Interactionist Zombies

Forthcoming, *Synthese*

Penultimate Draft

One of the most popular arguments in favor of dualism is the conceivability argument, advanced especially by Chalmers (1996).[1] The argument concerns the notion of a zombie; roughly, an exact physical duplicate of a conscious creature, but lacking phenomenal consciousness. In particular, the argument typically relies on the notion of a zombie *world-pair*, a pair consisting of (i) a world *w* at which phenomenal properties are instantiated and (ii) an exact physical duplicate of *w* at which no phenomenal properties are instantiated. The basic argument proceeds as follows:

CONCEIVABILITY ARGUMENT:

(1) **Conceivability**: Zombie world-pairs are conceivable.

(2) **Conceivability-Possibility Link**: If something is conceivable, then it is metaphysically possible.

(3) **Lemma**: From (1) and (2), zombie world-pairs are possible.

(4) **Global Supervenience**: For all A, B: A and B are non-distinct only if no two worlds differ with respect to their A-properties without also differing with respect to their B-properties.[2]

(∴) **Dualism**: Mental properties are distinct from physical properties.

Most opponents of dualism get off the boat at **Conceivability-Possibility Link**: it is often denied that conceivability is a reliable guide to metaphysical possibility.[3] This paper, though, is concerned

[1] See also Chalmers (2009) and Stoljar (2001). For criticism, see Hill (1997) and Hill and McLaughlin (1999). For recent criticism of conceivability arguments more broadly, see Campbell, Copeland, and Deng (2017). See Chalmers (2002) for much more on conceivability and the conceivability-possibility link.

[2] "Non-distinct," here, is an umbrella term that is meant to capture the ideas that mental and physical properties are identical, that the mental is metaphysically grounded in the physical, and other such proposals that are standardly thought to count as physicalist.

[3] See Balog (1999 and 2012). Some, including Frankish (2007), have offered what they take to be conceivability arguments in favor of *physicalism*, perhaps undermining the force of CONCEIVABILITY ARGUMENT. Baysan and Wildman (forthcoming) also offer a disjunctive account of phenomenal consciousness which they argue undermines both conceivability arguments for dualism and physicalism.

with a putative consequence of the possibility of zombies for interactionist dualism. It has been argued that the possibility of zombies entails that mental states/properties are epiphenomenal.[4] According to this line of thought, the possibility of an exact physical duplicate of our world which lacks phenomenal properties would entail that those phenomenal properties play no role in bringing about physical effects in the actual world. Hence, perhaps the greatest rhetorical arrow in the dualist's quiver – the conceivability argument – is inconsistent with interactionist versions of dualism.

This paper aims to defuse the objection to interactionist dualism. First, I will briefly review the argument in section I. Then, I will attempt to more clearly explore which counterfactual criteria are being employed as necessary conditions for causation, in section II. I will then describe a type of possible world-pair which is perfectly consistent with interactionism in section III, and which will serve as a basis on which to construct an interactionist-friendly conceivability argument for dualism, in section IV.[5] I will discuss a potential challenge for this account, concerning whether some of the psychophysical regularities in the relevant zombie worlds constitute new physical laws and, hence, undermine the physical isomorphism that was thought to exist between the two worlds, in section V. I conclude the discussion in section VI. To clarify the scope of discussion, I will not be making any global argument in favor of dualism. All I hope to establish is that, insofar as one finds conceivability arguments convincing, one is not thereby compelled to be an epiphenomenalist.

I. Epiphenomenal Zombies

If zombies are possible, then physicalism is false. But in order to qualify as zombies, these creatures must preserve our physical structure *in every detail.* There are many ways to cash out this idea, employing different notions of supervenience. But the most straightforward is that, for physicalism to be true, the mental must *globally supervene* on the physical: any world which is an exact physical duplicate of the actual world must also be an exact mental duplicate.[6] The conceivability of zombie-worlds threatens to undermine the global supervenience of the mental on the physical. That is good news for the dualist!

[4] Perry (2001, 2012) and Bailey (2006, 2009) are central advocates of this argument.

[5] My proposal, as will be seen, is in line with dualists who, in response to the causal exclusion argument popularized by Kim (1998, 2005), deny the causal closure of the physical, such as Lowe (2003) and Won (2021).

