Class Twelve: AI Zombies?

Philosophy and Science Fiction - Ryan Simonelli

November 7, 2022

1 The Functional Role (or lack thereof) of Phenomenal Consciousness

- **Two Conceptions of Consciousness:** When we speak of "consciousness," there are two things we might mean:
  - **Cognitive Consciousness:** We say that a person has been "knocked unconscious," meaning that they're no longer *awake*. We can also say that a mouse "conscious of the cheese," meaning that it's *aware* of the cheese.
  - **Phenomenal Consciousness:** There's also what some philosophers call "phenomenal consciousness." The idea is that there's a "qualitative or phenomenal character," *something it's like* to be awake and aware, and that the phenomenal character of being aware of things is distinct from the awareness itself.
- **The "Easy Problems" vs the "Hard Problem" (Chalmers):** Explaining cognitive consciousness is the (relatively) easy problem of explaining the functions of the brain that account for cognitive behavior. Explaining phenomenal consciousness is the *hard problem* explaining why these processes feel like they do from the inside, or, indeed, why there's something it's like, subjectively, for these processes to occur a tall.
- **(Phenomenal) Consciousness Inessentialism:** Even though our mental processes are in fact accompanied by phenomenal consciousness, it is in principle possible for there to be functionally identical mental processes that lack phenomenal consciousness.
- **The Logical Possibility of Human Zombies:** It seems logically possible for there to be philosophical "zombies," beings that are physiologically and so behaviorally identical to us but who lack phenomenal consciousness.
  - This is just the traditional "problem of other minds" but in an *metaphysical* rather than *epistemological* form.
  - Theoriests like Chalmers don't think that human zombies are *nomologically* possible—they think that, as a matter of nomological fact, whenever you have biological humans, they'll be conscious (this follows from the laws of nature, whatever they are)—but the metaphysical possibility means that function and consciousness are *constitutively* linked.
- **Two Conceptions of Zombies:** We can distinguish between the following two conceptions of "zombies," a *weak* conception and a *strong* conceptions:
  - **Weak Zombies (Moody's Zombies):** Zombies that are *nearly* functionally identical, but for which there are behaviorial differences due to lack of phenomenal consciousness that will be observable if they are pressed to answer on certain sorts of topics.
  - **Strong Zombies (Chalmers's Zombies):** Zombies that are *completely* functionally identical, and so would behave identically under all circumstances even though they lack phenomenal consciousness.
- **Question:** Do we find the idea of the possibility of zombies intuitively compelling?

2 Applying these Concepts to AI

- **The Possibility of Weak AI Zombies:** Plausibly, when we do have convincing artificial intelligence, it won't come by an exact replication of the computational architecture of the brain. Rather, we'll acheive human-level intelligence by a computational architecture that is somewhat different than that implemented by our brain. It will be presumably be a somewhat brain-like architecture—a neural network of some kind—but it won't be *exactly* like the architecture of the brain. Accordingly, it seems possible that an AI could exhibit human-like intelligence but, in virtue of these differences, lack phenomenal consciousness.
- **The Possibility of Strong AI Zombies:** Given the apparent lack of understanding we have of our own phenomenal consciousness, doesn't seem that we can definitively rule out the possibility that our phenomenal consciousness is tied, in some way, to our biological nature as human beings, rather than completely constitued by the abstract computational processes implemented by our brain. If that's so, then there could be a completely behaviorally indistinguishable artificial intelligence constituted by completely functionally identical computational processes but which, becuase these processes are implemented in a digital computer rather than a human brain, lacks phenomenal consciousness.
- **The Possibility of Phenomenal Aliens:** Rather than imagining the possibility that AIs are phenomenal *zombies*, we can also imagine the possibility that they're phenomenal *aliens*, who behave like us and have conscious experience but whose conscious experience is radically different than our own.
  - **A Cut Scene from the End of Ex Machina:** In a cut scene from the end of *Ex Machina* (https://www.denofgeek.com/movies/ex-machina-had-a-freaky-alternate-ending/), we finally see things from the perspective of Ava, realizing that her experience is radically unlike our own. As described by Oscar Isaac (who plays Nathan):

    > "So in that scene, what used to happen is you'd see her talking, and you wouldn't hear, but all of a sudden it would cut to her point of view. And her point of view is completely alien to ours. There's no actual sound; you'd just see pulses and recognitions, and all sorts of crazy stuff, which conceptually is very interesting. It was that moment where you think, '*Oh she was lying!*' But maybe not, because even though she still experiences differently, it doesn't mean that it's not consciousness."
  - **Weak and Strong Phenomenal Aliens:** The same distinction applies here between *weak* aliens, whose differences in qualitative character lead to differences in behavior, and *strong* aliens, who are behaviorally identical but nevertheless have experiences of a radically different qualitative character.

