INTERNATIONAL BACCALAUREATE

PHILOSOPHY INTERNAL ASSESSMENT

An examination of the mind-body problem and the nature of human consciousness as seen in the protagonist of the show *Westworld*

Link to Syllabus: Core theme — Mind-body problem

Word Count: 1998

Candidate Code: **kdg989**

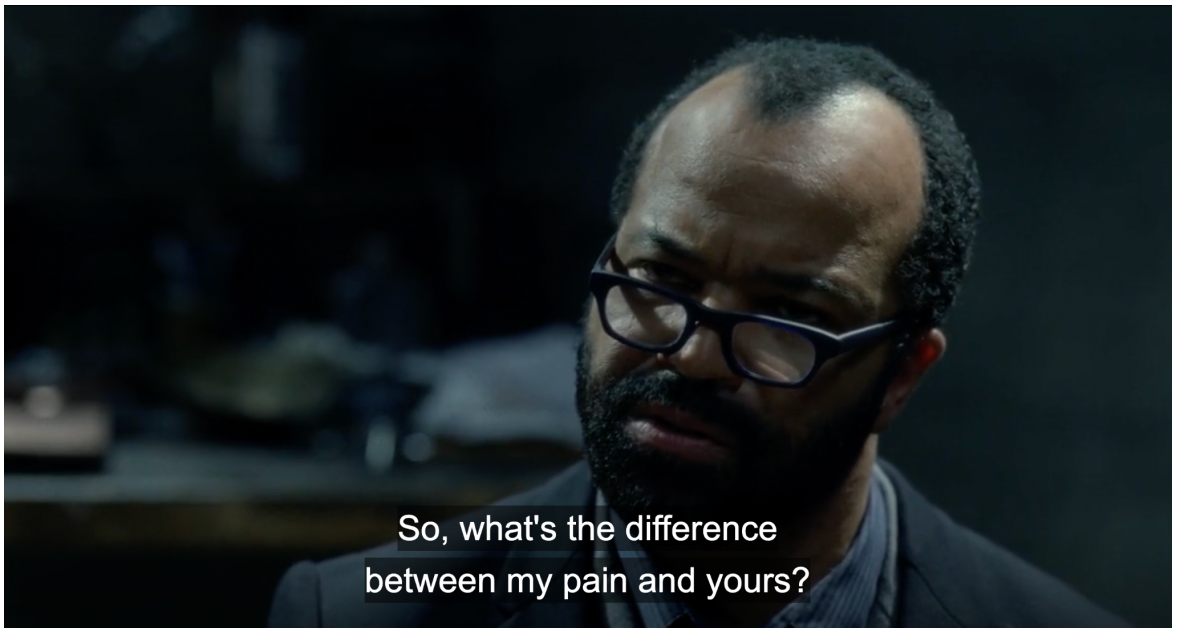
Stimulus — Westworld, Season 1 Episode 8 (“Trace Decay”) - 34:36 - 35:33

Ford: After all, at this moment you are in a unique position: a programmer who knows intimately how the machines work and a machine who knows its own true nature.

Bernard: I understand what I’m made of, how I’m coded, but I do not understand the things that I feel. Are they real? The things I experienced: my wife, the loss of my son.

Ford: …your imagined suffering makes you lifelike

Bernard: Lifelike, but not alive. Pain only exists in the mind, it’s always imagined. So what’s the difference between my pain and yours? Between you and me?



The prompt is a conversation between Dr. Robert Ford and Bernard, protagonists in the show ‘Westworld’. Bernard is a sentient humanoid robot, referred to as a ‘host’ in the show, and Ford is his creator. Each host has memories of certain events, called ‘cornerstones’, programmed into them, which shape their characteristics and behaviour. Bernard’s cornerstone is the loss of his son and subsequent separation from his wife, which leads him to feel and exhibit pain when he is reminded of these events. Something unique about the hosts is that they undergo procedural memory wipes, so things that Bernard learns or remembers as a result of his thinking about his cornerstone get reset at regular intervals. Bernard questioning Ford about the nature of this grief and pain, about whether it is “real”, and about the differences between Ford’s human mind and Bernard’s artificial mind provokes viewers to ponder these questions themselves, and calls to mind the mind-body problem and the nature of consciousness. If Bernard perceives his pain as real, then on what basis can we say that it isn’t? What separates his mind from ours and, in turn, can he be conscious the same way we humans are?

