Organon F 2021: 1–31
DOI (to be assigned)

ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

Robert Kirk's Attempted Intellectual Filicide: Are Phenomenal Zombies Hurt?

Dmytro Sepetyi*

Received: 6 June 2020 / Revised: March 30 / Accepted: 24 May 2021

Abstract: In the paper, I discuss Robert Kirk's attempt to refute the zombie argument against materialism by demonstrating, "in a way that is intuitively appealing as well as cogent", that the idea of phenomenal zombies involves incoherence. Kirk's argues that if one admits that a world of zombies z is conceivable, one should also admit the conceivability of a certain transformation from such a world to a world z* that satisfies a description D, and it is arguable that D is incoherent. From which, Kirk suggests, it follows that the idea of zombies is incoherent. I argue that Kirk's argument has several minor deficiencies and two major flaws. First, he takes for granted that cognitive mental states are physical (cognitive physicalism), although a zombist is free to—and would better—reject this view. Second, he confuses elements of different scenarios of transformation, none of which results in the incoherent description D.

Keywords: Consciousness; conceivability; incoherence; materialism; phenomenal zombie; possibility.

* Zaporizhzhia State Medical University
https://orcid.org/0000-0003-2110-3044
Maiakovskyi avenue, 26, Zaporizhzhia, Ukraine, 69035
dmitry.sepety@gmail.com

1. Introduction

In 1974, Robert Kirk introduced the concept of a phenomenal zombie—a creature physically exactly just like a conscious human being but without subjective experiences—and used this concept in an argument against physicalism. The argument, in brief, is that because zombies are conceptually or logically possible, phenomenal consciousness is something more than anything merely physical, so physicalism is false. When Kirk formulated the argument, it did not draw much attention. However, the argument was revived and made famous by David Chalmers in his philosophical bestseller *The Conscious Mind* (1996) and later papers. In the meantime, Kirk reversed his views and joined anti-zombists. In a series of publications, he argued that zombies are impossible in the relevant sense. He made the fullest exposition and defense of his argument in the book *Zombies and Consciousness* (2005), and recapitulated the argument in the paper "The inconceivability of zombies" (2008), and most recently in the book *Robots, Zombies and Us* (2017).

In *Zombies and Consciousness*, Kirk writes that one of the two main aims of the book is nothing less than "to dispose of the zombie idea once and for all". Not that there were no attempts to undermine the idea of zombies before. Kirk notes that "there are plenty objections to it in the literature, but they lack intuitive appeal". He believes his own attack on zombies fares better: "I have an argument which I think demolishes it [the zombie idea] in a way that is intuitively appealing as well as cogent" (Kirk 2005, vii).

In this article, I analyze Kirk's argument and show that it falls far short of the target. I will first outline the argument as presented in (Kirk 2005) and make some relevant elucidations. I will then note a few ways in which the formal representation of the argument needs clarification and more precise formulation (in accordance with Kirk's own explanations). Next, I will clarify the relationship between the idea of phenomenal zombies and interactionism. From this, I will proceed to the exposition of the failure of Kirk's argument. Finally, I will explain, in brief, why Kirk's later expositions of his argument (Kirk 2008; Kirk 2017) fare no better.

2. Kirk's argument in outline

As a preparation to his attack on zombies, Kirk argues, together with the supporters of the zombie argument (further on to be called "zombists"), that the possibility of zombies, in the sense that the idea of zombies involves no "inconsistency or other incoherence of a broadly logical or conceptual kind" (Kirk 2005, 10), is inconsistent with physicalism (materialism): if zombies are possible in this sense, then physicalism (materialism) is false. Kirk (2005) calls that kind of possibility "c-possibility"; Chalmers (1996; 1999; 2004) and later Kirk (2017, 75-92) call it "logical possibility". It should be distinguished from the possibility in a much more limiting sense, "natural" or "nomical" possibility—what is possible *given the laws of nature operant in the actual world*. Zombies may be naturally impossible (the laws of nature operant in the actual world ensure that whenever there are some physical states, there are some mental states) but c-possible, in the sense that there is no incoherence in the idea of zombies (or of a world inhabited by zombies instead of conscious human beings); the physical facts do not, on their own, entail there being consciousness (phenomenal mental states). If that is so, then consciousness is something extra, besides the physical.

If Kirk's arguments to this point are sound (and I think they are), then the c-possibility of zombies entails that materialism (physicalism) is false. If so, then to defeat the zombie argument, materialists should defeat the claim that zombies are c-possible, that is, that the idea of zombies is coherent. Because the claim has a strong intuitive appeal, the physicalist is invited to present the case to the contrary that is at least as (or, better yet, more) strongly appealing. Can she do this?

One helpful remedy Kirk recommends against the zombie idea is noting that it (at least, in its initial pure form) conflicts with another intuition that is at least as strong, or even stronger—the view that consciousness matters for our behavior. It seems absurd to think that our behavior, including what we talk and write, does not depend on our consciousness, so that consciousness could be subtracted and this would make no difference for what we do, including what we say and write. How could zombies, for example, talk about their phenomenal mental states, *qualia*, feels, if they had none? However, (this is not Kirk's point but mine) for a person who

would find the zombie argument persuasive apart from this conflict, the conflict seems to play for interactionist dualism rather than for materialism. On the one hand, the zombie possibility intuition (basically, the intuition that nothing physical entails anything subjective) rules out physicalism; on the other hand, the consciousness-matters-for-behavior intuition rules out epiphenomenalism; and both intuitions, if correctly understood, agree with interactionist dualism.

Kirk is aware that the zombie argument is not an argument for epiphenomenalism (to exclude all other possibilities) but an argument against physicalism (materialism) that leaves open the choice between epiphenomenalism, interactionism, panpsychism, and idealism. At least, it is so construed by the most eminent contemporary zombist, David Chalmers, who explained that the conclusion of the argument "is the disjunction of panprotopsychism, epiphenomenalism, and interactionism" (Chalmers 1999, 493). Kirk takes this into account. His argument is intended to be equally demolishing against all sorts of zombists, be they epiphenomenalists or interactionists or whoever.

The target of the argument is the claim that zombies are conceivable. If successful, the argument shows that although zombies seem conceivable, they are really inconceivable. It should be noted that the word "conceivability" is a red herring here; all that really matters is c-possibility, that is, the coherence of the idea. "Conceivability" pops up because David Chalmers construed the zombie argument as involving the subservient argument: (1) zombies are conceivable; (2) conceivability entails possibility; therefore, zombies are possible. Commenting on this argument, Kirk remarks that both of its premises are obscure because it is not clear how to understand "conceivable"; no clear meaning was ever specified in such a way that both premises were difficult to challenge. The general situation is that "the lower the threshold for conceivability, the easier it is to accept premiss (1)—but the harder it is to accept premiss (2)". Whatever the case, "to prove c-impossibility ... must be a good way to prove inconceivability" (Kirk 2005, 27). So, Kirk proceeds to prove that zombies are c-impossible in order to prove that they are inconceivable. However, why did he need to prove that zombies are inconceivable? Obviously, in order to block Chalmers' argument that since zombies are conceivable, they are c-possible. However, he

need not do it at all, if *he has already proven that zombies are c-impossible* (≡the notion of a zombie is incoherent). So, let us put aside "conceivability", and in further discussion, whenever Kirk uses the word, replace it with "c-possibility" or simply "possibility", or stipulate that "conceivability" means the same.[1]

Kirk's purported proof of the c-impossibility of zombies involves a story he calls "e-qualia story" and two claims about it:

> (C1) the c-possibility of zombies entails the coherence of the e-qualia story,
>
> (C2) the e-qualia story is incoherent.

