From P-Zombies to Substance Dualism

Journal of Consciousness Studies (forthcoming)
Penultimate Draft

Perry Hendricks
perryhendricks990@gmail.com

Abstract: P-zombies are creatures that are physically (functionally, behaviorally) like you and I and yet lack phenomenal consciousness. If such creatures are possible, it's (typically) taken to show property dualism is true: phenomenal consciousness isn't reducible to—nor does it supervene on—physical states. If inverted qualia are possible, it's possible that you and I have identical physical states and yet you see tomatoes as green and I see tomatoes as red. If this is the case, then (again) property dualism is (typically) taken to be true. In this article, I'll show that p-zombies and inverted qualia—if they are actually possible—prove more than previously thought: if one thinks p-zombies or inverted qualia are possible, then she should endorse one of the four following theses: (i) substance dualism, (ii) we have *even more* reason to reject p-zombies and inverted qualia since they entail an even more radical conclusion than previously thought (i.e. substance dualism), (iii) eliminativism about selves, or (iv) friends of p-zombies and inverted qualia have homework: they need to show a relevant, plausible disanalogy between arguments for p-zombies and inverted qualia as traditionally stated and my parallel arguments that entail substance dualism. My purpose here isn't to defend any of these particular options. Instead, my purpose is to highlight that these are the four options available to take.

1. Introduction

In terms of critiques of physicalism, two arguments loom large: The P-Zombie Argument and The Inverted Qualia Argument.[1] Proponents of these arguments typically take themselves to show that *property dualism* is true.[2] In this article, I show that proponents of these arguments are forced into one of the following four positions: either (i) substance dualism, (ii) we have even more reason to reject The P-Zombie Argument and The Inverted Qualia Argument, since they entail a more radical thesis than traditionally thought (i.e. substance dualism), (iii) eliminativism about selves, or (iv) those who endorse The P-Zombie Argument and The Inverted Qualia Argument have homework: they need to show a relevant, plausible disanalogy between these arguments as traditionally stated and my parallel arguments that entail substance dualism. I don't

[1] I'm *not* suggesting that these are the only arguments used against physicalism. There are other such arguments, e.g. Cutter (2020), Hasker (1999), Plantinga (2006), Swinburne (1997 and 2019), and so on. My only claim is that these two arguments are prominent.

[2] More exactly, proponents of The P-Zombie Argument take themselves to show that either panpsychism is true or (at least) property dualism is true. In this article, I'm focusing on those who think it shows (at least) property dualism is true.

make any claims about which horn is the right one to take. Instead, my purpose is just to illuminate the four options available. For ease of read, I will refer to the disjunction of these four options as a *tetralemma*.

The structure of this article is as follows. First, I'll briefly explain The P-Zombie Argument and what it's typically taken to entail (i.e. property dualism). After this, I'll show that a parallel argument can be made in favor of substance dualism, and that this forces proponents of The P-Zombie Argument into a tetralemma. Next, I'll briefly explain The Inverted Qualia Argument and what it's typically taken to entail (i.e. property dualism). I'll then show that a parallel argument can be made in favor of substance dualism, and that this forces the proponent of the argument into a tetralemma. However, before turning to these arguments, we first need to cover some terminological issues, and it's to this issue I now turn.

1.1 Defining Terms

For the purposes of this article, I'm going to take *me*—whatever it is that I am—to be a *point of view* or a *first-person perspective*. So, if some creature has a different *point of view* than me or a different *first-person perspective*, it's not me—it's someone else. (Or, alternatively, if it lacks a point of view or first-person perspective entirely, it's some*thing* else.) And I'm going to understand a *self* to be a point of view or first-person perspective, and I will assume that a self is not a mere property. Selves, understood this way, have properties: my first-person perspective might have the property of being in pain right now, or it might have the property of having persisted for some number of years, or it might have the property of desiring food. Lastly, I'm going to understand things that have properties and are not themselves mere properties to be *substances*. As such, selves—points of view or first-person perspectives—are substances. What distinguishes one self from another? Or, what distinguishes one point of view from another? It isn't merely a different set of beliefs or history. For example, another self could have the same beliefs, desires, and intentions as me, and it could perform the same actions as me, and yet not be me. It could be someone else—a different self.[3] Rather, what distinguishes facts about you and I—facts about personal identity—are facts about *what it's like to be me* and facts about *what it's like to be you*. For example, you and I might have identical beliefs, intentions, and experiences, but there can still be a difference between *me* and *you*, in that what it's like to be me is one thing and what it's like to be you is another thing. The facts about what it's like to be me, I'll take it, make it such that I am (i.e. are sufficient for something to be) me. Or, in other words, whatever it is that knows what it's like to be me *is* me—there couldn't be something that knows what it's like to be me and yet fail to be me. Put differently yet again, one can know all the facts about my phenomenal consciousness (e.g. that I experienced the taste of an apple at time $t1$, that I experienced pain at $t2$, and so on) and yet not know what it's like to be me, since—on the understanding laid out above—I'm a *point of view*, and my point of view isn't captured by my phenomenal consciousness (it's the thing having the phenomenal conscious experiences).

