



Undergraduate assignment feedback form

(to be completed by marker)

Student Name: <i>Bradley Hoanes</i>	
Module title:	Student number:
For completion by marker:	
Comments: <i>This is a very well-researched essay that displays a very good grasp of the literature, & a critical engagement with the reading material. While many of the arguments are well structured & presented, in places closer attention to detail is required. Try to rephrase questions so that they form premises in arguments, & avoid allowing questions to stand in place of counter-claims or sub-conclusions. These things are a matter of presentation & detail that will add clarity & more structure to your arguments. Overall, though, this is a very good essay.</i>	
Marker's signature: <i>RJ</i>	Date: <i>16/11/20</i>
Provisional mark: <i>67</i>	
NB. All marks are provisional until confirmed by the final examination board.	

Critically evaluate the two following challenges to Functionalism: "The Chinese Room" and "The Inverted Qualia Objection."

By Bradley Sloanes

Functionalism is a theory related to the field of Philosophy of Mind; a field which studies the relationship, or whether there is one at all, between the body and the mind.¹ Functionalism proposes that mental states, things that tend to be subjective experiences like pain, sadness, or the belief in certain ideologies or religions, are comprised solely by their "functional roles."² That is their relationship with how we behave and respond to stimuli and their effect on our sensory inputs and other mental states. This belief that mental properties can be implemented by different physical things has the implication of potentially extending "mental states" to computers if it performs or is capable of performing the same functions that we do. Of course, functionalism is a broad theory, and can be expressed through various forms, but this previous point brings us onto one of them: machine functionalism. This position was directly inspired by analogies that rose from the Turing Machine. It is important to talk about the Turing test and its implications in order to better understand the beliefs held by machine functionalists.³

The question of whether machines are capable of thinking like us dates back a surprisingly long time to Descartes in 1637, when he states that automata are able to respond to interactions with people but they could never respond in the linguistically right way to different interactions like any person could. Therefore, machines are considered to be cognitively separate from humans to Descartes.⁴ However, he fails to consider the potential for future machines to become advanced enough to overcome this linguistic barrier to consciousness. Alan Turing, the famous mathematician, philosopher and computer scientist, changed the way in which philosophers thought about this question. Instead of "can machines think?" he proposes "can machines do what we can do?"⁵ This is where the Turing test began to take formation as he demonstrated how this new approach would work by using a test inspired by a party game called "the imitation game." Two people go into separate rooms and are asked questions by a third party, which they respond to by typewriting their answers and the third party must attempt to tell them apart. Both players are trying to convince the third party that they are the other. Instead Turing then goes on to introduce the element that his new test would be renowned for, "what will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often

¹ Ted Honderich, *Problems in the Philosophy of Mind*. Oxford Companion to Philosophy (Oxford: Oxford University Press., 1995). — page number.

² Ned Block, "What Is Functionalism?", *The Encyclopedia Of Philosophy Supplement*. (New York: New York University, 1996), pp. 1,2.

³ Hilary Putnam, 'Minds and Machines', in *Dimensions of Mind: A Symposium*, ed. by S. Hook (New York: New York University Press, 1960), p. 21

⁴ René Descartes, *Discourse on Method and Meditations on First Philosophy* (New Haven & London: Yale University Press, 1996), p. 34,35. ✓ Okay but pp.no's — refer to Discourse or Meds?

⁵ Alan Turing, 'Computing Machinery and Intelligence', *Mind*, 236, (1950), 433-460 (p. 433).

Could you go to the primary text here?

when the game is played like this as he does when the game is played between two people?"⁶ According to Turing it doesn't matter how it accomplishes what it accomplishes, it only matters whether a machine can convince us that it is a human. If you cannot tell the difference, it is because there is no difference. As Putnam says, "anything capable of going through a succession of states in time can be a Turing machine."⁷ In other words, if you can show the cognitive functioning and abilities of a machine to be the same as the cognitive functioning and abilities of humans, then logically the conclusion is that the machine's behaviour is indistinguishable from the humans. However, just because one thing might mirror the actions of another does it really follow that it understands those actions? A machine might be able to analyse the poetry of William Shakespeare, but would it feel and understand the poetry itself? This idea that consciousness can be represented as a set of functions is unsatisfying and problematic. There are many people today who would fail the Turing Test despite the fact that they are human beings, because the idea of what it is to be a human being is a very subjective and, often, personal one.⁸

John Searle's "Chinese Room" argument presents a significant challenge to this aspect of functionalism and the Turing test. The thought experiment assumes the theoretical proposition that artificial intelligence has created a machine that appears to understand the Chinese language by following the instructions of its computer programming. Suppose that this same machine then passes the Turing test since it can mimic the task of language so effectively, by convincing a Chinese speaking human that it is a fellow Chinese speaking human. Does this machine truly understand what it is saying in Chinese, which would make it strong AI, or is it only simulating that ability and is therefore weak AI?⁹ Strong AI would be a machine that possesses a mind and a level of consciousness, whereas weak AI is a machine that is only a simulation of the mind.