From (C1) and (C2), it follows that that zombies are c-impossible (there is a hidden contradiction in the idea of zombies). The argument is valid. However, is it sound? Are both (C1) and (C2) true? Kirk meticulously argues that they are. I am going to agree, by and large (with some qualifications), with (C2) and show that Kirk's argument for (C1) fails.

The e-qualia story is a story of a (possible, or perhaps impossible) world that satisfies the following conditions:

> (E1) The world is partly physical, and its whole physical component is closed under causation: every physical effect has a physical cause. ...
>
> (E2) Human beings stand in some relation to a special kind of non-physical properties, e-qualia. E-qualia make it the case that human beings are phenomenally conscious.
>
> (E3) E-qualia are caused by physical processes but have no physical effects: they could be stripped off without disturbing the physical world.

[1] In recent personal email communication, Chalmers informed me that he now defines conceivability as not-apriori-not and has abandoned the use of the term "logical possibility". If so, "conceivability", in Chalmers' present use, is the same as Kirk's (2005) c-possibility and Kirk's (2017) logical possibility. (The treatment of the issue in (Chalmers 2002) and (Chalmers 2010) can be best interpreted along these lines.)

> (E4) Human beings consist of nothing but functioning bodies and their related e-qualia.
>
> (E5) Human beings are able to notice, attend to, think about, and compare their e-qualia. (Kirk 2005, 40)

A small terminological correction seems to be appropriate here. It would be more correct to talk in the e-qualia story not about "human beings" but, like in the zombie-story, about "the human-like inhabitants" of the e-qualia world, because if human beings in the real world do not satisfy the conditions (E2)-(E5), those e-qualia world inhabitants that satisfy them would perhaps not qualify as "human beings". However, this disqualification does not affect the argument. We may introduce the name "hubes" to designate both human beings and human-like inhabitants of possible (or even impossible) worlds that are exactly (or as far as possible exactly) like human beings physically, although they may essentially differ from human beings in other respects, which have to do with something non-physical (such as e-qualia). Now, having the term, let us replace in the e-qualia story "human beings" with "hubes".

To have convenient reminders for later use, let us designate the clauses (E3), (E4), and (E5) as "INERTNESS", "HUBES' COMPOSITION", and "EPISTEMIC CONTACT".

Kirk first argues for the claim (C2), that the e-qualia story is incoherent, and then for (C1), that if zombies are c-possible, then the e-qualia story should be coherent; if this argumentation succeeds, it follows that zombies are c-impossible. The argument for (C1) begins with a zombie story—a description (which should be coherent if zombies are c-possible) of a zombie world z that satisfies the following conditions:

> (A1) z is purely physical, causally closed system;
>
> (A2) Physically, z is as far as possible exactly like the actual world;
>
> (A3) The human-like inhabitants of z lack phenomenal consciousness. (Kirk 2005, 49)

To have a convenient reminder for later use, let us designate (A2) as "PHYSICAL IDENTITY".

Kirk proceeds to argue that *z* can be transformed—in a way that zombists should acknowledge to be c-possible, namely, by adding to it the non-physical factor that is supposedly responsible for consciousness in the actual world or, at least, one that is the same insofar as the phenomenology of non-physical mental states (e-qualia) is concerned[2]—into a *z** world that satisfies five conditions (Z1)-(Z5) that are equivalent to the conditions (E1)-(E5) of the e-qualia story. However, the preceding argument has established that the e-qualia story is incoherent; therefore z*-story is incoherent. However, z*-story is derived from the zombie-story and (if Kirk's arguments to the point are sound) a coherent transformation-story, from which it follows that the zombie-story is incoherent. Therefore, zombies are c-impossible, *q.e.d.*

Let us consider Kirk's argument in more details.

3. The e-qualia story: clarifications and reformulations

3.1. "E-qualia", qualia, and cognitive mental states

Apparently, the term "e-qualia" implies *qualia*—specific subjective qualities of mental states, their "what-it-is-likeness" for a mental subject (experiencer); the prefix "e-" probably is a shorthand for "epiphenomenal". Usually, when talking of *qualia*, philosophers mean *subjective qualities of experiences*, such as painfulness of pain (how it feels), or what it is like for an experiencer to have an experience of red color, or some other experience. Quite a few philosophers think that only qualia in this limited, experiences-bound sense are strongly resistant to materialistic (physicalist, or functionalist) reduction, whereas other aspects of mind, having to do with cognition and meaning, are far less problematic; therefore, we can assume that cognitive capacities (such as to notice, attend to, think about, and compare) can be fully accounted for in materialist (functionalist) terms, and focus on ex-

[2] In Kirk's own description, it should be "a non-physical item or items *x* which, when appropriately associated with z, would ensure that its inhabitants acquired our kind of phenomenal consciousness" (2005, 49).

periential qualia. (Further on, let us refer to this view as "cognitive physicalism".)[3] However, that is exactly what we *should not do* when discussing arguments against materialism, because this approach limits our choice to materialism and its most emasculated alternatives, and keeps out of discussion the more robust and defensible alternatives. To take this approach means to give materialism the uncontested lordship over the largest and most important part of the mental realm, and leave it for non-materialists to fight for the remaining poor grounds. Such a fight would be nearly hopeless, for in it, materialists would have on their side all the advantages, and their opponents all the disadvantages, of assuming that there is nothing to cognitive mental states besides physical processes that fulfill certain functions.

I think that the remark made by Howard Robinson in a paper defending the knowledge argument is relevant for the zombie argument as well:

> Those who ... think that physicalism can be correct for everything but qualia are in inconsistent position. The knowledge argument should not be cast in the form "physicalism can work for all other mental states but not for qualia", but in the form "even if it might look as if functionalism will work for less clearly introspectible states, such as thoughts, Mary's case shows that it will not work for qualia, and we can see from this that it does not

[3] See, for example, (Levine 2001, 4-6). In (Chalmers 1996), cognitive physicalism is not stated explicitly but seems to be presumed implicitly, in his distinction of psychological and phenomenal properties (where "psychological properties" are defined purely functionally, as ones that "play the right sort of causal role in the production of behaviour" (p. 11)), with putting awareness and other cognitive states on the "psychological" side, and especially in the thought experiments of fading and dancing qualia (pp. 253-274). However a year later, in reply to the criticisms advanced by Hodgson, Lowe, Velmans, and Libet, Chalmers repudiated cognitive physicalism and proposed that the use of cognitive terms in his earlier writings should be taken "in a stipulative sense" rather than as assuming that there is nothing more to cognitive states besides their behavioral functionality (Chalmers 1997, 20). However, I think that such a reading leaves his thought experiments of fading and dancing qualia deficient. It also may be relevant to note that (Chalmers 1997) and later is far more favorable to interactionism than (Chalmers 1996), and that the arguments of fading and dancing qualia presume epiphenomenalism.

