bring out the contours of the problem vividly, showing how difficult all the options were, guiding us through a character's challenges. The cognitive disharmonies would verifying that there truly is a problem. And a description of them be a way of analyzing it, showing all its imperfect paths forward, rather than just illustrating a situation that gets the problem started. A narrative like that would be a properly meaty thing, making philosophical analysis look like a skeleton. Skeletons tell you about the shape of things, but leave a lot to be filled in.

For instance, I have a friend who recently tried to quit smoking, after smoking for over half his of young life. This young man loves everything about smoking. He loves the way it divides up the day, the mini-vacation from socializing it allows ('be back in a minute, just going for a smoke'), the slow intake of heat, the chance to control a tiny bit of fire with his breath, the orange glow at the tip that twinkles like a star in his private universe. He loves the buzz that focuses his mind, clarifies his thoughts, and seems to enhance his memory. He loves smoking, even with all its health risks, which he knows well. The risks worry him. His complexion is sometimes green. Sometimes it's hard for him to breathe. His immune system is dicey and he often gets sick.

Given his love of cigarettes, it took a lot of effort for my friend to decide to quit. But he gathered his resolve and pitted it against his love of smoking. His resolve was wobbly but stable enough for him to face the fact that quitting is so hard that it rarely works. Hardly anyone can quit the first time, at least without drug enhancements. If you try to quit smoking, chances are high that you'll fail, at least at first.

Trying to do something you know has small chance of succeeding lands you smack in the middle of the problem of believing (and acting) against the evidence. This problem turns out to be the purview of a type of health professional: the *smoking-cessation specialist*. This person's job is to navigate the problem we've been discussing. They have to tell people to expect to fail in their attempts, while encouraging them to try.

A narrative focused on my friend or someone like him could trace the contours of this problem in a more vivid different way than a philosopher would. My friend's love of smoking is irrelevant to the problem, but a narrative that mentions it might endear you to him, the way the corresponding true facts about my friend endear him to me. His encounter with the cessation specialist was absurd – just what you'd expect from a philosophical problem in the flesh. My anecdote happened to be true, but you could write a story about someone in his situation, and it wouldn't matter whether the character was real or not. For instance, in George Orwell's novel *1984*, Winston resists believing the party dogma that 2+2=5 and as a result is tortured, but his neighbor lets himself become convinced, initially out of fear, and then never stops believing it.

When you look at realistic but fictional situations like these, you can see a problem in action. Unlike the case of Jackson's Mary, these examples illustrate the problem all on their own. And whereas Jackson's example of Mary could only play its stunning philosophical role by locating Mary in a highly unrealistic situation, examples that illustrate problems do so most forcefully when they locate their characters in situations we can easily picture ourselves being in.

But what exactly does it mean for an example to be realistic? It's clear enough what makes Jackson's Mary unrealistic. My friend's situation was real, so my example about him is realistic, and so are the two young people choosing career paths. But there are different dimensions of realism. We'll see next that these different kinds of realism can make a difference to how useful fictional examples are for testing normative hypotheses, which is a third role that fictional examples can play in philosophical inquiry.

Role 3: Test a normative hypothesis

Here's a hypothesis. In perception, you always have good reason to believe that things are the way they appear, unless you have special reason to think you're being misled. Most of the time, we don't have any such reason. If you want to know whether there's any mustard in the refrigerator, just open the door and look. In countless everyday situations, we rely on perception to find out mundane things.

We could call this hypothesis the of-course, go-ahead-and-believe-your-eyes hypothesis, but for short let's call it the simple hypothesis. The simple hypothesis is normative because it concerns pressures on what you should believe.

Is the simple hypothesis called into question by the phenomenon known as “cognitive penetration”, in cases where your experiences are unduly influenced by your prior unjustified beliefs, fears, suspicions, or hopes?

The answer might depend on what kind of example you choose. Let’s contrast two cases of cognitive penetration: one extremely unrealistic and abstract, and the other historically situated. The unrealistic one doesn’t do much to challenge the simple hypothesis, but the historically situated one arguably calls it into question, creating the need to refine it.