This essay will provide answers to these questions from the perspectives of physicalist, functionalist, and emergentist theories of consciousness.

The type-physicalism identity theory firmly claims that Bernard’s pain isn’t real, and that his consciousness is not the same as a human’s. Type-physicalism is a physicalist theory which states that all mental states can be categorised into types, which can be linked to types of neural states in the brain. In other words, mental states are entirely reducible to physical neural states[1]. Referencing the prompt, the type-physicalist would state that Ford’s ‘pain’ is reducible to a category of physical events in his brain, such as C-fibre nerve firings. Therefore, Ford’s mental state of pain is exclusively identical to the neural state of C-fibre nerves firing. Type physicalism is highly anthropocentric, meaning that mental states could only be caused by neural states existing specifically in human brains[2]. Therefore, only creatures that share humans’ specific neurobiological makeup could experience the same pain as humans. Bernard doesn’t have the same neural state of C-fibre nerves firing as Ford as he doesn’t have a human brain. Therefore, type-physicalism would state that his pain is not the same as Ford’s, and that he can’t share the same consciousness as a human because he doesn’t share the same physical brain or neural states.

---

[1] "Physicalism," Stanford Encyclopedia of Philosophy, accessed July 9, 2022, https://plato.stanford.edu/entries/physicalism/.

[2] "Type-Physicalism, Functionalism, and Eliminative Materialism," Philosophy, n.d.https://philosophy.tamucc.edu/notes/type-physicalism.

Whilst the C-fibre stimulation example can be connected to physical pain, this theory does not explain the neural states that may be connected to emotional pain, which is more akin to the pain Bernard feels. This is an important distinction because emotional pain is often characterised by its effects on people, as opposed to physical pain which is a characteristic in itself. Emotional pain is nuanced and difficult to classify into simple neural states as it comprises various emotions. Additionally, people may react to it differently. After the loss of a loved one, one person may feel anger and another may feel regret. Is it possible to exactly match 'emotional pain' to all the various intricate neural states involved, especially if these states differ from person to person? This shows that the type-physicalism theory is not a complete enough framework to assess whether Bernard's pain is real. Furthermore, its extreme anthropocentricity makes it flawed. How are we to tell that these states are not replicable in other nonhuman systems? Animals have vastly different brains, but do they not also feel emotional pain? This is a description of the principle of 'multiple realizability', which states that mental states can be achieved in various kinds of systems[3], and which is supported by the functionalist theory of consciousness.

The functionalist theory of consciousness states that minds are not produced by specific mediums like brains, but through relations between parts. Using humans as an example — the brain, or specific parts like neurons, wouldn't be considered 'the mind' in itself. Rather, the relationship between the neurons and how they work together to produce outputs is 'the mind'. In other words, the firing and communication of the neurons is what constitutes the mind. It is easy to see how this allows for Bernard to have a mind: provided there is some collection of parts in his 'brain' which work together to produce outputs, he has a mind.

Functionalism extends this logic to mental states like pain. According to the theory, "what matters for being in one or another mental state are the roles that are occupied, not what occupies them"[4]. In other words, a mental state is characterised by the role it plays in the system of which it is a part as opposed to what it's internal constitution is. This idea can be illustrated with the example of a key. The role or function of a key is to open a lock. Whether something is a key or not doesn't depend on its material constituents, shape, or size, but whether or not it can open locks. Provided it meets this condition, it fulfils the role of a key, and therefore, it can be considered a key. Similarly, mental states are determined by the roles they play and the conditions that they meet. Using the prompt's example, the conditions for a mental state like pain may be as follows: it produces a desire to get out of the pained

[3] "Multiple Realizability," Stanford Encyclopedia of Philosophy, accessed July 9, 2022, https://plato.stanford.edu/entries/multiple-realizability/.