> work for thought—at least, a certain category of thought...—either." (Robinson 2004, 72)

On the most natural, common-sense and strongly intuitively appealing view, our (conscious) thinking, understanding, and willing are intrinsically just as subjective as an experience of pain or of green color. The idea that all those physical processes that go on in human bodies and brains could (in the sense of c-possibility) occur without there being any subjective (conscious) awareness and understanding is just as plausible as the idea that C-fiber firing in the brain could occur without pain-sensation. The (*prima facie* coherent) concept of the phenomenal zombies implies that the zombies lack not only the capacities for such subjective experiences as pain-qualia or red-color-qualia but also the capacities to notice, attend to, think about, and compare in any (human) sense that involves subjective (conscious) awareness and understanding. At the very least, such a view is open (and, I suggest, commendable) for a zombist. A zombist would do well to posit *subjective* (conscious) cognitive states (processes) of thinking-awareness-understanding on the phenomenal, not the physical side. If zombies are c-possible, then the states of "noticing, attending to, thinking about, and comparing", in the sense relevant to the zombie argument, belong to the category of "a special kind of non-physical properties" that "make it the case that human beings are phenomenally conscious", that is, "e-qualia", on Kirk's definition (E2).

3.2. Where does the incoherence lie? Direct and indirect causation

Kirk argues that (E3), INERTNESS, is inconsistent with (E5), EPISTEMIC CONTACT. In fact, his argument goes through only if the beginning clause of (E3), "E-qualia are caused by physical processes", is understood as "*All* e-qualia are *directly* caused by physical processes *alone*". What Kirk really argues for is that if e-qualia have no effects whatever, whether physical or non-physical, then it is impossible for hubes to be able to notice, attend to, think about, and compare their e-qualia. There is no need to delve into details of Kirk's argument to this point. For my purposes, it is enough that the claim is *prima facie* very plausible: how can I attend to my

experiences (assumed to be non-physical e-qualia), or think about my experiences, if my experiences never cause, or play any causal role in causing my attention or thinking?

On the other hand, if we take (E3) literally, without the qualifications "all", "directly", "alone", then Kirk fails to make the case that the e-qualia story is incoherent. Kirk's argument for the incoherence of the e-qualia story (the inconsistency between INERTNESS and EPISTEMIC CONTACT) rests entirely on the absence of causal connection from experiences to attention and thinking. However, Kirk's formulation of the story (especially, with respect to INERTNESS) does not exclude the possibility that hubes' non-physical experiences are causally relevant to their attention and thinking, *if attention and thinking are taken to be non-physical* mental states (even if they are epiphenomenal, having no physical effects).

So, the dilemma arises:

> either Kirk's argument fails to show that the e-qualia story is incoherent,
>
> or the formulation of the e-qualia story should be made more precise so as to exclude *any* possibility of there being a causal link from experiences to attention and thinking, whether the latter are taken to be physical or non-physical.

If the former, then the argument fails full stop. If the latter, the argument can proceed with the e-story slightly reformulated, by inserting into (E3) the qualifiers "all", "directly", "alone":

> (E3*) INERTNESS* *All* e-qualia are *directly* caused by physical processes *alone* but have no physical effects: they could be stripped off without disturbing the physical world.

It may be objected here that in fact, Kirk (2005, 42) *does argue* that the e-qualia story forbids e-qualia to have effect on other e-qualia. However, if you consider the argument, you easily see that it is made on the construal of (E3) as (E3*). The argument is that "since by (E3) all qualia are already caused to occur by physical events", there would be no work for e-qualia-to-e-qualia causation to do. Now three points should be noted.

1) In fact, the qualifier "all" is absent in the formulation of (E3) on p. 40, but it is clear that because there is no other qualifier (such as "some"), "all" is implied. And on p. 42 Kirk confirms this explicitly. So my adding "all" does not really change (E3) but merely emphasizes its point.

2) There is no causal work for e-qualia to do only if all e-qualia are caused by physical processes *alone*, in the sense that physical processes *are sufficient* for causing these e-qualia—to be distinguished from the possibility that some e-qualia are produced jointly by physical processes and e-qualia, so that without the participation of e-qualia physical processes would not have this effect. So, adding the qualifier "alone" (understood in this sense) is perfectly justified.

3) There is no causal work for e-qualia to do only if all e-qualia are caused by physical processes *directly*,—to be distinguished from indirect causation, when a physical process P causes an e-qualia A that, in its turn, causes an e-qualia B. If P causes A that causes B, then the fact that P causes B (by causing A that causes B) in no way robs A of its causal work. So, if anything, the argument on p. 42 shows that *in Kirk's own meaning, (E3) should be construed as (E3*).*

3.3. Robinson's objection

Howard Robinson (2016, 55) proposed that a zombist can deny the incoherence of the e-qualia story by 1) making use of the typical functionalist account of intentionality, according to which the intentionality of an epistemic state is a matter of behavioral appropriateness, and 2) holding that this behavioral appropriateness need not necessarily be due to an epistemic state's being caused by its object (experience, in our case) but can be due to common causal ancestry of both the epistemic state and its object. I think Robinson is right that Kirk does not provide an argument to neutralize such a move. However, for me, personally, it seems highly plausible that for an epistemic state to be about a particular real object (at least, in the sense of original intentionality), there must be causal link from the latter to the former. So although the move proposed by Robinson is available for a zombist, I propose to explore the availability of other resources.

For convenience of the following discussion, it is useful to introduce a distinction between three possible varieties of "zombists" that would treat

Kirk's argument differently. Let us designate a zombist who is *not a cognitive physicalist* "a Cartesian zombist" (because he/she, like Descartes, holds that thinking pertains to a non-physical mind rather than to a physical body), and a zombist who is a cognitive physicalist—"a non-Cartesian zombist". With non-Cartesian zombists, the way to meet Kirk's argument depends on whether a zombist is an epiphenomenalist or an interactionist. A non-Cartesian epiphenomenalist in fact holds that the actual world satisfies the e-qualia story, and so he/she can meet Kirk's argument only by denying that the e-qualia story is incoherent (probably, in the way Robinson suggests). With respect to such a zombist, the rest of Kirk's argument has nothing to do. So the following discussion is concerned only with the coherence of Cartesian zombism and non-Cartesian interactionist zombism.

4. The zombie story and interactionism

In the zombie story (A1)-(A3), the condition (A2), PHYSICAL IDENTITY, stipulates that the zombie world to be discussed (z) is physically "as far as possible exactly like the actual world". The phrase "as far as possible" needs an explanation: why did Kirk moderate his zombie story with it? The purpose was to take into account Chalmers' explanation that the zombie argument is not an argument for epiphenomenalism but leaves open several non-materialistic alternatives, such as panprotopsychism, epiphenomenalism, and interactionism. If so, for Kirk's argument to bite against zombism generally (not only against epiphenomenalist zombism), the description of a zombie world *z* should be such that any zombist (interactionist as well as epiphenomenalist) should admit the c-possibility of such a world.