If two examples interact with a normative hypothesis in this way, which one should be given more weight? Or are they both useless?

Let’s return to this question after looking at the examples, starting with the unrealistic one.

Consider someone who thinks they’re seeing a red dot. It doesn’t matter who the person is, or when or where she lives, because anyone living in any time or place could see a red dot (...anyone except for Jackson’s Mary, in her colorless room!). As it happens, this person is hallucinating a red dot, and unbeknownst to her, her hallucination is happening because she has been wanting to see a red dot. The hallucination is brought about by her desire. You might call it “wishful seeing.”

According to the simple hypothesis, this hallucinator has reason to believe there’s a red dot in front of her. That is after all the way things look to her.

By contrast, if you think wishful seeing can remove the power of permeated experiences to support believing your eyes, then you’ll think this hallucinator does not have good reason to believe her eyes.

Which side is right? Here’s an extension of the example that might seem to favor the simple hypothesis. Suppose that without any change that the hallucinator can detect from the inside, she goes from hallucinating a red dot to actually seeing one. She is still pleased to be seeing a red dot, but her desire no longer plays any role in bringing about her visual experience. For all she knows, she has been seeing the same red dot the whole time.

Ho-hum. If you think that when (and because) this person’s hallucination is cognitively penetrated, she gets *less* reason from her experience to believe her eyes, then you're saying that she gains *more* reason to believe her eyes at the end of this seamless transition than she has at the start. The epistemic power of her experience changes, even though it seems the same to her all along.

That result might strike you as arbitrary score-keeping in epistemology.

But now see if the whole issue looks different when we consider a type of experience embedded in much more specific and real-world scenario.

The somewhat sterile, socially abstracted seamless transitions from hallucinating to seeing red dots probably has no actual instances, whereas this next type of scenario is a brutal and yet culturally normal one that recurs in the history of United States. It has all too many instances. It's a scenario in which someone - almost always a man, usually white, often armed and often a police officer - is acquitted for using force - often lethal force - against someone else who is black (usually a man or a boy), on the grounds that the shooter's belief that that man or boy posed imminent severe danger was reasonable. The acquittals lead to massive indignation. Cases like these reflect an ongoing political dynamic in the United States. Here are three examples of it from the last fifty years, each one sparking large and sometimes protracted political protests.

Harlem, New York City, 1964: Officer Thomas Gilligan shot and killed 15-year-old James Powell in Yorkville. They claimed he had a knife, but no knife was ever found.

Queens, New York City, 1973: Officer Thomas Shea (first NYC police officer tried for murder while on duty) was acquitted for shooting to death 10 year-old Clifford Glover. He claimed the 4th grader was reaching for a gun, but no gun was ever found.

Ferguson, Missouri, 2014: Police officer Darren Wilson told a grand jury that it was reasonable for him to shoot his gun at 18-year old Michael Brown, describing Brown as having "the most intense aggressive face. The only way I can describe it, it looks like a demon, that's how angry he looked."

During this time other similar cases never became part of public political life but showed the same pattern of acquittal and indignance.

Now that I've described the type of scenario, let's consider what the officers' perceptual experiences could have been like. Suppose these experiences result from cognitive penetration and fail to be a perception of things as they really are - like the red-sphere hallucination. We're trying out the hypothesis that the visual experiences of the men inflicting violence were cognitively penetrated by their own attitudes which are congruent with racism.

It isn't possible to know the exact contents of the visual experiences that these officers had. And we can't know whether any cognitive penetration of visual experiences actually occurred in these cases. But we can ask a hypothetical question that bears on the simple hypothesis. If any of the officers perceptually experienced threat or danger due to cognitive penetration by racist attitudes, would those perceivers have just as much reason to believe their eyes, as they could have if their perceptual experiences weren't influenced by racist attitudes?

The simple hypothesis would treat such experiences the same way they treat the case of seeing the mustard in the fridge. To the police officers, under the scenario we're considering, the people they're attacking look dangerous, and they feel as sure they are not being misled as you do when you look in the fridge for the mustard.

I think these cases speak strongly against the simple hypothesis, creating a need to refine it.