[6] On global supervenience, see Bennett (2004).

Yet, there is a snag: imagine the zombie duplicate of our world. This zombie world is *exactly* like our world throughout its entire history, in all physical detail. Thus, my zombie twin – even though the lights are completely out, internally – still writes a paper about dualist mental causation and winces at his untimely tendonitis "pain" while doing so. And if this is true, then it would appear that my own phenomenal properties make *no difference* as to whether I will display all of my seemingly mentalistic behavior. Therefore, the possibility of zombies entails that consciousness is epiphenomenal:

> **Epiphenomenal Zombies**: If zombie-worlds are possible, then mental properties are epiphenomenal.

Epiphenomenal Zombies may create trouble for different reasons. Most strongly, it may be thought that the possibility of zombies entails that our introspective reports about consciousness are somehow unreliable.[7] The idea, here, is that our zombie twins would produce exactly the same introspective reports, and would register all of the same information about conscious experiences as we do. Therefore, the evidence that we thought we had in favor of our being conscious turns out not to depend on the presence of consciousness in the first place, and so that evidence is unreliable.

But a softer approach, advocated by Perry (2001, 2012), Bailey (2006, 2009), and Mohammadian (2021) suggests that the possibility of zombies is merely inconsistent with *interactionist* dualism, and so interactionists cannot consistently invoke zombie-conceivability arguments against physicalism. It is this version of the argument that I will be primarily concerned with.

The basic idea behind **Epiphenomenal Zombies** is that, if zombie-worlds are possible, then phenomenal properties are causally redundant. See, for instance, Bailey:

> It is natural to think of zombies in the following way: zombies are what are left over when you subtract phenomenal consciousness from the actual world, or a relevant part of it, while leaving everything physical unchanged. To say that zombies are logically/metaphysically possible is to say that it is logically/metaphysically possible to perform this subtraction. But to allow this is precisely to allow

[7] Carroll (2021) makes an argument in this ballpark, and a similar case regarding the Knowledge Argument is given by Watkins (1989) and Moore (2012).

> that the presence of phenomenal consciousness is unnecessary for any physically specifiable event to occur, since ex hypothesi all these events (or states, or properties) *would occur anyway even if consciousness were not present* (2009, 131; emphasis added).

It is clear that Bailey has in mind a necessary criterion of *counterfactual dependence* for causation, as the emphasized portion suggests. Thus, we can state the argument as follows:

EPIPHENOMENAL ZOMBIES ARGUMENT:

(1) **Independence**: If a world with mental properties has a zombie duplicate, then the physical effects of that world are counterfactually independent of the mental properties.

(2) **Counterfactual Dependence**: A causes B only if B counterfactually depends on A.

(∴) **Epiphenomenal Zombies**: If zombie-worlds are possible, then mental properties are epiphenomenal.

This argument has intuitive pull, and seems to rest on a well-motivated necessary condition for causal connection. Yet, as we will see, **Counterfactual Dependence** is too strong;[8] and once we relax the requirement, interactive dualists will find an opening to make modified conceivability arguments.[9]

[8] It should be said that **Independence** is not obviously true either, because it is conceivable that there be a zombie duplicate of our world which is not the *closest* non-mental world to actuality. The possibility of a zombie duplicate of the actual world, then, does not immediately entail **Independence**. I will, however, grant **Independence** for the sake of argument. One reason for doing this is that **Independence** may well be true, and it is true for any world with deterministic laws, assuming that the nearest non-mental world is one in which mental properties are subtracted and the physical world is left unchanged with respect to laws and initial conditions. In that case, the deterministic laws will allow only one possible history given the initial physical conditions; so, the zombie-duplicate world would be the *unique* non-mental world which preserves the physical laws and initial conditions, while also adding no extra non-mental/non-physical properties. Hence, it would be better if an interactionist-friendly conceivability argument can be made consistent with **Independence**.