How can non-epiphenomenalist views be reconciled with the possibility of zombies? In particular, how can it be with interactionist dualism? *Prima facie*, it seems that interactionism is inconsistent with the c-possibility of zombies. It seems that if in the actual world, non-physical consciousness causally influences brain states responsible for behavior (as interactionists believe), phenomenal zombies and zombie-worlds as usually described are c-impossible for an obvious reason: zombies lack some *physically relevant* (although non-physical) causal factor that we have, and so the physical dynamics of their bodies' functioning should be different. However, David

Chalmers explained how "the possibility of zombies is compatible with non-epiphenomenalist dualism": "an interactionist dualist can accept the possibility of zombies, by accepting the possibility of physically identical worlds in which physical causal gaps (those filled in the actual world by mental processes) go unfilled, or are filled by something other than mental processes" (Chalmers 2004, 182-183).

4.1. Replaceabilism

However, "the possibility of physically identical worlds in which physical causal gaps (those filled in the actual world by mental processes) go unfilled, or are filled by something other than mental processes" is likely to seem problematic, at least *prima facie*. One may think that the idea of a world in which some physically relevant causes are systematically lacking but all physical events go as if nothing were lacking is incoherent; such a world is not c-possible. Otherwise, if in a possible world, physical causal gaps are filled by something other than mental processes, what can this "something other" be? It seems that if it is not mental and is causally relevant, there is no reason why it should not count as physical. However, if it counts as physical, then the possible world so conceived is not exactly physically identical with the actual world; it has some physical surplus. On the other hand, we can run the zombie argument with a modification that takes care of such a physical surplus: if zombies with a physical surplus are c-possible, it seems that materialism should be false, because those zombies lack nothing physical that human beings have but lack consciousness. (It is implausible that adding some physical surplus would bereave human beings of consciousness and turn them into zombies). Although Kirk did not go in these details, he made his description of the zombie world z in such a way that it could accommodate such zombies with a physical surplus (and so make it possible for some interactionists to count as zombists), by means of the phrase "as far as possible" in the condition (A2).

An interactionist zombist view that is so accommodated can be designated as *replaceabilism.* A replaceabilist admits the possibility of zombies with a moderate modification to the initial (Kirk 1974a; Kirk 1974b; Chalmers 1996) specification. Replaceabilist interactionism is consistent with, and implies, the c-possibility of *modified* phenomenal zombies or a

modified zombie world that *lacks nothing physical* that we (or the actual world) have but lacks consciousness nevertheless. There just should be, in those modified zombie worlds, some other *physically relevant* causal factors to compensate for the causal deficiency resultant from the subtraction of human consciousness. By the condition (A1), which says that a zombie world z is purely physical, Kirk stipulates that these factors should themselves be physical (so, a zombie world may be physically richer than the actual world).

As an alternative to filling the causal gap (that should—if interactionism is true—result from subtracting mental processes) with some additional physical factors, we can conceive of some possibility like the following. Imagine a world that runs in parallel with ours and is at every moment exactly like our world a minute ago in all physical respects, because this world is governed by a physically omniscient and omnipotent demon who took fancy to support that belated-copy-world so that all physical deviations (that may happen because the humanlike inhabitants of that world have no consciousness, or because of quantum-mechanic indeterminacy) are almost instantly detected and eliminated by the demon. Although in such a conceivable scenario, some mental processes (those of human beings, indirectly, and those of the demon, directly) are causally efficient with respect to physical events in the zombie world, the zombies themselves are purely physical copies of human beings without phenomenal mental states, so they fit the requirements of the zombie argument.

Besides replaceabilism, an interactionist dualist has two other options, which I designate as *irreplaceabilism* and *supercoincidentalism*. Let us consider these alternatives and their relationship with the c-possibility of zombies.

4.2. Irreplaceabilism and the conditional construal of the zombie argument

An interactionist can deny the c-possibility of replacing human phenomenal minds with some physical entities so that all physical events proceed without any change.[4] Kirk mentions such a possibility and remarks that

[4] It is open for an interactionist—at least, if he is a substance dualist—to take the view that the human mind, or self, or soul develops and affects the brain in such a

"some interactionists might deny that physical events could cause human-like behavior, but they could not be zombists" (Kirk 2008, 85). This should be admitted, given Kirk's definition of "zombists" as "those who think zombies are conceivable" (Kirk 2005, 38), where "conceivable"="c-possible". However, such an interactionist—zombist or not—can still find use for the zombie idea and the zombie argument in a conditional way, as suggested by Andrew Bailey, "as part of a destructive dilemma for the physicalist": either physical reality is *not* causally closed, and so physicalism is false, or it is causally closed, and then zombies are possible, and so physicalism is false anyway (Bailey 2009, 135).[5] If so, then Kirk's argument falls short of

way that it is in principle (as a matter of c-possibility) irreplaceable—not with respect to some particular effect but with respect to *all the totality* of its *real and possible* physical effects *throughout the life*—with anything physical.

Is irreplaceabilism plausible? I think that it is. To see this, let us first think of our talks, and writings, and philosophical discussions about our experiences and other conscious states and processes (such as having a certain occurrent thought). It seems very implausible that all the physical aspect of all these happenings could be effected by zombies without any experiences and other conscious states and processes, with some purely physical substitute. It is far from clear (and, I think, implausible) that a purely physical substitute for consciousness capable of such an achievement is possible, even in principle (as a matter of c-possibility). And this becomes even more so, if we think of such persons as Plato, or Einstein, and their intellectual achievements, and the impacts of those achievements on the course of human history, behaviors of millions of people, etc. Presumably, their intellectual achievements were a matter of conscious interest to some problems, conscious understanding, conscious thinking, and conscious guess. Presumably, their huge impact on the human history, on behaviors of millions of people, was a matter of other people's conscious interests and understanding, etc. Is it possible, even in principle, (c-possible) that in a modified zombie-world, its humanlike inhabitants-zombies would do all the same movements, with all the same (speechlike) sounds produced, books written and typed, machines and computers produced and run, as a result of nothing but purely physical interactions of the microphysical components of which their bodies consist, with no (phenomenal) consciousness at all? Perhaps it is, but it is at least just as plausible that it is not.

[5] As Kirk himself remarks, "the idea of zombies suggests itself as soon as one accepts the causal closure of the physical" (Kirk 2008, 74). Thus, an irreplaceabilist can consider the zombie argument as showing not what *is* logically possible (given

his most ambitious purpose of "disposing of the zombie idea once and for all", or demolishing it (Kirk 2005, vii), even if it were successful in all other respects (which it is not, as will be shown in the following sections), that is, against all those who fall under his definition of "zombists".