[3] This is similar to Rudder Baker's (2013) robust first-person perspective.

Moreover, since selves are substances and I'm a self (as defined above), *I'm* a substance: *I'm* a thing that bears properties and I'm not *myself* a property (e.g. I have the property of experiencing bliss when I taste chocolate or of being bored when I listen to opera, but I'm not merely a property). Those who think there are no such facts are what I will call eliminativists about personal identity or selves.[4]

2. The Zombie Wars

P-zombies are just like you and I physically, functionally, and behaviorally: if you burn them, they'll scream, if you greet them, they'll greet you, and if you ask them to pass the salt, they'll comply. But there's one crucial difference between p-zombies and you and I: whereas you and I have phenomenal consciousness, p-zombies are devoid of all conscious experience. So, for example, when they burn their hand on an oven, they say "Ouch!" and move away quickly, but they don't feel pain (or anything else)—p-zombies might act like you and I, but they completely lack conscious experience. If these creatures are possible, there's a significant upshot: if p-zombies are possible, phenomenal consciousness doesn't *supervene* on physical states. This is because—if p-zombies are possible—we can have two physically identical creatures (say, you and your p-zombie) that differ with respect to phenomenal conscious properties (namely: you have phenomenal consciousness but your p-zombie doesn't), and hence phenomenal conscious properties don't supervene on physical properties. However, physicalism *minimally* entails that phenomenal consciousness supervenes on physical states (e.g. Kim 1990 and 2007). So, physicalism is false. Furthermore, since—given the possibility of p-zombies—phenomenal consciousness doesn't supervene on the physical, it follows that phenomenal consciousness isn't itself physical. And this means that property dualism is true: phenomenal conscious properties are non-physical properties.

Of course, it's controversial whether p-zombies are possible: physicalists will deny their possibility, and some even deny their *conceivability* (e.g. Kirk 2008 and Tye 2006)! However, some think that p-zombies really are possible, and this commits them to property dualism.[5]

My purpose here is not to adjudicate the debate about whether p-zombies are possible—I know of no way to settle that dispute. Instead, I will show that those who think p-zombies are possible must adopt one of four theses laid out below.

---

[4] Some might think this is a question-begging way to view personal identity. However, I haven't begged any questions thus far, I've just stipulated my terms. I've not claimed that eliminativism about personal identity (as I've defined it) is implausible. Instead, I've only said what it amounts to.

[5] The most notable proponent of this is Chalmers (e.g. 1996 and 2010). Again, some take this to entail that either panpsychism is true or physicalism is false and (at least) property dualism is true. I'm focusing here on those who take this to show that (at least) property dualism is true.

2.1 From P-Zombies to Me-Zombies

Recall that a p-zombie is a creature that is physically (functionally, behaviorally) identical to a human but has no conscious experiences at all—it lacks phenomenal consciousness. Think of your zombie twin: she's identical to you in terms of molecular structure, function, and behavior. But she lacks conscious experience entirely—there's nothing it's like to be a p-zombie. One strand of evidence for the possibility of p-zombies is that they're (arguably) conceivable: we can imagine a creature physically (functionally, behaviorally) identical to you or me that's completely devoid of conscious experience, and we detect no contradictions when we imagine this being. This gives us (defeasible) reason to think p-zombies are possible. And, therefore, this gives us reason to think that physicalism is false (since the mental doesn't supervene on the physical) and reason to think that property dualism is true, since phenomenal consciousness isn't identical to, nor does it supervene on, the physical.

