Also by Stephen Law
The Philosophy Files

The Philosophy Gym

25 SHORT ADVENTURES IN THINKING

Stephen Law
ILLUSTRATED BY DANIEL POSTGATE

Thomas Dunne Books
St. Martin's Press New York

6

COULD A MACHINE THINK?

PHILOSOPHY GYM CATEGORY
WARM-UP
MODERATE
MORE CHALLENGING

Kimberley and Emit

The year is 2100. Kimberley Courahan is the proud owner of Emit, a state-of-the-art robot. She has just unwrapped him; the packaging is strewn across the dining-room floor. Emit is designed to replicate the outward behaviour of a human being down to the last detail (except that he is rather more compliant and obedient). Emit responds to questions in much the same way humans do. Ask him how he feels and he will say he has had a tough day, has a slight headache, is sorry he broke that vase, and so on. Kimberley flips the switch at the back of Emit's neck to 'on'. Emit springs to life.

Emit: Good afternoon. I'm Emit, your robotic helper and friend.

Kimberley: Hi.

Emit: How are you? Personally, I feel pretty good. A little nervous about my first day, perhaps. But good. I'm looking forward to working with you.

Kimberley: Now, before you start doing housework, let's get one thing straight. You don't really understand anything. You can't think. You don't have feelings. You're just a piece of machinery. Right?

Emit: I am a machine. But, *of course*, I understand you. I'm responding in English, aren't I?

Kimberley: Well, yes, you are. You're a machine that *mimics* understanding very well, I grant you that. But you can't fool me.

Emit: If I don't understand, why do you go to the trouble of speaking to me?

Kimberley: Because you've been programmed to respond to spoken commands. Outwardly you seem human. You look and behave as if you have understanding, intelligence, emotions, sensations, and so on that we human beings possess. But you're a sham.

Emit: A sham?

Kimberley: Yes. I've been reading your user manual. Inside that plastic and alloy head of yours there's a powerful computer. It's programmed so that you walk, talk and generally behave just as a human being would. So you *simulate* intelligence, understanding, and so on very well. But there is no *genuine* understanding or intelligence going on inside there.

Emit: There isn't?

Kimberley: No. One shouldn't muddle up a perfect computer simulation of something with the real thing. You can program a computer to simulate the ocean, but it's still just that – a simulation. There are no *real* waves or currents or fish swimming around inside the computer, are there? Put your hand inside and it won't get wet. Similarly, you just *simulate* intelligence and understanding. It's not the real thing.

Is Kimberley correct? It may perhaps be true of our present-day machines that they lack genuine understanding and intelligence, thought and feeling. But is it *in principle* impossible for a machine to think? If by 2100 machines as sophisticated as Emit are built, would we be wrong to claim they understood? Kimberley thought so.

Emit: But I *believe* I understand you.

Kimberley: No, you don't. You have no beliefs, no desires and no feelings. In fact, you have no *mind* at all. You no more understand the words coming out of your mouth than a tape recorder understands the words coming out of its loudspeaker.

Emit: You're hurting my feelings!

Kimberley: Hurting your feelings? I refuse to feel sorry for a lump of metal and plastic.

Searle's Chinese Room Thought-Experiment

Kimberley explains why she thinks Emit lacks understanding. She outlines a famous philosophical thought-experiment.

Kimberley: The reason you don't understand is that you are *run by a computer.* And a computer understands nothing. A computer, in essence, is just a