[9] Chalmers' (2004) response to the original challenge is to suggest that our zombie-duplicate world contains unfilled causal gaps. But it is not clear that such a world is physically possible. For one thing, if the laws are deterministic, and the physics of the world is causally closed, then only one future history is possible given the initial conditions. Consequently, if there is any world which evolves to exhibit the same mentalistic behaviors in the absence of phenomenal properties, then those behaviors are nomically necessitated by the initial physical states, irrespective of the instantiation of phenomenal properties. Such physical states, then, would presumably be causally sufficient for the relevant behaviors. Hence, the kinds of worlds that Chalmers posits are either physically impossible – thus requiring a physical difference, namely a difference in the physical laws, in violation of the supervenience requirement – or the mental properties of the actual world are genuinely causally redundant in the sense that the *nearest* non-mental world to actuality instantiates the same physical properties

II. Criteria for Causation

Consider the following two criteria as candidate necessary conditions for mental causation, where M represents the instantiation of a mental property (e.g. the feeling of pain), P is M's underlying physical realizer (e.g. the firing of C-fibers), and B is the candidate behavioral effect of M (e.g. screaming out in pain).

The Minimal Criterion: M causes B only if $(M \,\&\, {\sim}P) \mathrel{\Box\!\!\rightarrow} B$.

The Strict Criterion: M causes B only if (i) $(M \,\&\, {\sim}P) \mathrel{\Box\!\!\rightarrow} B$, and (ii) $(P \,\&\, {\sim}M) \mathrel{\Box\!\!\rightarrow} {\sim}B$.[10]

The idea behind **The Minimal Criterion** is that, if M happens to be co-instantiated with a (perhaps disjunctive) physical realizer P, we should be able to separate M from P and still have it be the case that B will occur in the closest possible M-world. Such a world would, of course, be nomologically impossible, since it would presumably violate a psychophysical law to the effect that P-instantiations necessitate M-instantiations.[11] Nevertheless, the causal *oomph* of interactive-dualistic mental properties arguably comes in part from the notion that the mental properties *alone* are sufficient to bring about the relevant behavioral effects. Hence, the nearest world in which these mental properties went unrealized (or in which they had alien physical or non-physical realizers) is still one in which the behavioral effects come to bear, if the mental properties are causally efficacious at all.[12]

throughout its entire history. Alternatively, you could put in some extra non-physical *and* non-mental stuff to close the causal gaps; but this seems to face a similar issue, in that the mental will still supervene on the non-mental, or the non-proto-mental. A more plausible response, which comes from personal communication with Chalmers, might be to restrict the supervenience base to physical *properties*, excluding physical laws, and make a no-supervenience argument for dualism from there. Such an argument, though, hasn't been spelled out in the literature, and if my arguments in the next section succeed, it will be unnecessary to restrict the supervenience base in this way.

[10] As I am using it, the semantics for "$\Box\!\!\rightarrow$" are such that $(P \mathrel{\Box\!\!\rightarrow} Q)$ is true at *w* iff there exists a possible (P & Q)-world which is closer to *w* than any (P & ~Q)-world.

[11] For this reason, in fact, dualists may find even **The Minimal Criterion** to be too strong – see Vaassen (2019). However, as I will show, dualists need not deny **The Minimal Criterion**, and maintaining it will allow for a more robust picture of dualist mental causation.

[12] Of course, this is only going to be a necessary condition for *dualist* mental causation, since physicalists will hold not just that the antecedent is nomically impossible, but also metaphysically impossible. The resultant counterpossible may be trivially true – depending on how you prefer to assess counterpossibles – but it wouldn't be true in any interesting sense for mental causation.

The Strict Criterion, by contrast, adds the condition that the behavioral effects counterfactually depend on the instantiation of mental properties, so that, were the mental properties not instantiated, the behavioral effects would not come to bear. On standard accounts of overdetermination (e.g., from Bennett 2003 and 2008), $(M \,\&\, {\sim}P) \Box\!\!\rightarrow B$ and $(P \,\&\, {\sim}M) \Box\!\!\rightarrow B$ jointly obtaining is what it takes for M and P to causally overdetermine B.[13] Consequently, **The Strict Criterion** amounts to the claim that causal overdetermination is metaphysically impossible. Hence, while the two accounts agree that causes must be counterfactually sufficient for their effects, they disagree on whether causal overdetermination and preemption are metaphysically possible.

Which of these accounts is correct? Outlawing overdetermination carries some intuitive plausibility. Yet, **The Strict Criterion** is too strong. Generally speaking, most in the causation literature agree that there are genuine albeit rare cases of causal overdetermination. For instance, consider the following example (Fenton-Glynn 2021, 35):

> *Symmetric Overdetermination*: Alice and Bob are hunting deer. Each has a loaded gun. They spot a deer. They both shoot, with their bullets piercing the deer's heart simultaneously. Each bullet alone would have been sufficient to bring about the deer's death.

