Frank Jackson, Latter Day Physicalist

James Garvey interviews Frank Jackson, originator of one of philosophy"s most brilliant thought experiments. (Originally published in 2011.)

Here is one of the best thought experiments in the whole of the philosophy of mind:

“Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room *via* a black and white television monitor. She specialises in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes.... What will happen when Mary is released from her black and white room or is given a colour television monitor? Will she *learn* anything or not?”

Well, what do you think? Take your time, because there’s a lot at stake: nothing less than the fundamental metaphysical nature of the universe itself. And don’t worry if you’re not sure what to say, because apparently there’s a lot to be said. There are more than a thousand published papers, innumerable conferences, and even several books addressing the question of what Mary did or didn’t know.

It’s Frank Jackson’s knowledge argument, and it appeared in 1982 in a paper with the agreeably strange title, “Epiphenomenal Qualia”. Qualia are the potentially spooky features of some conscious states, the so-called raw feels of our experiences – the pangs of jealousy, the hurtfulness of pain, the redness of red, the tang of the taste of a lemon, and so on. Epiphenomalism is the view that at least some mental properties have no physical effects. As Thomas Huxley vividly put it, such properties don’t do anything in the physical world, just as “the steam whistle which accompanies the work of a locomotive engine is without influence upon its machinery”. So just knowing the paper’s title, you know Jackson is talking about a serious sort of dualism, the view that there’s more stuff in the world than just physical objects. And some of that stuff has no effects in the physical world. As Jackson put it in the article, it’s hard to buy into the view “without sounding like someone who believes in fairies”.

Nevertheless, it’s had an enormous impact. A bit of unscientific Googling turns up 2.5 million pages for “Frank Jackson’s Knowledge Argument”. Compare that, entirely unfairly, to “Immanuel Kant’s Transcendental Deduction”, which barely limps past half a million. Perhaps his argument’s power lies in the fact that it just grabs you by the collar and forces a choice on you. Did Mary learn something or not? That translates roughly to, well, pick one – dualism or physicalism?

If you think that Mary knew all the physical facts but learns something when she first sees red, then there’s more to know than just physical facts – by hypothesis, she had all those already. So if she learns something, physicalism is false, because it leaves out part of the world she discovers on experiencing red. Maybe it leaves out epiphenomenal qualia.

You can try to deny the existence of nonphysical properties, keep your physicalist credentials, and say she would have somehow already known all about red in her black and white room, despite having never seen it. If you're a physicalist, knowing all the physical facts just is knowing everything there is to know, so she would learn nothing the first time she sees red. But that's a stretch, isn't it? As Jackson concludes in the original article, it "seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. *Ergo* there is more to have than that, and Physicalism is false."

Remarkably, Jackson has since somehow talked himself out of it all. He now resolutely rejects dualism. I wonder how hard that must have been – getting international attention with a pretty impressive argument in favour of a minority view, fighting your corner with gusto, only to swap sides comprehensively some years later. But we start at the beginning, and I ask him to tell me the history of the argument. How did it come to him? Was it a Eureka moment in a bathtub or did it take ages to puzzle out? His reply starts with a generous proviso.

"Almost everything I'm going to say about that argument is based on hundreds, perhaps thousands of discussions with friends and colleagues. If this were an academic article it would be bristling with footnotes and acknowledgements. This is an interview, so I'm not going to give you lots of names.

"The knowledge argument or the Mary argument or the black and white room argument actually has a long history. In the original paper I wrote on it, I footnoted a bit of the history, and of course since then I've discovered its history was richer and longer. In fact there's a little version of it in C D Broad and lots of other places."

Broad certainly was thinking nearby, but instead of a brilliant neuroscientist, he goes on about mucous membranes and offers us the slightly uninspiring image of an archangel with a grip on chemistry and "the further power of perceiving the microscopic structure of atoms". The creature would "know exactly what the microscopic structure of ammonia must be; but he would be totally unable to predict that a substance with this structure must smell as ammonia does when it gets into the human nose." Not bad, but not Mary either.

"I would take credit for putting it forcefully and clearly," Jackson says, "but I would like to say that of course the argument's got a long history." Fair enough, but, personally, what happened?