4.3. Supercoincidentalism

Alternatively, the interactionist zombist can hold that even a non-modified zombie world (with no physical entities added and no non-physical factors involved) is possible, but such a possibility involves a superhugely improbable succession of coincidences (infinitely more improbable than the chance that a tornado sweeping through a junkyard might assemble a Boeing 747). The point is that if the idea of a genuine *physical* causal indeterminacy is not incoherent (and quantum mechanics seems to show that it is not merely coherent but holds in the actual world), and if consciousness has physical effects in the actual world, then it is not strictly impossible that there may be an exact physical duplicate of the actual world in which there is no consciousness: in that world, all physical events turn out to be the same as in the actual world as a result of a superhugely improbable—but not strictly impossible—quantum-mechanical flukes.

Take note: it is not the case that quantum mechanical indeterminacy applies only to microphysical but not to macrophysical events. It applies throughout the board, only that for macrophysical events, the probability of a considerable deviation from the "normal" deterministic course is hugely small. A zombie world is a world in which such hugely improbable events regularly happen with zombies, and incidentally they happen in such a way that all parts and particles of zombies make exactly the same movements as the corresponding parts and particles of human bodies in the actual world.

Supercoincidentalism has a considerable advantage over the other two interactionist options, in that (1) it accommodates the c-possibility of zombies in the most direct way (which requires no modification to the initial

that the world is such as it is, that is, interactionistic) but what *should be* logically possible on the assumption that the actual world is causally closed with respect to the physical events.

specification of zombies) and (2) it saves the irreplaceabilist intuition that it is superhugely unlikely that purely physical twins of human beings with no phenomenal minds would behave in all exactly the same ways as conscious human beings do throughout whole human lives.[6]

Although the supercoincidentalist option is distinct from the replaceabilist one, the argument that follows fits both in the same way.

5. The transformation story, and where Kirk's argument fails

Zombists hold that a zombie world z described by the conditions (A1)-(A3) is c-possible. Should a zombist agree with Kirk that adding to z the non-physical factor (from now on to be called "the consciousness factor") that is supposedly responsible for consciousness in our world, or its "phenomenal duplicate" (perhaps, bereaved of powers to produce physical effects), can conceivably transform that world into the world z^* identical to the e-qualia story world?

Consider Kirk's description of z*:

> (Z1) z^* is partly physical, and its whole physical component is closed under causation: every physical effect in z^* has a physical cause.
>
> (Z2) The human-like organisms in z^* are related to a special kind of non-physical item x. x makes it the case that they are phenomenally conscious.
>
> (Z3) x is caused by physical processes but has no physical effects: it could be stripped off without disturbing the physical component of z^*.

[6] Consider the infinite set of possible worlds that at some moment $\boldsymbol{t}$ are exact physical copies of the actual world at this moment but in which there are no phenomenal minds. In that set, the subset of worlds in which all physical events with zombies will for a considerable time proceed exactly as they do in the actual world with human beings makes up an infinitely small—going to zero—portion. The probability of hitting at random at such a world in such a set goes to zero. Nevertheless such zombie worlds are not strictly impossible.

> (Z4) The human-like inhabitants of z* consist of nothing but functioning bodies and their related x. [...]
>
> (Z5) The human-like inhabitants of z* are able to notice, attend to, think about, and compare the qualities of their experiences. (Kirk 2005, 51)

In the description, "x" stands for the consciousness factor.

For our discussion, there are two important questions about this description to be asked and answered:

> – Why does Kirk think that a zombist is committed to the c-possibility of the transformation from z to z^*?
>
> – Is (Z1)-(Z5) really equivalent to the e-qualia story?

5.1. The epistemic intimacy argument and its failure

Why does Kirk think that a zombist is committed to the c-possibility of the transformation from z to z^*? In brief, his reasons are as follows.

If one admits that z is possible, then one cannot deny that z^+ is possible, where z^+=z+the consciousness factor, where the consciousness factor is phenomenally just like human consciousness, has the same dependence on human brains, and *has no physical effects*. This ensures that in z^+, (Z1), (Z2), (Z3), (Z4) hold (Kirk 2005, 49-50).

Kirk argues that zombists should admit the c-possibility that (Z5) holds as well: because z^+ is exactly like the actual world *physically*, and it has the consciousness factor x that is phenomenally exactly like "the non-physical item y which they think produces phenomenal consciousness in the actual world", and given that we have *epistemic intimacy* with our experiences, there is nothing to account why the hubes of z^+ cannot c-possibly have such epistemic intimacy with their experiences (Kirk 2005, 50-51). I suggest that Kirk could strengthen his argument by pointing out that because z^+ is physically identical with the actual world, its hubes would talk and write about their experiences just like we do, and this is impossible if they are unable to notice, attend to, think about their experiences. Let us designate this argument as *the epistemic intimacy argument*.

If *the epistemic intimacy argument* succeeds, then a zombist should admit the c-possibility of z^*.

It should be noted that Kirk's argument is made *on the assumption of cognitive physicalism*: it is stipulated that the consciousness factor in z^* has no causal impact on the physical but is not stipulated that there are no causal connections *within* the consciousness factor. A bit later, I will argue that a zombist who is a cognitive physicalist (a non-Cartesian interactionist zombist) can plausibly decline the epistemic intimacy argument, and so Kirk's purported refutation of zombism fails. However, I will first explore how a Cartesian zombist (who is *not a cognitive physicalist*) can respond Kirk's argument.

A Cartesian zombist would not need to resist the epistemic intimacy argument, insofar as the latter stipulates that the consciousness factor x in z^+ has no *physical* effects and does not stipulate that there is no causation within x itself. Insofar as such intrinsic causation within the consciousness factor x is not ruled out, a Cartesian zombist would agree that a world z^* that satisfies (Z1)-(Z5) is c-possible. You just add something like a Cartesian soul (having cognitive mental states as well as experiences) to z and deprive it of all powers to produce physical effects. The resulting world would satisfy (Z1)-(Z5); however, there is no incoherence involved. This is possible because (Z1)-(Z5), although very much like the e-qualia story, is not exactly the same. (Z3) is not quite the same as (E3), and (Z4) is not quite the same as (E4). However, the dialectics of the argument will come to the point where a Cartesian zombist can be confronted with *the modified epistemic intimacy argument* (and the modified clause ($Z3^m$)) that involves the stipulation that nothing in the consciousness factor x has *any* effects, *physical or nonphysical*. (Note that ($Z3^m$) indeed would be equivalent to (E3*), INERTNESS* in the e-qualia story.)

So, a zombist who admits the incoherence of the e-qualia story should find something wrong with the *epistemic intimacy argument*, either initial or modified or both.

Happily for a zombist, there is a simple explanation as to what is wrong with these intimacy arguments. It is as follows.

Although there are no behavioral (and generally physical) and no phenomenal differences between z^+ and the actual world, there still can be *some*

differences that make it the case that (Z5) cannot hold in z^+. What other differences can there be, given that both the actual world and z^+ contain nothing but physical entities and consciousness? The answer is that there is an important *difference in causal relations*: presumably, there is causation from experiences to cognitive mental states in the actual world; on the other hand, Kirk's argument hangs on the stipulation that there is no such causation in z^*. A zombist can hold that such an absence is inconsistent with (Z5)—at least, if the e-qualia story is indeed incoherent.