Examples like *Symmetric Overdetermination* are quite broadly thought to pose trouble for counterfactual dependence theories of causation in the vein of Lewis (1973).[14] While it is true that counterfactual dependence is present in typical instances of causation, it does not appear to be a necessary condition for causation. As Kroedel (2015) puts it, "while counterfactual dependence is *merely a sufficient condition for causation* and not a necessary one, in standard cases effects counterfactually depend on their causes" (361, emphasis added). This, it seems, constitutes evidence that most in the causation literature would rebuke a complete denial of causal overdetermination;

[13] See Kroedel (2015), though, for dissent.

[14] For instance, see Kroedel (2008), Moore (2009), McDermott (2002), Hitchcock (2007), and Halpern and Pearl (2005) on *Symmetric Overdetermination* and the corresponding problem for Lewis' account. See also Hitchcock (2001) for an interventionist account of causation which, like Lewis' account, is counterfactual in nature, but which, unlike Lewis' account, handles causal overdetermination. See also Sider (2003) for more on why we generally should allow for causal overdetermination.

hence, advocates of **Epiphenomenal Zombies** should not rely on the assumption that overdetermination is metaphysically impossible.

Rather than outlawing overdetermination completely, we can adopt **The Minimal Criterion**, but supplement it with the following:

> **No Widespread Overdetermination**: In modal space, genuine causal overdetermination is extremely rare, and in the actual world, there are no two event- or property- types ϕ and ψ each of which individually causes some effect E, yet both of which are always actually co-instantiated.

There is substantial evidence that the literature on mental causation and causal exclusion displays a preference for this position over **The Strict Criterion**. Consider, for instance, Kim's exclusion principle (2005, 42):

> **Exclusion**: No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

Clearly, this formulation of causal exclusion leaves open the possibility of genuine causal overdetermination. The exclusion argument against dualist mental causation is then supplemented by an additional premise that behavioral effects are not overdetermined by individually sufficient physical and mental causes. One of the central motivations for this premise is that physical/mental overdetermination would require widespread coincidences (Kim 1998, 52-53). Any genuine instance of causal overdetermination that we ever observe or have reason to accept is incredibly rare – seldom, for instance, do two bullets simultaneously puncture the heart of a deer. As another example of this line of thought in the literature, Papineau (2002, 18) explicitly allows that overdetermination sometimes occurs:

> Now, some events are indeed overdetermined… like the death of a man who is simultaneously shot and struck by lightning.

Hence, Papineau's argument against dualist mental causation makes use of a weaker premise (p. 18):

> The physical effects of conscious causes aren't *always* overdetermined by distinct causes [emphasis added].

Finally, in support of this premise, Papineau says (p. 28):

> We don't find any 'belt and braces' mechanisms elsewhere in nature—that is, mechanisms which ensure that certain classes of effects invariably have two distinct causes, each of which would suffice by itself.

The allegation from those who oppose dualist mental causation, then, is that it would require that behavioral effects are *always* or *frequently* overdetermined; this kind of systematic overdetermination is rejected as implausible. However, it leaves the door open for a conception of dualist mental causation on which (1) physical causes are *not* ordinarily sufficient for the relevant behavioral effects, and (2) physical causes can *sometimes* be sufficient, and when they are, this yields a rare occurrence of genuine causal overdetermination. And intuitively, if there are such rare occurrences of physical/mental overdetermination, we can imagine "patching up" those occurrences into a single (very distant) world which does host widespread overdetermination. Such a world would not violate **No Widespread Overdetermination**, since the latter only requires that overdetermination be rare in modal space and in the actual world. At that point, we can insist that the actual world (along with the overwhelming majority of possible worlds) does not have a zombie duplicate, but that a distant world which *does* host widespread overdetermination can have a zombie duplicate; and the pair formed by this world and its zombie duplicate can serve as a basis for interactionist-friendly zombie arguments for dualism. Crucial to this strategy will be to make plausible that (1) and (2) are compatible. This is my main objective in the next section.

