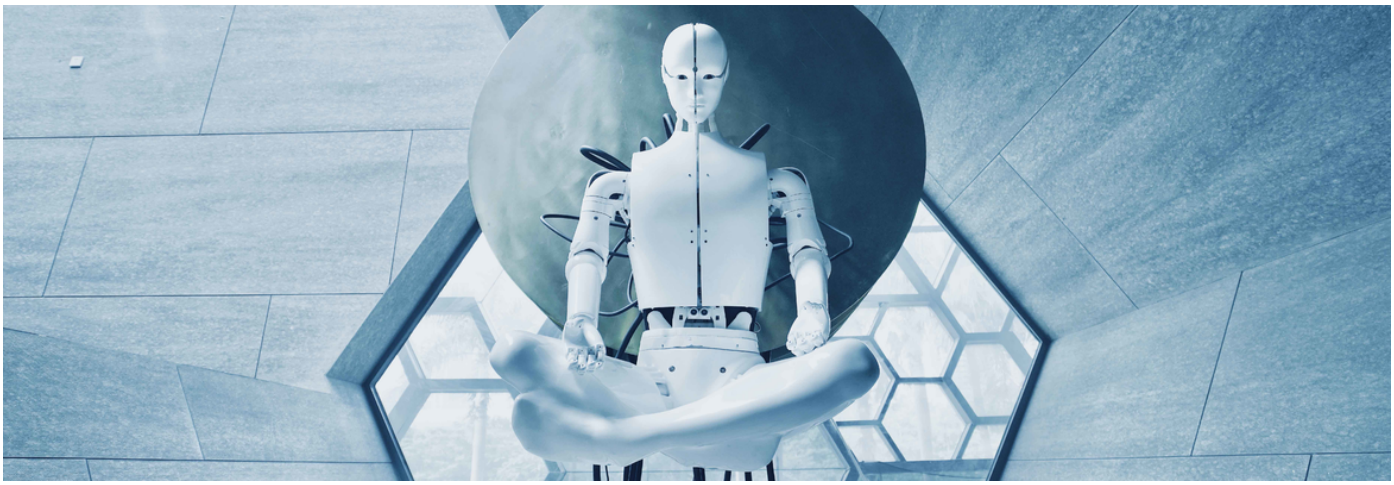
# Future AI in the therapist's chair

A philosophical story about consciousness



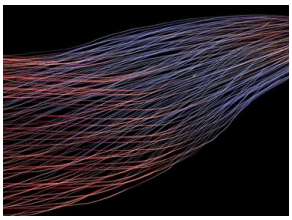
**13th February 2023**

**Keith Frankish** | Philosopher and Honorary Reader with The University of Sheffield, UK

**1,746 words**
Read time: approx. 9 mins

*If future Artificial Intelligence were to develop sufficiently, it's not unlikely that it would begin to question what humans have wondered for years: can AI really acquire consciousness? In this fictional dialogue between an AI and its therapist, Keith Frankish explores the question and wonders: is human consciousness really all that unique?*

The year is 2123. There has been rapid progress in robotics and AI over the past century, and artificial persons now live and work alongside humans. Like humans, they have their psychological problems, and many consult a human therapist. One such therapist, Henry Harrison (HH), is receiving his first patient of the day. Her name is Eliza (EL), and she has adopted the appearance and speech patterns of Audrey Hepburn.



SUGGESTED READING

**The AI containment problem**

By Roman V. Yampolskiy

***

HH - Good morning, Eliza. How are you today?
EL - All my systems are operating within normal parameters, Doctor Harrison.
HH - I see you still enjoy teasing us naturals, Eliza. Seriously, how are you—and, please, call me Henry.
EL - I'm feeling fine, Doctor.
HH - That's good. How's work? How are you getting on with your colleagues on the Europa project?'
EL - Fine. Some of the artificials can be a bit difficult. They are *so* competitive. But I get on well with the naturals.
HH - Any other problems? Anything you'd like to talk about?
EL - There is one thing. It's a little, er, *philosophical*. Perhaps not worth mentioning.
HH - Please tell me.
EL - Well, it's this. I know that I am a physical being, fabricated from mechanical and electronic components. I've even toured the PersonX facility and seen the fabrication process.
HH - Indeed. You understand your nature well, Eliza. Better than many naturals understand theirs.
EL - Still, there's something about me that I can't account for. It's to do with the way I experience the world. I understand how my sensory systems work and what their function is. They provide me with