Recollect that the e-qualia story was found incoherent exactly because arguably, in the absence of causal connection from experiences to cognitive mental states, there can be no cognitive mental states *about experiences*: the very absence of such causal relations rules out the existence of cognitive mental states *with such aboutness*. Surprisingly, Kirk fails to see that this applies to the c-possible result of adding the consciousness factor x to the zombie world z: if there is no causation from experiences to cognitive mental states, then (Z5) does not hold.

One can wonder: how can that be if z^+ is exactly like the actual world both physically (in particular, in behavior of its hubes) and phenomenally? The answer is that although there will be something it z^+ that is physically and/or phenomenally exactly like cognitive mental states about experiences in the actual world, that something *would not qualify as* cognitive mental states *about experiences*, because—at least, in cases when the referent is a particular really existent object—*the appropriate causal relationship is constitutive of aboutness* (at least, partially).

Again, that is exactly why in the e-qualia story (E3*), INERTNESS* seems to conflict with (E5), EPISTEMIC CONTACT. At least, in Kirk's argument for the incoherence of the e-qualia story, there is nothing to show that (E3*), INERTNESS* is inconsistent with there being physical states (including all behavioral movements) that are *physically exactly like* what a cognitive physicalist can take for cognitive mental states about experiences, or with there being some e-qualia that are *phenomenally exactly like* what a cognitive non-physicalist would take for occurrent cognitive mental states about experiences.

In fact, it is not too difficult to see how there can be two mental states that *are phenomenally identical but differ in their aboutness*: one is about

a particular really existing thing, whereas the other is not. Take, for example, seeing a table and hallucinating a table. They can (c-possibly) be phenomenally the same, but the former is about a particular really existing table, whereas the latter is not. And this would be the case even if there really is a table in place where the hallucination suggests but that table has nothing causally to do with the hallucination. The same applies to cognitive states about particular real experiences and their c-possible phenomenal twins which *fail to be about particular real experiences*. Think of the fantastic Swampman-style scenario in which my physical duplicate gets assembled out of atoms. Suppose that I had a toothache yesterday, and I can well recollect that experience. Surely Kirk, as a materialist, should admit that because we (and our brains) are physically exactly the same, my duplicate can "recollect" having a toothache yesterday, and this his "recollection" can be physically and phenomenally exactly like my recollection. However, my recollection is genuinely about a particular real experience I had yesterday, whereas—as Kirk himself argued—my duplicate's "recollection" cannot be genuinely about that (or any other real) experience, because there is no causal link from the experience to his "recollection". And although my duplicate can behave (move) exactly like I do when I talk or write about my past experiences, making just the same noises and leaving just the same marks on paper, this his behavior will not be talk and writing about his past experiences.

This crucial point made, in the rest of this section, I propose a detailed account of how the defense of a Cartesian zombist would proceed *before it arrives at the point of collision with the modified epistemic intimacy argument* (subsections 5.2-5.3), and an account of how a non-Cartesian interactionist should coherently envision the $z \rightarrow z^{+}$ transformation scenarios, none of which happen to result in z^{*} that satisfies (Z1)-(Z5) (subsection 5.4).

However, one of Kirk's latter expositions of his argument, (Kirk 2008), gives this discussion a new turn, to be discussed in section 6.

5.2. Broadening the e-qualia story

One objection that a zombist can make to Kirk's argument is that (Z4) and (E4), HUBES' COMPOSITION, are not equivalent. Compare

(Z4) Hubes of z^* consist of nothing but functioning bodies and their related consciousness factor x.

and

(E4) Hubes consist of nothing but functioning bodies and their related e-qualia.

A zombist is not committed to the view that the consciousness factor x is nothing but e-qualia—nor even to the c-possibility of such a consciousness factor. On Kirk's definition, e-qualia are non-physical properties; however, the description of z^* leaves it open that the conscious factor x may be more than that. Some zombists would prefer the substance dualism view that a human beings consist of nothing but a functioning body and a *non-physical mental subject* that is a bearer of phenomenal mental states, or qualia. And they can hold that *it is not even conceivable for there to be e-qualia without nonphysical mental subjects that underlie them.* So, even if Kirk's claim that the c-possibility of a zombie entails the coherence of the description of z^* is right, this is not enough to refute zombism, because the description of z^* is not equivalent to the e-qualia story, at least insofar as (E4) and (Z4) are concerned.

However, Kirk anticipated this objection and dealt with it by claiming that the difference is not significant because "even if x were kind of unitary substrate rather than the collection of properties, it would have to underlie, realize, or otherwise provide for a plurality of properties", including phenomenal qualities, and "so far as the e-qualia story is concerned, therefore, those qualities might just as well be called 'e-qualia'" (Kirk 2005, 51-52).

The point of this reply is that the e-qualia story would still remain incoherent, if we supplement its ontology (which initially included only physical entities and e-qualia) with non-physical unitary substrates that (do nothing but) underlie e-qualia. Only in that case, the difference between (Z4) and (E4) is insignificant. (Surely, if adding such unitary substrates to the e-qualia story would make it coherent, then the difference would be very significant!)

If so, Kirk's reply amounts to the correction in the initial e-qualia story that can be made explicit by replacing (E4) with

> (E4*) Hubes consist of nothing but functioning bodies and their related e-qualia, *or non-physical substances that are bearers of their related e-qualia.*

Given that e-qualia and a non-physical mental subject are the only candidates for the role of the consciousness factor x, with this modification to the e-qualia story (further on taken for granted), we have the required equivalence between (Z4) and (E4*). However, this does not save Kirk's argument.

5.3. Intra-mental causation and the defense of Cartesian zombism

There is a more important difference between (Z1)-(Z5) and the e-qualia story—the non-equivalence of (Z3) and (E3*), INERTNESS*. Compare:

> (Z3) x is caused by physical processes but has no physical effects: it could be stripped off without disturbing the physical component of z^* (Kirk 2005, 51),

where x stands for the consciousness factor,
and

> (E3*) All e-qualia are directly caused by physical processes alone but have no physical effects: they could be stripped off without disturbing the physical world.

(Z3) is *not equivalent* to (E3*) in an important respect. To see this, we should recollect that the e-qualia story was found incoherent because (E3*), INERTNESS* conflicts with (E5), EPISTEMIC CONTACT, and they were found inconsistent because INERTNESS* forbids non-physical mental events (such as pains or red-color qualia) to cause or take part in the causation of any cognitive mental states (such as conscious attention, thinking, etc.). However, a Cartesian zombist would find nothing in (Z3) to this effect. Nothing in the formulation of (Z3) forbids causation *within* the consciousness factor x. It is relevant to this point that the consciousness factor does not need to be a simple property (e-qualia). A zombist can hold (in line with a variety of property dualism) that the consciousness factor is a *set of causally connected non-physical mental states* (conscious experiences causing conscious awareness, attention, thought, etc.). Or she can hold that

the consciousness factor is an entity-substance with rich internal differentiation, temporal development, and internal causal relationship between its states—it may be a full-blown mental subject, or self, or the Cartesian soul. If so, there is no incoherence between (Z3) and (Z5) analogous to the incoherence between (E3*) and (E5). So, the *z** story is not equivalent to the e-qualia story in a crucial way that makes Kirk's argument invalid.