III. Interactionist-Friendly World Pairs

The way an interactionist should respond to the epiphenomenal zombies argument is to agree with their opponents that there are no zombie duplicates of our world. However, interactionists can consider other world-pairs which are consistent with dualist mental causation, and which may serve as a perfectly good basis for modified conceivability arguments.

III.1. CSTM

Before presenting the world-pairs I have in mind, it will be worth pausing to introduce the basic ideas behind Classical Statistical Mechanics (CSTM), especially as presented by Albert (2000, 2015).[15] The basic postulates of CSTM include:

(i) Deterministic dynamical laws; e.g., the classical laws of motion,

(ii) A uniform probability distribution over the possible microstates of the universe compatible with its macrostate, and

(iii) A characterization of the initial state of the universe as a special low-entropy condition.

CSTM, as presented, dominates the philosophical literature on statistical mechanics.[16] It presupposes a classical understanding of the laws of motion only because doing so simplifies matters in a way that typically does not implicate the philosophical discussions at issue. Moreover, it is generally acknowledged that Quantum Statistical Mechanics (QSTM) will need to accommodate roughly the same probabilistic predictions as CSTM.[17]

CSTM has received so much attention in part because of its role in accounting for thermodynamic phenomena. In particular, postulates (ii) and (iii) are crucial for understanding the Second Law of Thermodynamics, which states (roughly) that the entropy of an isolated system tends to increase over time, where entropy (in the Boltzmannian treatment) is defined as follows:

[15] See also Loewer (2001).

[16] For a sample of the physical and philosophical literature on CSTM, see Feynman (1967, ch. 5), Brown and Uffink (2001), Uffink (2001), Loewer (2012), Callender (1997, 2011), Cohen and Callender (2010), Frisch (2010), Leeds (2003), North (2011), Weslake (2014), and Winsberg (2004).

[17] See Chen (2021) and Chen and Tumulka (2022) for more on this.

$$S = k_B\log(W)$$

Here, k_B is a constant, while W is the number of microstates that can realize some particular macrostate of a system. For example, a gas of certain pressure, temperature, etc., packed tightly into the corner of a box can be realized by relatively few microstates, compared to a gas which is spread widely throughout the box. Hence, the entropy of the first sort of gas is much lower than the entropy of the second.

Boltzmann (1895) famously proved that, for overwhelmingly most microstates a system can take, the classical laws of motion will evolve the system to have higher entropy in the future. For instance, overwhelmingly most of the possible ways that the particles in an ice cube (sitting in a room temperature environment) can be arranged – where the "ways the system can be microscopically arranged," the system's possible *microstates*, are the set of possible values for the positions and momenta of all of its particles – correspond to that ice cube melting toward the future. Similarly for gas dispersing, glass breaking, and so forth. Postulate (ii) explains why this happens: by applying a uniform distribution over the possible microstates a system (in this case, the entire universe) can assume, we are in a position to explain why it is highly probable that the system's entropy will increase, since overwhelmingly most microstates correspond to increasing entropy.

Postulate (iii) is required to ensure that CSTM makes accurate statements about the *past* history of a system. This is because the classical laws of motion exhibit a feature called time-reversal invariance, which intuitively says that you can perform a simple operation on the microstate of a system – in this case, reversing the momenta of all the particles – and get the result that the system would evolve *backwards* in time in precisely the same way that it in fact evolves forward in time. But if this is so, then it appears that any statistical argument that entropy will increase toward the future is also an argument that entropy will increase toward the past, given the time-reversible dynamics. That means that we should expect a half-melted ice cube to be fully melted not only an hour in the future, but also an hour in the past! This is, of course, an absurd result, and formed the basis of the Reversibility Objection to Boltzmann's approach.[18] Postulate (iii) is intended to avoid the

[18] See Brown, Myrvold, and Uffink (2009), Davies (1974), and Sklar (1993) for more on the Reversibility Objection.