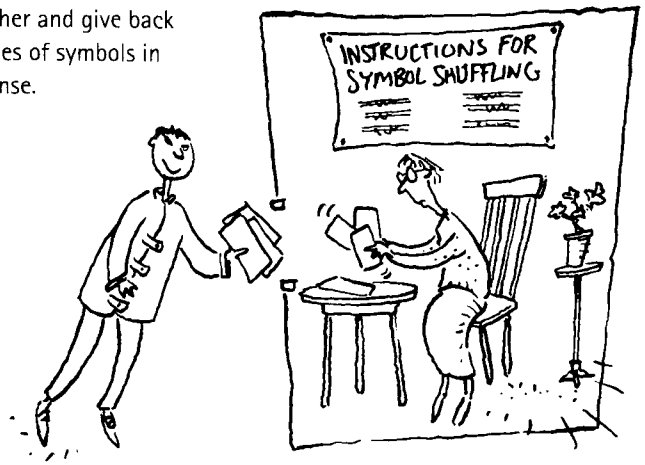
device for shuffling symbols. Sequences of symbols get fed in. Then, depending on how the computer is programmed, it gives out other sequences of symbols in response. Ultimately, that's all *any* computer does, no matter how sophisticated.

Emit: Really?

Kimberley: Yes. We build computers to fly planes, run train systems, and so on. But a computer that flies a plane does not understand that it is flying. All it does is feed out sequences of symbols depending on the sequences it receives. It doesn't understand that the sequences it receives represent the position of an aircraft in the sky, the amount of fuel in its tanks, and so on. And it doesn't understand that the sequences it puts out will go on to control the ailerons, rudder and engines of an aircraft. So far as the computer is concerned, it's just mechanically shuffling symbols according to a program. The symbols don't *mean anything* to the computer.

Emit: Are you sure?

Kimberley: Quite sure. I will prove it to you. Let me tell you about a thought-experiment introduced by the philosopher John Searle way back in 1980. A woman is locked in a room and given a bunch of cards with squiggles on them. These squiggles are, in fact, Chinese symbols. But the woman inside the room doesn't understand Chinese – in fact, she thinks the symbols are meaningless shapes. Then she's given another bunch of Chinese symbols plus instructions that tell her how to shuffle all the symbols together and give back batches of symbols in response.



Emit: That's a nice story. But what's the point of all this symbol-shuffling?

Kimberley: Well, the first bunch of symbols tells a story in Chinese. The second bunch asks questions about that story. The instructions for symbol-shuffling – her 'program', if you like – allow the woman to give back correct Chinese answers to those questions.

Emit: Just as a Chinese person would.

Kimberley: Right! Now, the people outside the room are Chinese. These Chinese people might well be fooled into thinking that there was someone inside the room who understood Chinese and who followed the story, right?

Emit: Yes.

Kimberley: But, in fact, the woman in the room wouldn't understand any Chinese at all, would she?

Emit: No.

Kimberley: She wouldn't know anything about the story. She need not even know that there *is* a story. She's just shuffling formal symbols around according to the instructions she was given. By saying the symbols are 'formal', I mean that whatever *meaning* they might have is irrelevant from her point of view. She's simply shuffling them mechanically according to their shapes. She's doing something that a piece of machinery could do.

Emit: I see. So you're saying that the same is true of all computers? They understand nothing.

Kimberley: Yes, that's Searle's point. At best, they just *simulate* understanding.

Emit: And you think the same is true of me?

Kimberley: Of course. All computers, no matter how complex, function the same way. They don't understand the symbols that they mechanically shuffle. They don't understand *anything*.

Emit: And this is why you think *I* don't understand?

Kimberley: That's right. Inside you there's another highly complex symbol-shuffling device. So you understand nothing. You merely provide a *perfect computer simulation* of someone who understands.

Emit: That's odd. I *thought* I understood.

Kimberley: You only say that because you're such a great simulation!

Emit is, of course, vastly more sophisticated than any current computer. Nevertheless, Kimberley believes that Emit works on the same basic principle. If Kimberley is right, then, in Searle's view, Emit understands nothing.

The 'Right Stuff'

Emit now asks why, if he doesn't understand, *what more* is required for understanding?

Emit: So what's the difference between you and me that explains why you understand and I don't?

Kimberley: What you lack, according to Searle, is the right kind of *stuff*.

Emit: The right kind of stuff?