"I had been a dualist for years. I was taught by Michael Bradley, and he had some good arguments for dualism. I always thought it was a plausible view. As I say in the beginning of 'Epiphenomenal Qualia', we dualists don't really need an argument to say that consciousness doesn't fit into the physicalist world view. It's just intuitively obvious. When you hunt for arguments you hunt for arguments that physicalists are going to have trouble resisting. I got a telephone call – this was before the days of email – from the psychology department at Monash University, asking me to give a lunchtime talk. They didn't quite say it this way, but they sort of said we understand that you're one of the few dualists left on the planet. Would you like to give a talk saying why? So I had to write something, something

informal. And that was the first draft of 'Epiphenomenal Qualia'. It seemed a bit of a waste not to do something with it, so I added in some stuff and made it longer. I wrote it reasonably quickly." There's a pause. He puts on a wistful face and stares off into the middle distance. "I was younger then." He half-inhales an infectious, staccato laugh.

"The follow up article ('What Mary Didn't Know') came about after Paul Churchland wrote a not terribly friendly piece about the knowledge argument. I thought it was a bit offhand. I didn't worry about him saying he didn't believe it, that's fine, but he sort of suggested it was making some kind of elementary error which anyone could pick up. Not quite as bad as affirming the consequent but pretty bad all the same. That riled me slightly, and I regret to say the slight tone of irritation shows in the piece." He actually says that with a slight tone of irritation. He looks a little riled now.

When I have another look at the papers I see what he means. It was never going to be particularly convivial. Who could possibly have less sympathy for dualism than Churchland? His view, eliminative materialism, has it that that our psychological categories might be eliminated by a mature neuroscience – beliefs, hopes, desires and so on might not map on to an empirically informed theory of the brain's functions, so we might end up having to revise, even eliminate our everyday view about the mind. Like witchcraft and phlogiston, beliefs and desires might end up consigned to the conceptual scrapheap, once we get a grip on how the central nervous system really works. It's about as materialist as materialism gets.

Churchland presents "a conveniently tightened version" of the knowledge argument, which in itself must have been a little exasperating for Jackson. (What? It wasn't tight enough the first time around?) In one of Churchland's reconstructed premises Mary knows about brains states, but in another she doesn't know about sensations. Churchland argues that "the defect ... is simplicity itself". Jackson is equivocating, using "knows about" in two different ways, talking about two different kinds of knowledge, and this renders the argument invalid. Once you spot this, Churchland beams, the argument is "a clear non sequitur .... Such arguments show nothing". God, he even has a bit of fun with a parallel argument about ectoplasm. It doesn't quite call for pistols at dawn, but I can see how Churchland might be read as being dismissive of the misguided little dualist. Maybe Jackson did well to be merely riled.

In Jackson's reply, he says with an audible huff that Churchland's reformulation of the argument "may be convenient, but it is not accurate". It's not the *kind of knowledge* Mary has but *what she knows* that matters. He produces "a convenient and accurate" version of the argument which appears to sidestep Churchland's objection.

"That's the biographical background to it," he continues. "Now, exactly why that particular version of the knowledge argument popped into my head – I do not know," he says, genuinely mystified. Maybe he read Broad's short argument many years earlier, and although he forgot about it, it might have exerted some unconscious influence. But he certainly had seen Thomas Nagel's 1974 essay, "What is it like to be a bat?", and maybe that did figure in somehow. There, Nagel writes about batty subjectivity – what it's like to be a bat and experience a

sonar image of the world – which he argues is only accessible to bats. He concludes that "it is a mystery how the true character of experiences could be revealed in the physical operation of that organism." The conclusion is importantly different to Jackson's: it's not that physicalism is false, but that we can't understand what it might mean to say that it's true. Jackson says something of Nagel might have been on his mind, maybe he was trying to make a similar point without all of the complexity of Nagel's piece.

Whatever the argument's origins, it's had an extraordinary history over the past 30 years – objections, replies, countless reformulations on behalf of well-wishers and hostile interpreters alike. As Jackson has authorial privilege, I ask him how he understands the argument. What's his interpretation of it? What's its real point?

"Although I now think it's mistaken," he begins, "the essential thought behind the argument is simply that when Mary has colour experiences, her conception of the kinds of properties that are instantiated in our world gets dramatically expanded. In theory it's no different than coming across a new sort of animal. How many different sorts of dogs are there? People think they've gotten on top of it, but they turn the corner, and the see a completely different dog from any dog they've got on their inventory. So they enlarge their conception of how many kinds of dogs there are. What happens to Mary is that she has a certain view of what the world's like, a black and white view, and all the stuff that comes to her from the physical sciences. And when she sees colour for the first time I think the plausible thing to say is that she gets an enlarged idea of what kinds of properties there are to be encountered in the world. She comes across new properties."

When Jackson lays it out like that, crystal clear, it's hard not to feel a certain insecurity about physicalism. What else can you say, except that Mary learns about a new part of the world when she sees colour for the first time? But Jackson is a latter day physicalist. How did he talk himself out of dualism?