Reversibility Objection, by conditionalizing all statistical-mechanical predictions on the universe having been in a very low entropy state at some point in the distant past, around the time of the Big Bang. Conditional on *that* postulate, it turns out that entropy is very likely to have increased between the past and present states, in line with our experience.[19]

The point of my invoking CSTM is only to point out that it is now well-understood that the dynamical laws, on any of our best fundamental physical theories, will not suffice to rule out many *seeming* physical impossibilities, from entropy spontaneously decreasing to a rock in projectile motion spontaneously decomposing into statuettes of the British Royal Family, or reciting the Gettysburg Address.[20] The addition of CSTM to our physical picture of the world serves not to rule out such scenarios – indeed, they remain genuine nomic possibilities. Rather, CSTM renders such scenarios *extremely* unlikely, insofar as they would be realized by highly atypical microstates. It is this broad lesson that one should bear in mind as I introduce, in the remainder of this section, the world-pairs that I think can serve as a basis for an interactionist-friendly zombie argument.

III.2. Miraculous World-Pairs

The basic interactionist story that I want to put forward is one on which the mental is not only causally sufficient to bring about the relevant sorts of behavioral effects, but also that the mental is *ordinarily* a causal difference-maker, in that the effects standardly counterfactually depend on the mental properties and are not causally overdetermined. However, I will argue that there are worlds in which this is not always the case; and *only these types of worlds* have zombie-duplicates. Given the existence of these world-pairs, a non-supervenience argument for dualism can be constructed.

Start, first, with a causal chain like the following.

[19] For literature on CSTM and the low-entropy boundary condition in postulate (iii), see Penrose (1994), Callender (2004a, 2004b), Frigg (2009a, 2009b), Goldstein (2001), and Parker (2005).

[20] The colorful examples are due to Albert (2015).

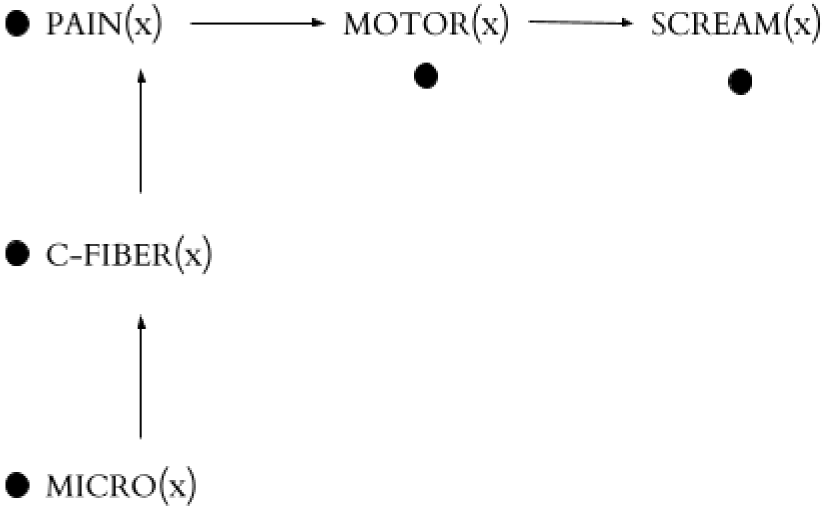

Figure 1

In this diagram, the vertical arrows represent some kind of *non-causal* dependence-relation, such as strong emergence or realization. The horizontal arrows represent causal dependence. MICRO(x) refers to the instantiation of some precise microphysical state for some system x, C-FIBER(x) refers to the instantiation of the brain state of C-fiber activation, PAIN(x) refers to the instantiation of the phenomenal property of pain, MOTOR(x) refers to the instantiation of some motor response in response to pain, and SCREAM(x) refers to the behavioral effect of a conscious being screaming out in pain.

In order to satisfy **No Widespread Overdetermination**, the interactionist must build into their model of mental causation that, standardly, behavioral effects like SCREAM(x) counterfactually depend on phenomenal properties like PAIN(x), and that, were the antecedent physical facts held fixed while the phenomenal properties were removed, the given effects too would disappear. This is reflected in the following diagram.