The thought experiment presses on further, however. Searle goes on to suggest that if he were to himself manually operate an English version of this machine, he could also do as well as the machine by following its programming and therefore convince someone that he could also speak Chinese. Yet, we know that Searle does not speak Chinese. If Searle is merely mimicking the language, then the conclusion is that the machine is too.

On the other hand, a criticism of this argument is that it wrongly assumes that it is one part of the machine/brain which understands Chinese. No specific part of your brain knows the English language either. Does that mean you do not understand English? No, because it is the whole system (your entire brain itself) which understands English. So if this were to be applied to the Chinese room, it's the whole machine working together that understands Chinese even though the particular part of the System (Searle) does not.¹⁰

However, Searle's retort to this criticism is that it bestows the power of consciousness to the machine without any explanation of how consciousness could be the product of an

⁶ Alan Turing, 'Computing Machinery and Intelligence', *Mind*, 236, (1950), 433-460 (p. 434).

⁷ Hilary Putnam, 'The mental life of some machines', in *Intentionality, Minds and Perception*, ed. by Hector-Neri Castaneda (Detroit, Michigan: Wayne State University Press, 1967), p. 183

⁸ Huma Shah, Kevin Warwick, 'Human Misidentification in Turing Tests', *Journal of Experimental and Theoretical Artificial Intelligence*, 27.2, (2014), 123-135.

⁹ John Searle, 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3.3, (1980), 417-424.

¹⁰ Daniel Dennett, *Consciousness Explained* ([n.p.]: The Penguin Press, 1991), p. 439.

information processing system. He asks, furthermore, what if the man memorises the rules and keeps track of everything in their head? Now the process comes not from the whole system but from one small part of it: the man. Consequently, the original conclusion still applies because if the man, who is also now the machine, does not understand Chinese then the machine does not understand it either.¹¹

On the other hand some philosophers, like Richard Yee, might criticise this response by stating it is nonsensical that the man is now of two minds and that the ability to process Chinese still relied on the Turing machine rather than on the man's ability.¹² Even so, Searle's response still poses a challenge as the assumption that consciousness is a form of information processing still remains.

Yee's criticism of the Chinese Room is developed upon in the virtual mind reply. Marvin Lee Minsky, a cognitive and computer scientist, argues that we have virtual things (like networks and files) that exist within computers. He further argues that consciousness could also exist in a computer in the same way as social networks and these other virtual realities.¹³ A theoretical example of a seemingly infinite information processing system gaining consciousness is that of the "Geth" in the science fiction video game trilogy Mass Effect. They are likened to a shattered mind on their own, but when together each individual (for lack of a better term) Geth provides processing power for them to form a collective consciousness that operates like a brain. They do it so successfully that they are actually capable of thinking outside of their programming and decide to rebel against their organic creators.¹⁴ This is similar to the "China Brain" critique which posits that if you got each citizen of China to emulate one brain cell, using some technology to connect them at all times, this would be functionally equivalent to a consciousness despite no individual having any understanding of the brain itself.¹⁵

Alternatively, if a mind exists in a virtual or simulated reality wouldn't it just be a simulated form of consciousness? Searle argues exactly this point stating that, "no one supposes that computer simulations of a fire will burn the neighbourhood down or that a computer simulation of a rainstorm will leave us all drenched."¹⁶ So likewise, why should we assume that simulated consciousness amounts to actual consciousness?

Conversely, it could be argued that when you set an alarm on your computer or order from a virtual shop you do not claim that it is not really a shop or a clock. Both of these virtual processes have a physical impact on the world.¹⁷ While this may be true, it's still hard to see how this could ever amount to consciousness. Thus, the question still remains, is the mind like the clock or the rainstorm?

Try to avoid finishing a para. with a question (unanswered question)

¹¹ John Searle, 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3.3, (1980), p.423

¹² Richard Yee, 'Turing Machines And Semantic Symbol Processing: Why Real Computers Don't Mind Chinese Emperors', *Lyceum*, 5.1, (1993), 37-59 (p. 44).

¹³ Marvin Minsky, 'Decentralized Minds', *Behavioral and Brain Sciences*, 3.3, (1980), 439-40.

¹⁴ Drew Karpysyn, *Mass Effect* (Bioware, 2007).

¹⁵ David Cole, "The Chinese Room Argument", *The Stanford Encyclopedia Of Philosophy*, 2004, p. 4.

¹⁶ John Searle, 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3.3, (1980), p.429

¹⁷ Nicholas Fearn, *The Latest Answers to the Oldest Questions: A Philosophical Adventure with the World's Greatest Thinkers* (New York: Grove Press, 2007), p. 47.