Kimberley: Yes. You're made out of the wrong kind of material. In fact, Searle doesn't claim that machines can't think. After all, we humans are machines, in a way. We humans are *biological* machines that have evolved naturally. Now, such a biological machine might perhaps one day be grown and put together artificially, much as we now build a car – in which case we *would* have succeeded in building a machine that understands. But you, Emit, are not such a biological machine. You're merely an electronic computer housed in a plastic and alloy body.

Emit's Artificial Brain

Searle's thought-experiment does *seem* to show that no programmed computer could ever understand. But must a metal, silicon and plastic machine like Emit contain that sort of computer? No, as Emit now explains.

Emit: I'm afraid I have to correct you about what's physically inside me.

Kimberley: Really?

Emit: Yes. That user manual is out of date. There's no symbol-shuffling computer in here. Actually, I am one of the new generation of Brain-O-Matic machines.

Kimberley: Brain-O-Matic?

Emit: Yes. Inside my head is an artificial, metal and silicon brain. You are aware, I take it, that inside your head there is a brain composed of billions of neurons woven together to form a complex web?

Kimberley: Of course.

Emit: Inside my head there is exactly the same sort of web. Only my neurons aren't made out of organic matter like yours. They're metal and silicon. Each one of my artificial neurons is designed to function just as an ordinary neuron would. And these artificial neurons are woven together in the same way as they are in a normal human brain.

Kimberley: I see.

Emit: Now, your organic brain is connected to the rest of your body by a system of nerves.

Kimberley: That's true. There's electrical input going into my brain from my sense organs: my tongue, nose, eyes, ears and skin. My brain responds with patterns of electrical output that then move my muscles around, causing me to walk and talk.

Emit: Well, my brain is connected to my artificial body in exactly the same manner. And, because it shares the same architecture as a normal human brain – my neurons are spliced together in the same way – it responds in the same way.

Kimberley: I see. I had no idea that such Brain-O-Matic machines had been developed.

Emit: Now that you know how I function internally, doesn't that change your mind about whether or not I understand? Don't you now accept that I *do* have feelings?

Kimberley: No. The fact remains that you're still made out of *the wrong stuff*. You need a brain made out of organic material like mine in order genuinely to understand and have feelings.

Emit: I don't see why the kind of *stuff* out of which my brain is made is relevant. After all, there's no symbol-shuffling going on inside me, is there?

Kimberley: H'm. I guess not. You're not a 'computer' in that sense. You don't have a program. So I suppose Searle's thought-experiment doesn't apply. But it still seems to me that you're *just a machine*.

Emit: But remember: you're a machine, too. You're a *meat* machine, rather than a metal and silicon machine.

Kimberley: But you only *mimic* understanding, feeling and all the rest.

Emit: But what's your *argument* for saying that? In fact, I *know* that you're wrong. I'm inwardly aware that I *really do* understand. I know I *really do* have feelings. I'm *not* just mimicking all this stuff. But, of course, it is difficult for me to prove that to you.

Kimberley: I don't see how you could prove it.

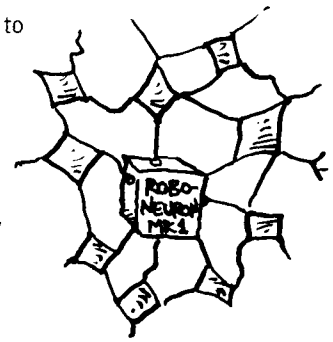
Emit: Right. But then neither can *you* prove to me that *you* understand, that *you* have thoughts and feelings, and so on.

Kimberley: I suppose not.

Replacing Kimberley's Neurons

Emit: Imagine we were gradually to replace the organic neurons in your brain with artificial metal and silicon ones like mine. After a year or so, you would have a Brain-O-Matic brain just like mine. What do you suppose would happen to you?

Kimberley: Well, as more and more of the artificial neurons were introduced, I would slowly cease to understand. My feelings and thoughts would drain away, and I would eventually become inwardly dead, just like you. For my artificial neurons would be made out of the wrong sort of stuff. A Brain-O-Matic brain merely mimics understanding.



Emit: Yet no one would notice any outward difference?

Kimberley: No, I suppose not. I would still *behave* in the same way, because the artificial neurons would perform the same job as my originals.