A *me-zombie*, let's say, is a physically (functionally, behaviorally) identical creature to me that *isn't me*—it has a different personal identity or it has *no* personal identity (on my definition of personal identity). Put differently, a me-zombie is a physical duplicate of my body that lacks a first-person perspective.[6] A me-zombie and a p-zombie look to be *equipossible*: our case for thinking one is possible seems to be as strong as our case for thinking the other is possible. For example, we can imagine a creature physically (functionally, behaviorally) identical to me and yet it lacks phenomenal consciousness. That's a p-zombie. But we can imagine further that the creature is physically (functionally, behaviorally) identical to me and yet it lacks my personal identity—it's devoid entirely of a point of view. Think about it this way: in the same way that it's coherent to suppose that a p-zombie lacks phenomenal consciousness, it's coherent to suppose that a me-zombie, in addition to lacking phenomenal consciousness, also lacks a first person perspective—there doesn't appear to be a logical contradiction there.[7] And this supports the view that a me-zombie is conceptually coherent, which supports the view that a me-zombie is conceivable, which supports the view that me-zombies are logically possible. However, if a me-zombie is possible, then it follows that *I* don't supervene on the physical, which means that *I* am non-physical. And as we saw above (Section 1.1), I'm a substance, and so substance dualism is true.

Of course, one might claim that to be me *just is* to have certain mental states or brain states. I don't dispute this claim. Instead, my point is that a parallel response can be made with respect to The P-Zombie Argument. That is, one might claim that to have phenomenal consciousness *just is* to be physically (behaviorally, functionally) identical to a creature like you and I. And this means that p-zombies and me-zombies are in roughly the same evidential boat.

---

[6] In this sense, a me-zombie is *more* than a p-zombie: in addition to lacking phenomenal consciousness, it lacks a first-person perspective or has a different first-person perspective.

[7] Indeed, I'm inclined to think that having a first-person perspective entails having phenomenal consciousness, meaning that lacking phenomenal consciousness entails lacking a first-person perspective. However, I won't argue for this view here.

However, this leaves the proponent of The P-Zombie Argument with four options: she may either (i) endorse substance dualism, since she sees that the case for p-zombies and me-zombies is equally strong, and she thinks the case for p-zombies is strong; or she may (ii) take this to be a reason to reject The P-Zombie Argument, since it has more radical implications than she thought: if p-zombies and me-zombies have roughly equal support and it's crazy to think me-zombies are possible, she might see that as reason to reject p-zombies as being possible; or she may (iii) be an eliminativist about personal identity or selves: perhaps she doesn't think that there are such things as first-person perspectives or selves, and this gives her reason to think me-zombies are philosophically innocuous; or she may (iv) understand this parallel between p-zombies and me-zombies to provide her with homework: she needs to find a plausible disanalogy between p-zombies and me-zombies that allows her to accept the possibility of the former but not the latter.

We can state the argument more formally as follows:

(1) P-zombies are possible.
(2) If p-zombies are possible, me-zombies are possible.
(3) If me-zombies are possible, then substance dualism is true (i.e. (i))
(4) Therefore, substance dualism is true.

This argument forces us to either accept substance dualism (i.e. (i)), reject The P-Zombie Argument since it leads to radical conclusion (i.e. (ii)), reject the existence of selves as I've defined them (i.e. (iii)), or point out a relevant difference between p-zombies and me-zombies (i.e. (iv)).

This is a no-judgment zone: I don't provide any answers for what's the best route to take in response to this tetralemma. Rather, my purpose is just to bring to light this tetralemma facing proponents of The P-Zombie Argument.