If someone hallucinated a threatening situation when they saw you, and if the situation they hallucinated jibed with an entrenched cultural stereotype that imposed limitations on you, you'd be justifiably offended. "But it's just an unwitting hallucination – like seeming to see a red dot", someone might say. This would be a poor defense, because in the context we're considering, it isn't just a hallucination. It's an experience that manifests a cultural situation, with a political dynamic operating through it. A cultural myth is operating through individual mind.

When I focus on the cultural myth operating in individual minds, it doesn't seem incumbent on someone to whom the fear was directed to excuse the officers upon learning that their racist outlook had reached all the way to their perceptual experiences. If a racist hallucinates me as dangerous, I have several reactions. I'm terrified (especially when this person is armed), angry, offended, and I'm more inclined to think the hallucinator has an epistemic problem than I am to think they're being reasonable because their twisted outlook has infiltrated all the way to their perceptual experience. It redounds poorly on him if the hallucinator can't see an ordinary young person for what he is, whether he is shoplifting or just going about his business.

Here's where I think the type of realism in examples can matter. When asking about what's rational for our fellow human beings to believe and to do, it makes sense to consider cases that are close to the ones in which these normative notions are ultimately meant to apply. Social relationships are deeply relevant to the justification of some beliefs. If we're asking what would be a reasonable state of mind for someone to be in, our verdicts and predictions about that need to stand up to human situations in their full complexity.

So it seems important to test hypotheses about what's reasonable against perceptual situations that don't abstract from the kinds of complexities missing from cases like the red sphere. Those complexities might rightly be missing when we test for visual acuity or when we ask how most generally the mental relates to the physical. But it's less obvious that examples lacking in social complexity should be our paradigms when we're asking about justificatory power and rational standing. The features in perception we abstract away from might be relevant to what we're trying to discover. If that idea is correct, then when considering problems about what we have reason to believe, examples that are realistic should carry more weight than examples that are abstracted from any social context.

But what makes a fictional example realistic? We can distinguish at least two kinds of realism. Sociological realism concerns whether a scenario reflects the social and political dynamics at a given time and place. A scenario is sociologically realistic only relative to a historical situation. Psychological realism concerns whether it is psychologically possible for someone's experiences or lifetime of experiences to be caused in the way an example describes.

Jackson's Mary is neither psychologically nor sociologically realistic. The same holds for seamless transitions from hallucinating to properly perceiving a red dot.

The idea that perceptions of danger or threat (including misperceptions) can be brought about by fears or beliefs that are congruent with racism is sociologically realistic. For all we know, it occurs all the time. Whether cognitive penetration is psychologically realistic, by contrast, is not known. It's known that perceptual experiences can be influenced by long-term changes in perceptual expertise, which is the ability to discriminate between different stimuli. You gain perceptual expertise when you learn to play chess, when you taste many subtle differences between wines, or when you get used to differentiating people with similar facial features to one another. These forms of expertise have lasting effects on your perceptual system. It's less clear in what ways perceptual experiences can be influenced by one-off states of mind such as emotions or beliefs that are congruent with racism. On that question, there's a lot of controversy in psychology. So we don't know right now which kinds of cognitive penetration are the most psychologically realistic. These types of influence on perception have long been subject to debate in perception science.

If a field of inquiry only ever posed loaded questions or analyzed problems, you might think it was perpetually in preparation, pursuing questions but never purporting actually to answer them. While some types of inquiry in the philosophy of mind use fictional examples as springboards for inquiry, in normative contexts such examples can also move us from questions to answers. And they can do the same for claims about what's necessary. Even if it turns out that racist attitudes never cognitively penetrate visual experience, we might still learn from reflecting on examples that it is possible.

This observation highlights another role for fictional examples, this time one that does not lie within any clear disciplinary boundaries. Some fictions we think up might turn out to be psychologically possible. In his children's story "The Golden Key," the 19th-century Scottish writer George MacDonald describes a rainbow containing hues beyond violet – colors no human had ever seen. Over a century later, psychologists investigate whether there can be novel colors. When fiction or philosophy generate hypotheses about the mind that can be tested, they shade into theoretical psychology.