"I've always thought that if you're a dualist you should be up front about the metaphysics. And you should say, of course these properties are epiphenomenal. We know enough about the world to know that these extra properties which I believe in aren't guiding my pen as I write the article saying qualia are left out of our physical picture of the world. In 'Epiphenomenal Qualia' I explain why it's not such a disaster being an epiphenomenalist, but I came to think of this as a triumph of philosophical ingenuity over common sense. This is what someone who's done a good philosophy degree can somehow make seem all right, but if you look at it in a more commonsensical way it's actually pretty implausible. So the epiphenomenal stuff was just very hard to believe.

"For a while I was at the stage of people who say, there must be something wrong with the knowledge argument. It's not obvious, despite the fact that some people jump up and down and say it's obvious, because look at all these smart people giving quite different diagnoses of what's wrong. That tells you it's not obvious what's wrong with it. I was in that situation, thinking there's got to be something wrong with it but not sure what it was. And then I decided that the best way out is to think in representationalist terms about phenomenal experience. When you think in those terms, what you're thinking is that when

something looks red to you, don't think of that as a relationship between you and an instance of some special property. Think of it as representing things as being a certain way. You don't think of it in relational terms, you think in propositional terms, as a kind of intentional state.

"When you think in those terms, it's a mistake to wonder where the special redness is. What you have to ask yourself is, when something looks red, how am I representing the world to be? And if you're convinced that you're representing the world such that it has some special property outside the physical picture of the world, and you think physicalism is plausible, then of course you think it's a case of false representation. Then you better have some story about how looking red represents things to be, and what that to be is, and how it can be found in a physical picture of what the world's like."

That actually helps break the spell a little. Maybe it's a mistake to think of Mary as bumping into a new thing, like turning the corner and seeing a new kind of dog. Instead, she's got a representation of how things are. But how do representations work, on this view? What is it to see red if it's not to be in a relation with something red?

"When I'm talking about representation I'm talking about a state where you're invited to have a certain view about how things are. Of course you may reject it. When you have those famous perceptual illusions, and you know they're illusions, you're in a state which invites you to think that some line is curved. You know perfectly well it's not curved. Nevertheless you're in a state which sort of says to you, 'This is the way things are! This is the way things are!' That's what I mean by a representation. So when something looks red, I think you're in a state which almost shouts at you, 'This object has a really striking surface property!' The experience of something's looking red doesn't say something about you. It says something about the object. Dispositional theories are theories that say in one way or another that we should think of colour as a relation between you and the object. But I think that when something looks red, you're representing the way it is, not the way you are."

I'm still not sure I see exactly how a shift to representationalism gets us clear of trouble with Mary. I ask Jackson for his new, physicalist answer to the question posed by his former dualist self: does Mary learn something or not? He takes a deep breath – I get the feeling he's spent more time thinking about this question than just about anyone else. It turns out that the physicalist has not one, but two ways out.

"Looking red I think is clearly a representational state. I think the idea that perceptual states in general are representational states is extremely plausible. If you think that and you're a physicalist what you have to say is, right, Mary clearly enters a new representational state when she leaves the room. That should be common ground. If you're a physicalist, then you've got two things to say. You're either going to say, why doesn't she get new knowledge? Well, she already had it. If she already had it then you have to answer the question, what property do her newer experiences represent things as having which she knew about in the room? Maybe she didn't know about it under the name 'red', but if she's in a new representational state, and things are as they're being represented to be, and she doesn't learn anything new about the world, you need to give an answer to

what looking red represents things as being, where the content of the representation can be expressed in physical terms. Alternatively, you can say it's a false representation. Colour is an illusion. You have to say one or the other."

This seems to be about as far as Jackson cares to go outside the black and white room. Once there's an escape route, he seems satisfied to leave it at that. I ask him which course he takes. If it's a new representation, how does he understand it? If seeing the red of the rose is an illusion, what's illusory about it?

"On Mondays, Wednesdays and Fridays I go for the illusion view," he laughs, "the other days I say, what she's representing is certain complex similarity and difference relationships between the light accessible properties of objects. Now she doesn't know what properties stand in those similarity relations, it's up to optical science to tell us what they are, but when something looks red, it's represented as being strikingly similar to blood and strikingly different from the sky, as being more similar to pink things than to black things, as having a property of grabbing your attention in a distinctive way, a way in which dark blue does not, etc, etc, etc.