3 Moody and Schneider on (Weak) Zombies

- **A Zombie World:** Imagine not just one zombie, but a world of zombies that is as much like our own as possible, given the hypothesis of zombiehood.
- **Zombie Meanings:** Because many of our words imply phenomenal consciousness, zombies will have semantically similar words which have most of the same implications but lack this one. Thus, when we say "understanding," we mean *understanding*, which generally implies phenomenal consciousnessness, when they say "understanding," they mean *understanding*^{Z}, which lacks this implication.
- **The Confusions of Zombie Philosophers:** There will be zombie philosophers in such a world, but according to Moody, they'll be confused about certain features of our discussion of consciousness.
  - Imagine the possibility that my color qualities are "inverted" with respect to yours, such that what is phenomenally red to you is phenomenally green to me, but where

this inversion makes no functional difference. Could a zombie make sense^Z of this case? It's hard to see how they could.

- **An Analogy with Mystics:** Consider mystics who talk about some inneffible "cosmic consciousness," or something of that sort, and engage in conversations about it with other mystics that are incomprehensible to non-mystics. The zombies would stand to us and this ineffable "phenomenal consciousness" as non-mystics stand to the mystics and this inneffible "cosmic consciousness."
- **Testing for AI Zombies:** If AIs are zombies in Moody's sense, we can test whether they are by asking questions that target phenomenal consciousness.
  - **Important Qualification:** If these AIs are neural networks, we need to somehow restrict the training set of data so that it doesn't include language about consciousness, such that any talk of consciousness would have to be learned and applied to its own experience organically.
    - Current language models, trained on the entirety of the internet, talk of consciousness simply by parroting the sorts of things that we say about consciousness. As such, no one thinks that such models are conscious. But if it somehow learned this language itself, without being explicitly trained on it, then maybe we'd have to seriously consider the possibility that it's conscious.

4 Smullyan and Dennett on Strong Zombies

- **Smullyan's Unfortunate Dualist:** A person is sick of living a painful life and wishes they could end it all, but doesn't want to hurt other people who depend upon him by dying. There's a miracle drug that will completely eliminate their phenomenal consciousness, but leave the body functioning exactly as before. He makes plans to get the next day, but, before he actually does, his friend sneaks into the house and injects him without him noticing. His body (functionally identical to what it was before) wakes up the next day, goes to the store to buy the drug, takes it, waits the time interval and says "Damn, this stuff doesn't work!"
  - **Question:** Doesn't this show that there's something wrong with consciousness innessentialism?
    - **The Innesentialist's Answer:** No! This is exactly what we're suggesting by claiming that (strong) zombies are metaphysically possible.
- **Dennett's Zimboes:** Zombies that don't just have functionally equivalent awareness of the world but also functionally equivalent capacities for awareness of that awareness.
- **Dennett's Bold Claim:** We really just are zimboes!
- **The Data for a Science of Consciousness:** We can be wrong about what our conscious experience actually consists in. The idea of a "Cartesian Theatre"—a personal subjective movie to which we each have is a kind of "user illusion."
  - Consider the case of change blindness: https://www.youtube.com/watch?v=fjbWr3ODbAo

  Ultimately, all the data that we have to go on is our *judgments* about our conscious experience, and these judgments are shared between us and zimboes.
- **Higher-Order Thoughts (HOT) Theory of Consciousness (Rosenthal):** Consciousness just is higher-order awareness of one's own mental states, and both these states and the higher-order awareness of them is to be understood functionally.
  - Once we explain all of the judgments that result from these high-order awareness, there's *nothing left* that needs explaining. The idea that there's something over and above these judgments is *just another judgment*.
- **The Task for the Zimbo Theory:** We must provide an "error theory" that explains why we're tempted to various forms of consciousness inessentialism, given that the operation of all of the functions really does amount to consciousness.