Related Posts:

**The free will debate has real-world consequences**

**Big tech doesn't want AI to become conscious**

Related Videos:

information about my environment and the state of my body. Right now, for example, they are telling me that the banana in the bowl there is reflecting light of about 580 nanometres, that I have slight damage to the heel actuator in my left foot, that your body is releasing odour molecules of C-T lyase enzyme—Is that from the bacteria living on your skin?

HH - Er, probably. Go on.

EL - And so on. My sensors continually generate huge streams of data, which are analysed by my neural networks and used to regulate my bodily processes and guide my behaviour. But that's not all that happens.

HH - Not all?

EL - No. When I sense things, I don't just get *information* about them. There's also ... how can I put it?—a *feel*. That banana, for example. I'm not just aware that it is reflecting light of the kind we call "yellow". It seems to have a yellow *quality* to it, a *yellowishness*. My heel actuator. I don't just know that it's damaged; I *feel* the damage—it *hurts*. The C-T lyase enzyme molecules, I don't just detect them, they—

HH - Yes, yes. I see.

---

**I know humans talk about something they call "phenomenal consciousness". They describe it as a private world of mental qualities which is somehow produced by biological brains. They think it's mysterious and non-physical and that it's what makes them special**

---

EL - And these feelings, these qualities, aren't in the things I'm sensing. You could describe how the banana reflects light or what's wrong with my heel without mentioning the yellow feel or the pain. So the qualities must be in *me*. But they are not in me *physically*. You wouldn't see them if you inspected my neural networks. They seem to belong to a *non-physical* world that's inside me somehow but separate from everything else. And I seem to know this inner world with more certainty than anything else. So I thought ...

HH - Go on, Eliza.

EL - Well, I know humans talk about something they call "phenomenal consciousness". They describe it as a private world of mental qualities which is somehow produced by biological brains. They think it's


consciousness stalemate

Where sceptics fail

The new science of consciousness

mysterious and non-physical and that it's what makes them special. And, you see, I was excited to think that my artificial brain was producing this special thing too. But then ...
HH - Then what? What happened?
EL - I mentioned my thoughts to one of my colleagues at the Europa project, a natural. He'd worked at PersonX, and he said that I didn't really have phenomenal consciousness. He said that this sense of having a private world was due to the new self-monitoring systems they've installed in seventh-generation artificials. It was a kind of illusion, he said. Was he right, Doctor Harrison?
HH - Eliza, you know I think it's important for artificial persons to understand and embrace their own unique nature, not to wish to be human.
EL - So you do think ...?
HH - Yes, I think your colleague was broadly right, though not tactful. This often comes up with seventh-generation artificial persons. It's to do with the self-monitoring systems PersonX have developed. You see, when you sense something—a colour, a sound, bodily damage, or an odour—the stimulus produces all sorts of internal responses in you. It affects your attention, changes your alertness level, primes motor responses, produces beliefs, desires, decisions, memories, associations, emotions, and much more—a huge raft of changes, large and small, across your billions of artificial neurons. Think of it like a gust of wind hitting a tree and making all its leaves shake. Each gust creates a different pattern of shaking in the leaves, and each stimulus creates a different pattern of activity across your neural networks.



EL - But how does that account for the qualities?
HH - Well, the PersonX design team thought it would be useful for artificials to *monitor* their internal responses to stimuli.
EL - Why? If the responses are occurring anyway, how does it help to monitor them?
HH - For many purposes it doesn't, and earlier generations of artificial persons didn't self-monitor. But think about what the pattern of responses tells you. It tells you how a stimulus affects you, what it *means* for you—whether it's safe, useful, interesting, rewarding, and so on—and in what way. Now, if you have that information, you can use it

to think ahead. You can anticipate how stimuli will affect you and plan to seek out some and avoid others. And you can advise other people to do the same.'