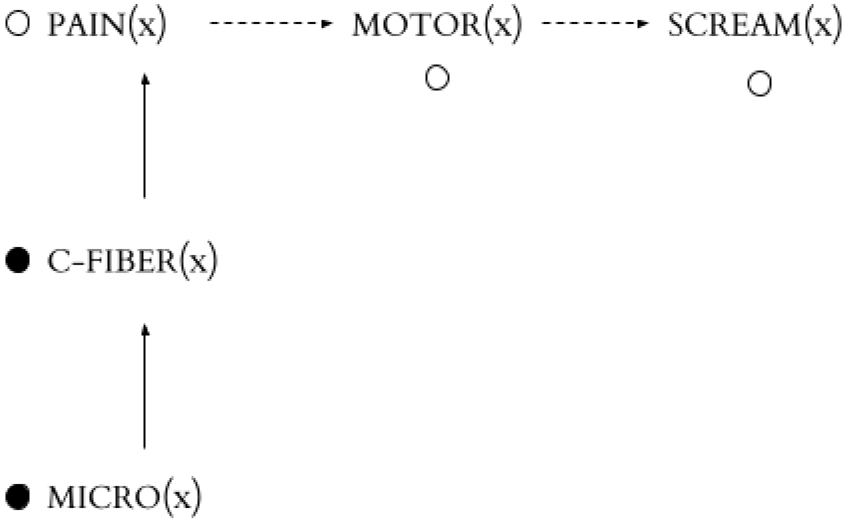

Figure 2

In *Figure 2*, a dark circle indicates that the given instantiation in fact took place, whereas a light circle indicates that it did not. The dotted horizontal arrows indicate the causal processes that *would* have taken place had PAIN(x) been instantiated but which, because PAIN(x) was not instantiated, were inhibited. Thus, *Figure 2* depicts a process in which MICRO(x) and C-FIBER(x) are instantiated without PAIN(x) (perhaps in violation of some contingent emergence-law), and as a result, the neural response and behaviors that would have resulted from the instantiation of PAIN(x) do not occur. In order to accommodate **No Widespread Overdetermination**, dualists should maintain that for overwhelmingly most microstates, a process like that in *Figure 2* is what happens when mental properties are removed from the picture. In other words, for typical MICRO(x): MICRO(x) & ~PAIN(x) □→ ~SCREAM(x).

Furthermore, in order to satisfy **The Minimal Criterion**, PAIN(x) must be causally *sufficient* to bring about the relevant effects. In other words, even if PAIN(x) had gone unrealized, or had been realized by some other micro/macrophysical states, the effects would still have occurred: PAIN(x) & ~MICRO(x) & ~C-FIBER(x) □→ SCREAM(x). Such causal processes would look as follows:

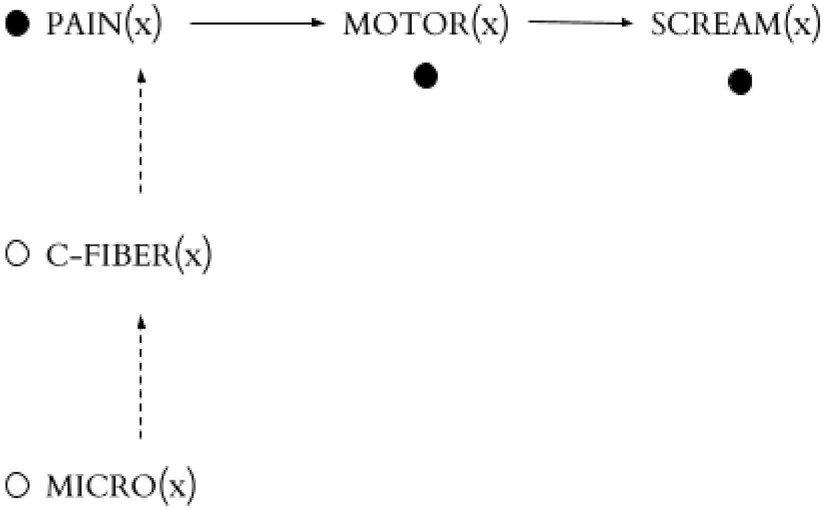

Figure 3

Yet, for all that has been said, dualists can consistently maintain that there are *some* possible initial microstates which are highly atypical, in the sense that they would evolve to produce a sequence of states in which a brain state like C-FIBER(x) is immediately followed by a brain state like MOTOR(x), which then goes on to produce the behavior SCREAM(x), even without instantiating PAIN(x). Such sequences of states, as it were, come entirely from the microphysics, and by pure accident. It just so happens that, for some very special MICRO(x), the laws of motion will evolve to have completely unconscious creatures screaming out in what appears to be pain. This would not be due to any genuine higher-level causal links between C-FIBER(x) and MOTOR(x), but rather due entirely to the highly atypical features of the initial microstate. Just as our fundamental physical laws do not suffice to rule out bizarre processes such as glass unbreaking, ice unmelting, and rocks reciting the Gettysburg Address, even an interactionist can admit that there are some very special microstates which would evolve to produce what appear to be mentalistic behaviors, *with or without* consciousness being present! This idea comes naturally from CSTM, which renders these bizarre microstates unlikely without making them nomically impossible. Such causal processes would look as follows.