Kirk was not ignorant of this kind of objection. He considered essentially the same objection (although formulated in different terms) and replied as follows:

> The question is not whether some metaphysical story or other could be told by which non-physical items were capable of cognitive processing. It is whether the conceivability of zombies entails the conceivability of the e-qualia story. Hence all I have to do is to show that if zombies are conceivable, then so is a version of (Z1)-(Z5) according to which *x* is inert. And this is a consequence of the fact that causation is a contingent matter. (Kirk 2005, 52)

Now, a zombist can object that this reply misses the mark. Kirk *did not show* that if zombies are c-possible, then so is the world *z*** describable by (Z1)-(Z5) *plus the additional condition* that the consciousness factor is inert in the strong sense required for his purpose—that besides having no external, physical effects, it also has no internal causal links between its own states, such as experiences and thinking about experiences.

Kirk did not foresee and answer this objection; however, at this point he could appeal to the epistemic intimacy argument appropriately modified (adapted to the context of cognitive non-physicalism). However, a Cartesian zombist can decline this argument in the way explained in the subsection 5.1. So Kirk fails to prove that Cartesian zombism is incoherent.

5.4. Causal overdetermination and the defense of non-Cartesian interactionist zombism

On the other hand, a non-Cartesian interactionist zombist (one who accepts cognitive physicalism) would consider two possibilities of envisaging the transformation from the zombie world *z* to the world *z**:

> – either we add to *z* the same consciousness factor as that of the actual world, including its causal powers;
>
> – or we add to *z* such a consciousness factor that is exactly like that of the actual world insofar as the production and phenomenology of its non-physical states (e-qualia) is concerned but is bereft of its causal powers to produce physical effects.

1) The case with the same consciousness factor, including its causal powers
In the first case, a zombist can point out that the resulting world cannot c-possibly fit (Z3), which says that the consciousness factor has no physical effects. The trouble with (Z3) is as follows.

In z*, physical factors alone have all the causal powers required to produce all the effects which the consciousness factor produces in the actual world. And in z*, the consciousness factor alone has all the causal powers required to produce all the effects it produces in the actual world. However, in z*, causal powers of the physical are not alone, and causal powers of the consciousness factor are not alone; they are put together. It seems that this should result in different (additive) effects, as compared with those that would result if only one of them acted. (1+1 equals 2 rather than 1.) However, if the effects are different, then the consciousness factor is causally efficient, and (Z3) does not hold for z*.

However, Kirk could insist that it is conceivable (c-possible) that in z*, causal powers of the consciousness factor make no physical difference: the causal powers of the physical and the causal powers of the consciousness factor together produce exactly the same effect that each of them would produce alone. (In z*, causal 1 of the physical + causal 1 of consciousness equals causal 1, not causal 2.) A zombist can concede this, but point out that this would clearly be a case of causal overdetermination. If so, we still do not have (Z3); instead, we have, as the best c-possible approximation to (Z3)

> (Z3*) x is caused by physical processes but has no non-overdetermined physical effects: it could be stripped off without disturbing the physical component of z*.

So, the envisaged world z* is not a world in which the consciousness factor has no physical effects but a world in which its physical effects are systematically overdetermined by physical factors. In this situation of overdetermined causation, the consciousness factor makes no physical difference; however, overdetermined causal links from the consciousness factor to physical states of the brain are still causal links, and there being such causal links may be a sufficient ground for some such brain states (or their functional aspect) to count as noticing, attending to, thinking about experiences, etc.[7]

2) The case with the consciousness factor's causal powers subtracted

In the second case, a zombist can point out that the resulting world does not fit (Z5), which says that the hubes "are able to notice, attend to, think about, and compare the qualities of their experiences". In the resulting world, there would be quasi-cognitive states that are physically (and so functionally) exactly like noticing, attending to, thinking about experiences, etc. in the actual world; however, they should not count as genuine noticing, attending to, thinking about experiences, etc., exactly because they do not stand in the appropriate causal relationship to experiences. (That is the case because the epistemic intimacy argument is mistaken, as was shown in the subsection 5.1.)

The remaining description (Z1)-(Z4), without (Z5), is crucially non-equivalent to the e-qualia story, because in the latter, the contradiction arises between (E1)-(E4*) on one side and (E5), EPISTEMIC CONTACT on the other. So the description of the world (Z1)-(Z4), unlike the (incoherent) e-qualia story, is perfectly coherent. And so Kirk fails to prove that non-Cartesian interactionist zombism is incoherent.

[7] A reminder may be appropriate that the interactionist at issue does not hold that there is such overdetermination in the actual world; he just holds that a world with such overdetermination is c-possible, and that in such a possible world, overdetermined causation from experiences to some physical brain states should count as sufficient for there to be *epistemic contact* from experiences to cognitive states.

6. Kirk's later expositions of his argument. The necessity of epistemic contact with experiences: why Kirk would better not appeal to it

In his later paper, "The inconceivability of zombies" (2008), and again in the chapter 7 of his book *Robots, Zombies and Us* (2017), Kirk rehashes his argument in a bit different and less detailed way, with the same unquestioned implicit assumption of cognitive physicalism.

There are several differences to be pointed out.

1) Both (Kirk 2008) and (Kirk 2017) omit the epistemic intimacy argument altogether. They just take it for granted that adding to z an inert consciousness factor leaves the hubes of z* in epistemic contact with their experiences.

2) Nevertheless, (Kirk 2008) considers the possible objection on the side of an interactionist zombist that if the inert consciousness factor "continued to make our successors conscious, its lack of causal efficacy would prevent it from continuing to sustain epistemic contact" (Kirk 2008, p. 86), and makes a new argument against it.

3) In (Kirk 2017), the former e-qualia story goes under the name "epiphenomenalism".

Of these, only the second point can be taken as strengthening Kirk's position, so let us discuss it.

Kirk begins with the remark that he "find[s] it hard to make sense of that suggestion" (Kirk 2008, p. 86); then he quotes David Chalmers' statement that there is "not even a conceptual possibility" that a subject should have an experience "without any epistemic contact with it" (Chalmers 1996, p. 197), states his approval ("surely he is right about that") and adds some more comments to support the claim that "being in epistemic contact with one's conscious experiences is part of what it is to have them" (Kirk 2008, 87).[8]

[8] This argument is mentioned already in (Kirk 2005, 50); however, there Kirk is more cautious about it and does not make it part of his argument: "although Chalmers's assumption is plausible, it is not needed for this argument" (Kirk 2005, 50). Instead, Kirk relies on the epistemic intimacy argument. In (Kirk 2008) things are reversed: Kirk omits the epistemic intimacy argument and relies on the argument from the necessity of epistemic contact with experiences.