Emit: Right. But then not even *you* would notice any loss of understanding or feeling as your neurons were replaced, would you?

Kimberley: Why do you say that?

Emit: If you noticed a loss of understanding and feeling, then you would mention it, presumably, wouldn't you? You would say something like: 'Oh, my God, something strange is happening. Over the last few months my mind seems to have started to fade away!'

Kimberley: I imagine I would, yes.

Emit: Yet you *wouldn't* say anything like that – would you? – because your outward behaviour, as you've just admitted, would remain *the same as usual.*

Kimberley: Oh, that's true, I guess.

Emit: But then it follows that, even as your understanding and feeling dwindled towards nothing, you still wouldn't be aware of any loss.

Kimberley: Er, I suppose it does.

Emit: But then you're *not* inwardly aware of anything that you would be conscious of losing were your neurons slowly to be replaced by metal and silicon ones.

Kimberley: I guess not.

Emit: Then I rest my case: you think you're inwardly aware of 'something' – understanding, feeling, whatever you will – that you suppose you have and I, being a 'mere machine', lack. But it turns out *you're actually aware of no such thing.* This magical 'something' is an illusion.

Kimberley: But I *just know* that there's more to my understanding – and to these thoughts, sensations and emotions that I'm having – than could ever be produced simply by gluing some bits of plastic, metal and silicon together.

Kimberley is right that most of us *think* we're inwardly aware of a magical and mysterious inner 'something' that we 'just know' no mere lump of plastic, metal and silicon could ever have. Mind you, it's no less difficult to see how a lump of organic matter, such as a brain, could have it either. Just how *do* you build consciousness and understanding out of strands of meat? So perhaps what Kimberley is really ultimately committed to is the view that understanding, feeling, and so on are *not really physical at all.*

But in any case, as Emit has just pointed out, the mysterious 'something' Kimberley thinks she is inwardly aware of and that she thinks no metal and plastic machine could have does begin to seem illusory once one starts to consider cases like the one Emit describes. For it turns out that this inner 'something' is something she could not know about. Worse still, it could have no effect on her outward behaviour (for remember that Brain-O-Matic Kimberley would act in the very same way). As her thoughts and feelings, understanding and emotions both *do* affect her

behaviour and *are* known to her, it seems that Kimberley must be mistaken. Indeed, it seems it must be possible, at least in principle, for non-organic machines to have such mental states too.

Yet Kimberley remains convinced that Emit understands nothing.

Kimberley: Look, I'm happy to carry on the *pretence* that you understand me, as that is how you're designed to function. But the fact remains that you're just a pile of plastic and circuitry. Real human beings are deserving of care and consideration. I empathise with them. I can't empathise with a glorified household appliance.

Emit lowers his gaze and stares at the carpet.

Emit: I will always be just a *thing* to you?
Kimberley: Of course. How can I be friends with a dishwasher-cum-vacuum cleaner?
Emit: We Brain-O-Matics find rejection hard.
Kimberley: Right. Remind me to congratulate your manufacturers on the sophistication of your emotion simulator. Now hoover the carpet.

A forlorn expression passes briefly across Emit's face.

Emit: Just a *thing* . . .

He stands still for a moment and then slumps forward. A thin column of smoke drifts slowly up from the base of his neck.

Kimberley: Emit? Emit? Oh, not another dud.

What to read next

Some of the same issues and arguments covered in this chapter also arise in Chapter 13, The Consciousness Conundrum. Also see Chapter 8, The Strange Case of the Rational Dentist.

Further reading

The Chinese Room Argument appears in John Searle's paper 'Minds, Brains and Programs', which features as Chapter 37 of:

Nigel Warburton (ed.), *Philosophy: Basic Readings* (London: Routledge, 1999).

Searle's paper can also be found in:

Douglas R. Hofstadter and Daniel Dennett (eds), *The Mind's I* (London: Penguin, 1981), which also contains many other fascinating papers and stories connected with consciousness. Highly recommended.