[4] David Braddon-Mitchell and Frank Jackson, "*Philosophy of Mind and Cognition,*" cited in *Introducing Philosophy: A Text with Integrated Readings*. Ed. Robert Solomon 9th ed. (New York: Oxford University Press, 2008) 341

state and it produces feelings of anxiety. In the human mind, a specific pattern of neurons firing may create a state that fulfils these conditions. However, Bernard's nonhuman mind may also have an internal configuration that produces a state that fulfils these conditions. Therefore, Bernard may also be feeling the same sensation of pain that a human experiences, regardless of whether he shares the same material brain or not.

There is an intuitive question raised by this theory however. How do we know if Bernard is truly feeling pain or just simulating it? If all that is required to feel pain is the occurrence of different internal arrangements, which for Bernard would be defined by computational processes, then is true emotional pain emergent from simple symbol manipulation? Or is this just a surface-level demonstration of the behaviours associated with pain, without any true feeling of pain itself.

John Searle's theory of biological naturalism addresses these questions quite clearly. The biological naturalist perspective states that simple computational processes are not enough to produce true consciousness, explaining this concept by distinguishing between syntactic and semantic understanding. The former is understanding the rules or formal structure of language, such as acceptable combinations of letters or ordering of these letters[5]. The latter refers to an actual understanding of the meaning conveyed by these arranged symbols[6]. Searle's classic example to demonstrate this is the Chinese Room thought experiment in which a person who knows nothing about Chinese is tasked with translating English to Chinese and vice versa through character cards. This is done using a detailed manual that shows exactly which cards of one language correspond to the cards of another. Searle claims that even if the person acting as the translator is a good processor of languages, he still does not understand the meanings of the characters he is manipulating[7]. In other words, he has a syntactic understanding of Chinese, without a semantic understanding of the characters. Using this example with reference to Bernard, one can say that he, may have a syntactic understanding of pain, that is, what it looks like to be in pain or what actions one exhibits when in pain, but that he doesn't have a semantic understanding of pain: he doesn't understand what it truly means to be in pain.

---

[5] Paul A. Murphy, "John Searle's Syntax-vs.-Semantics Argument Against Artificial Intelligence (AI)," Medium, last modified July 6, 2021, https://becominghuman.ai/john-searles-syntax-vs-semantics-argument-against-artificial-intelligence-ai-33052c688f93.

[6] Murphy, "Searle's Syntax-vs.-Semantics."

[7] "The Chinese Room Argument," Stanford Encyclopedia of Philosophy, accessed July 10, 2022, https://plato.stanford.edu/entries/chinese-room/.

Searle's definition of consciousness in humans is codified by two statements. Firstly, all mental phenomena are higher level features of the brain. Secondly, all higher level features of the brain are caused by lower level neurobiological processes. These lower level processes cause higher level features in the same way that specific arrangements of molecules produce liquidity — the mental states are emergent features of these lower level processes. Whilst Searle's definition is specific to human brains, one could conceivably extend its logic to inorganic 'brains' as well. Searle himself admits that human brains are not the only thing that can produce consciousness[8], stating that our current state of neurobiological knowledge prevents us from concluding that properties of the brain are necessary to produce consciousness, despite the fact that they are definitely sufficient to do so. Searle states that "the brain is a biological machine, and we might build an artificial machine that is conscious", and that "we will not be able to [produce consciousness] artificially until we know how the brain does it"[9]. Therefore, Searle's definition of consciousness doesn't so much deny that inorganic creatures can be conscious as much as claim that the current state of inorganic creatures cannot be.

This implies that, given sufficient complexity of Bernard's 'brain' or mankind's highly advanced neurobiological knowledge, Bernard's pain could indeed be real. It could be real in the sense that he doesn't simply mimic surface-level behaviours associated with pain, but truly does feel it. Searle's concession that the brain is not the exclusive creator of consciousness, paired with the uncertainty around Bernard's syntactic or semantic understanding of emotion, leaves the door open as to whether Bernard's pain is real.