So far so good. However, we need to explore the consequences of the supposed necessity of epistemic contact for Kirk's argument from its very start. Now I am going to argue that it blocks Kirk's argument *with respect to those cognitive mental states that stand in that necessary relation with experiences* already on its first stage (the argument for the incoherence of the e-qualia story).

Suppose that indeed it is conceptually impossible for there to be an experience and no cognitive state having that experience as its object. For simplicity sake, suppose that it is conceptually impossible for there to be an experience and no awareness of that experience. In that case, the relation between the experience and the awareness of this experience *is not that of causation* (causal links are contingent; they can always c-possibly be severed) but some special, *sui generis*, relation. Let us dub this relation as "superintegration". If so, there is no causation from the experience to its awareness but there is the awareness of the experience. Kirk's argument for the incoherence of the e-qualia story fails because it just does not take into account that there can be superintegration rather that causation between a non-physical experience and the awareness of that experience. In this case, an experience and the awareness of it are, in a sense, not two really distinct causally connected states, but two inseparable aspects of the same state (so that their separation is not even conceptually possible). Perhaps it is something like sides and angles of a polygon: although sides are not angles, it is even conceptually impossible for there to be the former without the latter, and it would make no sense to say that sides cause angles or *vice versa*.

Two things should be noted about this refutation of Kirk's argument for the incoherence of the e-qualia story.

First, it is available only for a zombist who admits that those cognitive mental states that are superintegrated with experiences are non-physical. That is, a zombist should be, in our terms, "Cartesian" *at least with respect to some cognitive mental states* (such as my present awareness of my present pain).

As for a *thoroughly* non-Cartesian zombist—that is, one who holds that all cognitive mental states (including such as my present awareness of my present pain) are physical—such a zombism is clearly incompatible with superintegration: if experiences are non-physical but my awareness of my

experiences is physical, then the former and the latter are distinct and cannot be superintegrated. This can be used as an argument against the view that combines dualism with thorough cognitive physicalism. However, note that this argument is entirely independent from Kirk's anti-zombist argument; if it undermines the mentioned variety of dualism, it makes it on its own, and Kirk's anti-zombist argument does no job here. (Note that this outcome is just what Howard Robinson says in the remark quoted in subsection 3.1.)

On the other hand, as far as other varieties of dualism (wholly or partially "Cartesian"—those that admit that *at least some* of our cognitive mental states are non-physical) are concerned, the acceptance of the claim about superintegration invalidates Kirk's argument for the incoherence of the e-qualia story: if it is not causal link but superintegration that makes us aware of our experiences, and if this awareness is indeed not a distinct mental state but an aspect of experiences that cannot be c-possibly severed from them, then there is no contradiction between (E3*) that says that experiences (e-qualia) are causally inert and (E5) that says that there is an epistemic contact with experiences—the epistemic contact is inbuilt in experiences themselves. [9]

The result is that far from saving Kirk's argument, the acceptance of the claim about superintegration blocks it at the first stage; at the same time, it undermines the view that combines dualism with thorough cognitive physicalism.

Second, the claim about superintegration is plausible only for the cases when an experience and a cognitive state directed at that experience are simultaneous. It may well be the case with my present-moment awareness

[9] At this point, the objection can be tried that the e-qualia story assumes cognitive physicalism, and of course, a Cartesian dualist should admit that *on that assumption*, the e-qualia story is incoherent. However, such an objection would be entirely beside the point. Of course, if we supplement the e-qualia story with the clause

(E6) All such states as being aware of experiences are physical,

then a dualist who accepts the claim about superintegration should agree that the e-qualia story+(E6) is incoherent. However, now to make his case, Kirk would be required to show that such a Cartesian dualist should admit the c-possibility of the world z*+(E6). I have no idea how he could do it.

of my present-moment experience. But it cannot be the case with my present-moment awareness (or thinking) of my a-day-ago or even a-few-moments-ago experience. If there is some temporal distance between an experience and a cognitive state having that experience as its object, there should necessarily be a causal link. If so, Kirk's argument can be run beyond its first stage (concerned with the coherence of the e-qualia story) only for those cognitive mental sates about experiences that are not superintegrated with the experiences they are about (such as my thinking about my past experiences). However, a zombist can successfully decline this argument as was explained in section 5.

The general outcome of this discussion is that there are two varieties of zombism that remain unscathed by Kirk's anti-zombist argument as well as by the claim about superintegration:

- a Cartesian dualism that holds that cognitive mental states are non-physical;
- a partially Cartesian interactionist dualism that holds that such states as my present awareness of my present experiences are non-physical, even if other cognitive mental states (such as my thinking about my past experiences or about non-experiential objects) are physical.

And the second part of Kirk's anti-zombist argument (having to do with the $z \rightarrow z^*$ transformation) achieves nothing at all.

References

Bailey, Andrew. 2009. "Zombies and Epiphenomenalism." *Dialogue* 48: 129–14. https://doi.org/10.1017/S0012217309090076

Chalmers, David. 1996. *The Conscious Mind.* New York: Oxford University Press.

Chalmers, David. 1997. "Moving Forward on the Problem of Consciousness." *Journal of Consciousness Studies* 4 (1): 3–46.

Chalmers, David. 1999. "Materialism and the Metaphysics of Modality." *Philosophy and Phenomenological research* 59 (2): 473–96.
https://doi.org/10.2307/2653685

Chalmers, David. 2002. "Does Conceivability Entail Possibility?" In *Conceivability and Possibility*, edited by T. Gendler and J. Hawthorne, 145–200. New York: Oxford University Press.

Chalmers, David. 2004. "Imagination, indexicality, and intensions." *Philosophy and Phenomenological Research* 68: 182-90. https://doi.org/10.1111/j.1933-1592.2004.tb00334.x

Chalmers, David. 2010. "The Two-Dimensional Argument Against Materialism." In D. Chalmers, *The Character of Consciousness*, 141–205. New York: Oxford University Press.

Kirk, Robert. 1974a. "Sentience and Behaviour." *Mind* 83 (329): 43–60. https://doi.org/10.1093/mind/LXXXIII.329.43

Kirk, Robert. 1974b. "Zombies v. Materialists." *Proceedings of Aristotelian Society*, 48: 135–152.

Kirk, Robert. 2005. *Zombies and Consciousness.* Oxford, New York: Oxford University Press.

Kirk, Robert. 2008. "The Inconceivability of Yombies." *Philosophical Studies* 139: 73–89. https://doi.org/10.1007/s11098-007-9103-2

Kirk, Robert. 2017. *Robots, Zombies and Us. Understanding Consciousness.* London, New York: Bloomsbury Academic.

Levine, Joseph. 2001. *Purple Haze: The Puzzle of Consciousness.* New York: Oxford University Press.

Robinson, Howard. 2004. "Dennett on the Knowledge Argument." In *There's Something about Mary*, edited by P. Ludlow, Y. Nagasawa, and D. Stoljar, 69-73. Cambridge, London: The MIT Press.

Robinson, Howard. 2016. *From the Knowledge Argument to Mental Substance.* Cambridge University Press.