3. Inverted Qualia

The Inverted Qualia Argument, very roughly, goes like this. Jane lives a particularly mundane life on Earth: she wakes up every morning and goes to work at the tomato factory, gets home, eats dinner, and goes to bed. Janice lives on Twin Earth, which is physically identical to Earth, and Janice is physically (functionally, behaviorally) identical to Jane, and she too lives a mundane life. She wakes up every morning and goes to work at the tomato factory, gets home, eats dinner, and goes to bed. But here's the kicker: when *Jane* looks at a tomato, it appears red to her and when *Janice* looks at a tomato, it appears green to her. But if this scenario is possible, then qualia don't supervene on the physical, since it's possible to have a difference in qualia

without having a physical difference. And therefore physicalism is false, and qualia are non-physical. And this means that property dualism is true.[8]

As with p-zombies, there have been detractors of The Inverted Qualia Argument. For example, Churchland (2006) argues that if qualia are inverted, then there will have to be some sort of physical change, and so the scenario above just isn't possible. My purpose here isn't to defend this argument, and so I won't adjudicate this dispute here.

3.1 From Inverted Qualia to Inverted Selves

So, some think inverted qualia are possible, and this entails that qualia are non-physical. However, it looks equally possible to invert *selves*. For example, suppose that Sophia lives a relatively mundane life. She goes to work at the fidget factory, gets home, eats, and goes to bed. She eventually dies at the age of 90. It seems possible that we could have held all the physical facts the same and yet a different self—first-person perspective—could have occupied Sophia's body, performed all the actions she performed, believed everything she believed, desired everything she desired, and so on. However, if that's the case, then selves don't supervene on the physical—we can switch out Sophia (a self) with a different self—and this means that selves are non-physical, and substance dualism is true. Or consider *me*: I have various desires, beliefs, have performed various actions, and so on. But it's possible that a different first-person perspective has the same desires, beliefs, performs the same actions, and yet *isn't me*. Having my beliefs (etc.) is one thing. Being *me* is another. And so it looks like you can invert me, a self, in the same way you can invert qualia. And this means that I—a substance—am non-physical, and substance dualism is true. Put differently, inverted qualia and inverted selves appear equipossible. Think about it this way: a human that is microphysically identical to me might act, believe, and intend in all the same ways I do and yet not be me—some other self could do this. And this means that you can have a creature physically (behaviorally, functionally) identical to me and yet not be me (since it's a different self). And hence *I* don't supervene on the physical, and hence *I'm* not physical.

Arguments against inverted qualia will likely equally apply to inverted selves. For example, one might claim that if you *really* invert selves, then there *must* be a physical change—there's just no way to do that and for there to not be a physical difference. This strikes me as the most plausible response to the possibility of inverted selves. I don't dispute this objection here. Rather, my claim is that this objection appears to be roughly as strong as it is to inverted *selves* as it is to inverted *qualia*: it cuts against both roughly equally.

Since support for inverted qualia and inverted selves is roughly the same and objections against both are roughly equally strong, this poses the following tetralemma to one who endorses The Inverted Qualia Argument: she may either (i) embrace substance dualism, since the case for

[8] For defenses of The Inverted Spectrum Argument, see e.g. Block (1990), Chalmers (1996), and Kim (2007).

inverted qualia and inverted selves are roughly equally strong and she thinks the case for inverted qualia is strong; or she may (ii) take this as a reason to reject The Inverted Qualia Argument, since it has an even more radical conclusion than she previously thought; or she may (iii) be an eliminativist about personal identity or selves: perhaps she doesn't think that there are such things as first-person perspectives or selves, and this gives her reason to think inverted selves are philosophically innocuous; or she may (iv) understand this parallel between inverted qualia and inverted selves to provide her with homework: she needs to find a plausible disanalogy between inverted qualia and inverted selves that allows her to accept the possibility of the former but not the latter.

The argument can be stated more formally as follows:

(5) Inverted qualia are possible.
(6) If inverted qualia are possible, inverted selves are possible.
(7) If inverted selves are possible, then substance dualism is true.
(8) Therefore, substance dualism is true.

The argument here forces us to either accept substance dualism (i.e. (i)), reject The Inverted Qualia Argument since it leads to radical conclusion (i.e. (ii)), reject the existence of selves as I've defined them (i.e. (iii)), or point out a relevant difference between inverted qualia and inverted selves (i.e. (iv)).