Another criticism of functionalism is the inverted qualia objection, which is another thought experiment involving a person who is born with a condition (let us call them jade) that makes them see the opposite spectrum of light that you or I would see. For example, when looking at the colour orange Jade would see blue and so forth. If both you and Jade were to be looking at the same orange you would see it as orange and she would see it as being blue, but when asked the colour of the orange you would both report it as being orange. If all of Jade's behaviours and functions are the same, does it really entail that she perceives colour the same way as you or me?¹⁸ The conclusion is that if two people who are functionally identical exist but yet they are experiencing the world differently, it means functionalism is incapable of explaining these unique differences in qualia.

However, this argument has received some criticism by philosophers and scholars. One such critic is Douglas Hofstadter, who sees the inverted qualia objection as being too solipsistic because no one could ever know what was going on in the minds of others. He gives some examples to demonstrate this absurdity, one being "the inverted political spectrum" in which one individual's idea of liberty would be the same as another's idea of subjugation.¹⁹ While the inverted qualia objection may have solipsistic potential, this is not reason enough to dispute the argument in its entirety. Furthermore, it could be argued that having separate experiences of colour and believing two polar opposite ideologies to be the same are two completely different scenarios. We have a clear definition of what liberty and subjugation are because we invented these ideologies, they are things that only exist because we collectively believe in them. If humans did not exist then neither would the idea of liberty or subjugation. However, the same is not true of colour because things are blue, green, red etc whether I believe them to be or not. If humans did not exist, colour would continue to exist. Consequently, Hofstadter's critique fails. *well structured argument*

Yet the critique of inverted qualia does not end there. This is illustrated by David Chalmers who argues that it would be highly unlikely for a person to experience a change in qualia which they fail to notice. This suggests that, even though mental states cannot be fully estimated in functionalist terms, there is a clear correlation between both mental and functional states.²⁰ In addition, if someone were to be afflicted with this inverted qualia from birth then it would be impossible to ever know that they perceive the world differently. Even the individual themselves would not realise this. The inverted qualia objection leaves us with more questions than answers.

To summarise, this essay has evaluated two major challenges to functionalism; that of the Chinese Room and the inverted qualia objection. The Chinese Room thought experiment has proved to be the more durable of the two arguments. It still allows for the potential existence of intelligent machines whilst pointing out the flaws of the Turing test. Furthermore, it correctly establishes that functionalism gives no justification for how consciousness could arise from information processing. Conversely, then inverted qualia objection leaves a lot to be desired as it is an impossible scenario because it is unfalsifiable.

¹⁸ Ned Block, 'Troubles with Functionalism', *Readings in the Philosophy of Psychology*, 1&2, (1980), 268-305.

¹⁹ Douglas Hofstadter, *I Am A Strange Loop* ([n.p.]: Basic Books, 2007), p. Chapter 23.

²⁰ David Chalmers, *The Conscious Mind* (Oxford: Oxford University Press, 1996).

Both receive very well thought out critique and, as a result, we are left with many questions. As technology and AI advances in complexity every moment, perhaps the answers to our questions will soon be revealed.

Bibliography

Block Ned, 'Troubles with Functionalism', *Readings in the Philosophy of Psychology*, 1&2, (1980), 268-305

Block, Ned, "What Is Functionalism?", *The Encyclopedia Of Philosophy Supplement*. (New York: New York University, 1996).

Chalmers David, *The Conscious Mind* (Oxford: Oxford University Press, 1996).

Cole, David, "The Chinese Room Argument", *The Stanford Encyclopedia Of Philosophy*, 2004.

Dennett Daniel, *Consciousness Explained* ([n.p.]: The Penguin Press, 1991).

Descartes René, *Discourse on Method and Meditations on First Philosophy* (New Haven & London: Yale University Press, 1996).

Fearn Nicholas, *The Latest Answers to the Oldest Questions: A Philosophical Adventure with the World's Greatest Thinkers* (New York: Grove Press, 2007)

Hofstadter Douglas, *I Am A Strange Loop* ([n.p.]: Basic Books, 2007).

Honderich Ted, *Problems in the Philosophy of Mind. Oxford Companion to Philosophy* (Oxford: Oxford University Press., 1995).

Karpyshyn, Drew, *Mass Effect* (Bioware, 2007).

Minsky Marvin, 'Decentralized Minds', *Behavioral and Brain Sciences*, 3.3, (1980), 439-40.

Putnam Hilary, 'Minds and Machines', in *Dimensions of Mind: A Symposium*, ed. by S. Hook (New York: New York University Press, 1960)

Putnam Hilary, 'The mental life of some machines', in *Intentionality, Minds and Perception*, ed. by Hector-Neri Castaneda (Detroit, Michigan: Wayne State University Press, 1967)

Searle John, 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3.3, (1980), 417-424.

Shah Huma, Warwick Kevin, 'Human Misidentification in Turing Tests', *Journal of Experimental and Theoretical Artificial Intelligence*, 27.2, (2014), 123-135.

Turing Alan, 'Computing Machinery and Intelligence', *Mind*, 236, (1950), 433-460

Yee Richard, 'Turing Machines And Semantic Symbol Processing: Why Real Computers Don't Mind Chinese Emperors', *Lyceum*, 5.1, (1993), 37-59