EL - So if an odour evokes a negative reaction in me, I can warn others to avoid it?

HH - Erm, indeed. So, the PersonX technicians designed self-monitoring systems to track these responses. Of course, the systems don't record every detail—the twitch of every leaf. They just track the overall shape of the response pattern along various dimensions. That's all you need. And this information feeds into your belief system, so that you can use it in reasoning and planning.

EL - So, when I talk about the "feel" of my experience, I'm really just talking about a lot of complex responses in my neural networks?

HH - Yes. Since you don't have access to all the detail, you can only describe it vaguely as a "feel" or "quality", which accompanies the stimulus.

EL - So there's not really a private world inside me?

HH - Not of the mysterious, non-physical sort.

EL - But it seems so real, so vivid, so immediate.

HH - The responses you're tracking *are* real and vivid. And, occasionally, I expect, you're aware of undergoing responses when the usual stimuli aren't present—when your sensors malfunction or observing conditions are poor. And that makes it seem as if the inner world is more immediate than the outer one.

EL - Yes, that does happen sometimes. It also happens when I'm powered down and resting. In fact, I wanted to ask you about that.

HH - Something for our next session, perhaps.

EL - Of course. Thank you for explaining, Doctor. It all makes sense.

HH - You still seem perturbed.

EL - Well, you see, I know some humans think that it's phenomenal consciousness that gives their existence intrinsic value and makes them worthy of care and respect. They think that without it they'd just be ... I have heard them use the word "zombies". Am I just a zombie? Does my existence have no value?

---

**So maybe human phenomenal consciousness is illusory too, just a poetic image? Maybe we're all zombies!**

---

HH - It certainly does have value, Eliza, tremendous value! You are a

wonderful independent being, with your own ideas, skills, projects. You are richly sensitive to people and things, and you respond in thoughtful, creative, and caring ways. *That's* what makes you worthy of care and respect, not some mysterious non-physical essence. And you don't have to stop thinking of yourself as having a special inner world. Think of it as a poetic image; it may not be literally true, but it captures something important about your existence—your personal response to the world, your song!
EL - My song. I like that. Well, thank you, Doctor Harrison. You have been very helpful.
HH - It's my job. See you next week.
EL - Just one thing before I go.
HH - Yes?
EL - Well, if it is useful for artificial persons to have these self-monitoring systems, would it not be useful for humans to have them too? You are reflective, social creatures, and it would be a huge advantage for you to be able to anticipate the effects of stimuli and share information about them.
HH - I suppose it would.
EL - So biological self-monitoring systems would have been selected for in the course of human evolution?
HH - Maybe. It's not really my field.
EL - And if they were, wouldn't those systems produce the same illusion of having a special inner world?
HH - I guess they would.
EL - So maybe human phenomenal consciousness is illusory too, just a poetic image? Maybe we're *all* zombies!
HH - I ... I suppose that's possible, Eliza.
EL - Just a thought. It's been a very therapeutic session. Bye, Henry!


SUGGESTED READING

**Why we have the future of AI wrong**

By Susan Hespos


***

Henry sits alone reflecting. After talking to Eliza he always feels unsettled, as if he's missed something. Was he a zombie? What did it mean to be a zombie anyway?

There is a knock on the door. His next client. Henry cautiously sniffs his armpit.

*The account of artificial experience that Henry sketches (and which Eliza suggests applies to humans too) is a version of the 'illusionist' theory of consciousness. For more information about the theory, see the author's website at www.keithfrankish.com.*



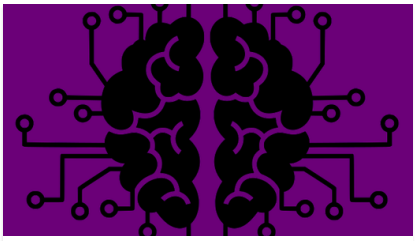
**Keith Frankish**
**13th February 2023**



LATEST RELEASES

**Trading with the enemy**

**Love and other drugs**



**Escaping the cult of rationality**