Again, this is a no judgment zone: proponents of The Inverted Qualia Argument may take whatever horn of the tetralemma they see fit. My purpose is just to bring to light this tetralemma.[9]

Importantly, my tetralemma will apply to other arguments for property dualism as well. For example, it will also apply to The Knowledge Argument, which claims (very roughly) that one can know all the physical facts there are to know and yet not know facts about phenomenal consciousness (e.g. what it's like to see red).[10] However, one can run my tetralemma as follows: we can know all the physical facts without knowing what first-person perspectives there are. And from this, it follows that first-person perspectives—selves—are non physical. And this forces us to choose between (i) endorsing substance dualism, (ii) rejecting the knowledge argument since it leads to more a radical conclusion than previously thought, (iii) being an eliminativist about selves, or (iv) taking this to show that we need to identify a relevant difference between knowledge of first-person perspectives and knowledge of qualia, such as the experience of the color red.

[9] The difference between The P-Zombie Argument and The Inverted Qualia Argument, then, is that the former is about *absent* phenomenal consciousness whereas the latter is about (in a sense) *switching* it.
[10] For discussions of The Knowledge Argument, see.g. Jackson (1982) and (2003), Ludlow, Nagasawa, and Stoljar (2004), and Nagasawa (2008).

4. Objections

Perhaps one might object that there's a difference in epistemic access between our knowledge of mental states and our knowledge of first-person perspectives. Roughly, the idea is that we have direct access to mental states but not to our first person perspective. So whereas we can—according to proponents of the P-Zombie Argument and The Inverted Qualia Argument—directly see that we have phenomenal consciousness and can see that a p-zombie wouldn't have it or that qualia could be inverted, we *can't* (at least as) easily see that a me-zombie is possible or that it's possible to invert selves.[11]

This is true enough: there's controversy about whether there are selves, and some argue that we don't have direct access to such things (if they exist).[12] I grant that this is a difference proponents of The P-Zombie Argument and The Inverted Qualia Argument could cite in an attempt to ward-off substance dualism. Of course, there are strong arguments in favor of first-person perspectives,[13] and if we have good reason to think that there are such things, even though we lack direct access to them, this doesn't seem to me to dramatically diminish our justification for thinking first-person perspectives could be inverted or lacking in the case of p-zombies.[14] Furthermore, if we don't have access to a first-person perspective, that might actually *increase* our confidence that a me-zombie is possible, since it would reduce our confidence that there are first-person perspectives at all. (Of course, this would push one toward the eliminativist position (i.e. (iii)). But those who are inclined toward realism about first-person perspectives won't accept this option.)

My purpose here isn't to adjudicate any of the above disputes. The point here is that to deny selves is to take horn (iii) of my tetralemma, and to argue that our epistemic access to selves and phenomenal consciousness is importantly different is to take horn (iv) of my tetralemma. So, this isn't an objection to the argument I've presented in this article: my purpose isn't to advocate a particular horn of the tetralemma. Rather, my purpose is to make clear the available options.

Another objection one might make is that to claim that a phenomenal duplicate of me *just is* me.[15] In other words, to have the same phenomenal conscious states that I do is to be me. This would mean that inverting selves doesn't make sense and isn't possible. This is no doubt a possible move that can be made on behalf of the proponent of The Inverted Qualia Argument. However, this doesn't conflict with the argument I've made in this article. Rather, this is just to endorse option (iv) of my tetralemma: it's an attempt to identify a relevant difference between

---

[11] Thanks to a reviewer for bringing this point to light.
[12] Hume (1986) famously argues for this point.
[13] E.g. Builes (2023) and Rudder Baker (2013).
[14] Alternatively, one might argue that we do have direct access to selves: I can directly access *me* having a pain, or *me* having a pleasurable sensation. Thanks to a referee for suggesting this point.
[15] Thanks to a reviewer for raising this objection.

inverted qualia and inverted selves by claiming that inverted selves aren't possible. And so this point doesn't conflict with my argument.[16]

Of course, more objections could be raised. However, these kinds of objections won't be objections *to my argument*. Rather, they will turn out to be reasons to endorse a particular horn of the tetralemma I've laid out above.