"But if we do the physics we may not find that the properties of the surfaces stand in these similarity relations. Well, in that case I'd be an eliminativist. I think we'd have to say, right, colour is an illusion, a very useful illusion, but an illusion all the same." So for Jackson it really could still go either way – "but that's mostly a matter for physics, not philosophy."

That thought about science brings us neatly to another point against physicalism made by Jackson in his dualist days. Physicalism is an extraordinarily optimistic view of our mental capacities – in principle, we've pretty much got a grip on all that there is, the physical stuff that makes up our world, and we're on our way to understanding it. But if our understanding is shaped by the need to survive – our brain is an evolved thing, after all – isn't it likely that there are vast parts of the universe that we'll never get a grip on, just because it never mattered in our evolutionary history? Doesn't this suggest that physicalism almost certainly leaves some of the universe out? Maybe the mental side of us or some part of it?

Dualist Jackson once made the point by imagining sea slugs burbling around in our deepest oceans – perhaps they evolved rationality and developed sciences, suitably restricted compared to ours, given their limited environment, but sciences that work pretty well for them where they are. They have philosophers too: tough-minded slugists who say that the restricted terms of their science can explain everything there is, and soft-minded slugists who suspect there may be some mysterious residue left out by slug science. With a richer grasp of the world and a larger science, we can see where the tough-minded slugists go wrong. But of course a being with a more comprehensive grip on things might make human physicalists look just like slugists. We could be making the tough-minded slugists' mistake. Maybe some part of the mind lies beyond the reach of physicalism, just as parts of the world are beyond the slug's view. Does he still have some sympathy with the humility of his earlier reflections, despite his conversion to physicalism?

"Yes I do. There's a position I call Kantian physicalism. What it says is this. Isn't it common sense that there are things that we don't know about the world? Even

the most enthusiastic physicalist has to say there are gaps in our knowledge. It's at least plausible that it goes much beyond that – it's not just that there are problems in quantum mechanics. It might be that there's a whole range of properties that we don't and can't know about because they don't impinge on us. Or if they do they impinge on us in a way that has no relevance to survival, so we didn't evolve in such a way that we can pick them up. Isn't that right?

"What the Kantian physicalist says is, yes, that is right. But those properties don't matter for mentality. In other words, if you took a world just like this, duplicated it in all the physical respects, but changed its fundamental nature in all sorts of dramatic ways, the pains would hurt just as much. You'd be screaming just as loud, you'd be pleading for the surgeon to stop operating without anaesthetic. So the mental side of things, the phenomenal side of things would be unaltered.

"There's an interesting paper by David Lewis called 'Ramseyan Humility'. You would think of Lewis as being the paradigmatic physicalist, and certainly in his earlier writings that's what you get. But in this paper he suggests there might be a whole range of properties we can't know about, because permuting them doesn't make any difference at the level in which we interact with the world. It's a bit like that thought experiment: maybe there's a matter version of our world, and an antimatter version, and there are duplicates of you and me, but one's made of matter and the other's made of antimatter. You can't know whether we're in the matter world or the antimatter one. There's a whole range of things you can't know. Imagine these duplication cases, where you duplicate the world physically, but change all the rest of the stuff – our conversation would go exactly the same way. Quine still writes 'Two Dogmas of Empiricism', or maybe it's Quine2 in the duplicate world, but the words on the page are exactly the same. The same number of people agrees with it, the same number of people disagrees with it.

"The slugists were wrong. They thought that they knew more than they do in fact know. But as far as mentality goes, the physicalist can say that the physical story is enough for mentality."

I take the point that a physicalist can be humble, but I'm still left with doubts about Mary. In the end, somehow, I don't entirely buy Jackson's new reply to that old question: does she learn anything or not? I'm still back where he was some years ago – I've got the feeling something's wrong with the argument, but I don't know what it is.

Maybe the long representationalist story is right, but somehow it just doesn't quite fit the simple, elegant question raised by Mary seeing red for the first time. The part of me in favour of parsimony would very much like a simple answer to a simple question. It's ingenious, all that talk about representing complex similarity and difference relationships between the light accessible properties of objects. I wouldn't go so far to say it seems ad hoc, but it does feel a little contorted, an unnatural stance taken up to squeeze out of the tangle of the knowledge argument. And maybe that long story works with seeing red – I think Jackson is right to say that perceptual states are essentially representational – but I'm left wondering about other states with qualitative feels that don't obviously represent anything. What is this pang of regret supposed to represent? My stupid decision to study philosophy when I could have been a well-heeled lawyer instead?

Jackson might have talked himself out of the knowledge argument's conclusion, but I still don't know. I'm no dualist, but there's something about Mary.