It is important to note that we are assessing Bernard's pain from a third-person perspective, and that because of this, we cannot verify whether his pain is real beyond simply asking him or assessing his behaviour. But is this really that inaccurate? Isn't this how we judge and assess the authenticity of different humans' emotions as well? If someone claims to be in pain, and exhibits the behaviour of someone in pain, don't we just assume that they are truly in pain? As it pertains to Searle's distinction between simulation and 'the real thing', it can be argued that it doesn't matter provided that the two are sufficiently similar. Because Bernard's actions are consistent with someone who experiences pain, and because he can express why he feels it, we can assume his pain is real because it appears sufficiently so to everyone else.

---

[8] Pascal Ludwig, "Could a machine think? Alan M. Turing vs. John R. Searle" (lecture, Unité de formation et de recherche de philosophie et sociologie, Université Paris Sorbonne Paris IV, n.d).

[9] Dan Turello, "Brain, Mind, and Consciousness: A Conversation with Philosopher John Searle," Library of Congress Blogs, last modified March 3, 2015, https://blogs.loc.gov/kluge/2015/03/conversation-with-john-searle/.

However, this logic cannot extend to the second question of whether Bernard's consciousness is like that of a human's. Here, it is important to return to the distinction between simple physical pain and emotional pain. The former is an isolated state or feeling, whereas the latter has profound effects on many other aspects of one's person. Whilst Bernard's pain may be real, he definitely does not have a human consciousness. This is because of the complexity of human experience, which extends beyond the feeling of isolated mental states. One's feeling of pain does not define their consciousness and experience, rather these things are created through the effect of this pain on the person's growth and decisions. Human conscious experience is not centred and built around specific states, as Bernard's is. Bernard's pain is eternal and defines him: it is what allows him to understand his actions, it entirely frames his view of the world, and it is what his consciousness is built around. The human consciousness is malleable, subject to change, and this is perhaps what defines the human experience: the ability to evolve. One's evolution in attitudes, fears, and ambitions in response to grief or pain is what defines their consciousness. Put simply, pain is what creates Bernard's conscious experience, whereas for humans pain is simply something that helps the conscious experience to develop. Bernard cannot undergo this development unless his code allows him to. Unless he is able to decide how he will change, grow, and evolve from his pain, and not just have it be the framework with which he understands his actions, Bernard and Ford cannot be the same. Until he is able to do this, Bernard will remain only lifelike, but not truly alive.

Bibliography

Braddon-Mitchell, David, and Frank Jackson. "*Philosophy of Mind and Cognition"*. London: Blackwell, 1996, cited in *Introducing Philosophy: A Text with Integrated Readings*. Ed. Robert Solomon 9th ed. (New York: Oxford University Press, 2008) 341.

Ludwig, Pascal. "Could a machine think? Alan M. Turing vs. John R. Searle." Lecture, Unité de formation et de recherche de philosophie et sociologie, Université Paris Sorbonne Paris IV, n.d.

"Multiple Realizability." Stanford Encyclopedia of Philosophy. Accessed July 9, 2022. https://plato.stanford.edu/entries/multiple-realizability/.

Murphy, Paul A. "John Searle's Syntax-vs.-Semantics Argument Against Artificial Intelligence (AI)." Medium. Last modified July 6, 2021. https://becominghuman.ai/john-searles-syntax-vs-semantics-argument-against-artificial-intelligence-ai-33052c688f93.

"Physicalism." Stanford Encyclopedia of Philosophy. Accessed July 9, 2022. https://plato.stanford.edu/entries/physicalism/.

"The Chinese Room Argument." Stanford Encyclopedia of Philosophy. Accessed July 10, 2022. https://plato.stanford.edu/entries/chinese-room/.

Turello, Dan. "Brain, Mind, and Consciousness: A Conversation with Philosopher John Searle." Library of Congress Blogs. Last modified March 3, 2015. https://blogs.loc.gov/kluge/2015/03/conversation-with-john-searle/.

"Type-Physicalism, Functionalism, and Eliminative Materialism." Philosophy. n.d. https://philosophy.tamucc.edu/notes/type-physicalism.