4. Conclusion

I've claimed that defenders of The P-Zombie Argument and The Inverted Qualia Argument each face a tetralemma: they must either (i) endorse substance dualism, (ii) take my argument as reason to reject The P-Zombie Argument and The Inverted Qualia Argument, (iii) be eliminativists about personal identity, or (iv) take it as homework to find relevant disanalogies between p-zombies and me-zombies and between inverted qualia and inverted selves. This is a no judgment zone: I've not advocated for any particular horn of the tetralemma. Instead, I leave it to the reader to decide which horn is most plausible.[17]

5. References

Builes, D. (2023). "Eight Arguments for First-Person Realism." *Philosophy Compass* 19(1), n.p. DOI: https://doi.org/10.1111/phc3.12959

Chalmers, D. (1996). *The Conscious Mind.* New York: Oxford University Press.

Chalmers, D. (2010). *The Character of Consciousness*. New York: Oxford University Press.

Churchland, P. (2006). "Chimerical colors: some phenomenological predictions from cognitive neuroscience." *Philosophical Psychology.* 18 (5), pp. 527–560. DOI: https://doi.org/10.1080/09515080500264115

Cutter, B. (2020). "The Modal Argument Improved" *Analysis* 80 (4), pp. 629-639. DOI: https://doi.org/10.1093/analys/anaa023

Dennett, D. (1991). *Consciousness Explained.* New York: Basic Books.

[16] While I won't dispute this objection here, I'll say a brief word about why it doesn't strike me as the most plausible case of a relevant difference. It doesn't strike me as terribly plausible because (arguably) phenomenal duplicates can have different first-person perspectives. Imagine Sally gets hooked up into an experience machine that will allow her to feel like she is the character Jesse in *Toy Story 2*. She has the phenomenal experience of performing every action Jesse does in *Toy Story 2*, and eventually is let out of the experience machine. When Sarah comes and plugs into the same experience machine and undergoes the same program of having the phenomenal experiences of Jesse in *Toy Story 2*, she doesn't *become* Sally because they had identical phenomenal experiences. Of course, one could press back against my argument. However, it's not my purpose to defend it here. Instead, I'm just registering one worry I have about this objection.

[17] Thanks to two anonymous reviewers for comments on this article. And thanks to Parker Settecase for discussions about these matters. Finally, thanks to G.L.G.—Colin Patrick Mitchell—for particularly insightful comments.

Hasker, W. (1999). *The Emergent Self*. Ithaca: Cornell University Press.

Hume, D. (1986). *A Treatise of Human Nature*. London: Penguin Classics.

Jackson, F. (1982). "Epiphenomenal Qualia" *Philosophical Quarterly*, 32, pp. 127–136. DOI: https://doi.org/10.2307/2960077

Jackson, F. (2003). "Mind and Illusion," in Anthony O'Hear (ed.) *Minds and Persons: Royal Institute of Philosophy Supplement 53*. Cambridge: Cambridge University Press: 251-271.

Kim, J. (1990). "Supervenience as a Philosophical Concept." *Metaphilosophy* 21, pp. 1-27. DOI: https://doi.org/10.1111/j.1467-9973.1990.tb00830.x

Kim, J. (2007). *Physicalism; or Something Near Enough.* Princeton: Princeton University Press.

Kirk, R. (2008). "The Inconceivability of Zombies." *Philosophical Studies* 139, pp. 73-89. DOI: https://doi.org/10.1007/s11098-007-9103-2.

Lewis, D. (1990). "What Experience Teaches" in William Lycan (ed.) *Mind and Cognition.* Hoboken: Wiley-Blackwell: 29-57.

Ludlow, P & Nagasawa Y., & Stoljar D. (2004). *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Cambridge: MIT Press.

Nagasawa, Y. (2008). *God and Phenomenal Consciousness*. Cambridge: Cambridge University Press.

Plantinga, A. (2006). "Against Materialism." *Faith and Philosophy* 23(1), pp. 3-32. DOI: https://doi.org/10.5840/faithphil20062316

Rudder Baker, L. (2013). *Naturalism and The First-Person Perspective*. New York: Oxford University Press.

Swinburne, R. (1997). *The Evolution of The Soul* (Second Edition). Oxford: Oxford University Press.

Swinburne, R. (2019). *Are We Bodies or Souls?* Oxford: Oxford University Press.

Tye, M. (2006). "Absent Qualia and the Mind-Body Problem," *Philosophical Review* 115, pp. 139–68. DOI: https://doi.org/10.1215/00318108-2005-013