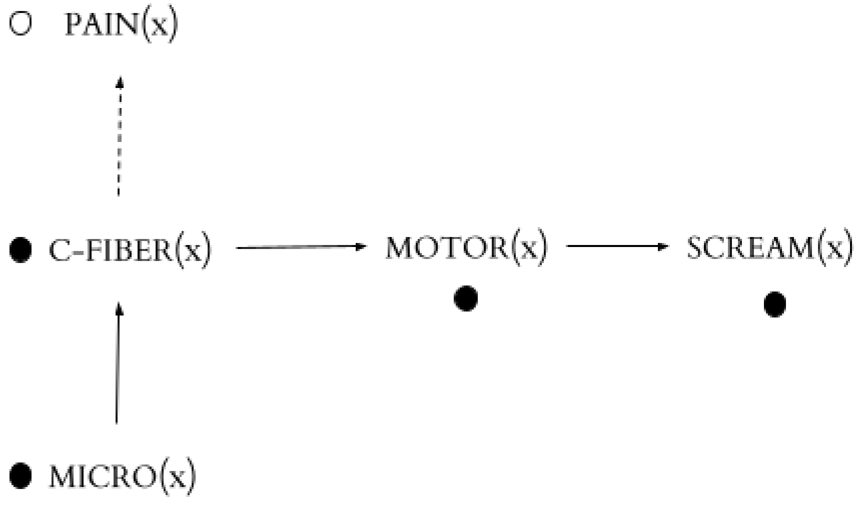

Figure 4

In *Figure 4*, PAIN(x) is not instantiated despite the relevant physical properties being instantiated – perhaps because a psychophysical law of the actual world has been violated. Nevertheless, because of a very atypical and seemingly miraculous microstate MICRO(x), there is nevertheless a highly coincidental sequence of states whereby MOTOR(x) and SCREAM(x) are instantiated just following C-FIBER(x).

Consider the kinds of causal process envisaged by *Figure 4*. Now imagine a world which has such an unfathomably atypical initial microstate that it hosts *many* causal processes which have this structure: i.e., many causal processes in which a precise initial microstate, by pure coincidence, evolves to realize a sequence of brain states which has all the markings of mental causation, *even though no mental properties are instantiated*.

We are now in a position to construct a world pair, $\langle w^*, w^*_Z \rangle$, on which an interactionist-friendly zombie argument can be constructed. w^* is a world that hosts mental property-instantiations, in which these properties give rise to causal processes with the structure pictured in *Figure 1*. Yet, w^*_Z is a world which lacks mental property-instantiations entirely, but which has a striking feature: wherever in w^* there is a causal process like that in *Figure 1*, w^*_Z hosts a causal process like that in *Figure 4*. In other words, w^*_Z ends up looking just like w^*, because wherever the mental properties in $w^{\star}$ are playing a causal role, the spooky microphysics of $w^{\star}_{Z}$ steps in to produce exactly the same sequences of neural and behavioral responses.[21] We can further suppose that $w^{\star}$ and $w^{\star}_{Z}$ share the same precise initial microstate, and that said microstate evolves deterministically throughout all of history in exactly the same way in both $w^{\star}$ and $w^{\star}_{Z}$.[22] Consequently, the two worlds are quark-for-quark physical duplicates, and yet one hosts phenomenal properties while the other is a zombie world. As a result, $w^{\star}$ also hosts an atypical microphysics which is sufficient, on its own, to bring about the relevant behavioral effects. Hence, the relevant effects are all causally overdetermined in $w^{\star}$ by the mental and microphysical properties.

For vividness, consider an analogy. Suppose you had a world in which, just by chance, every time a deer's heart was punctured by a bullet, it happened that there were two bullets puncturing the heart simultaneously – i.e., every instance of deer-shooting in that world was also an instance of *Symmetric Overdetermination.* And now imagine that you had a second world in which, every time a deer was shot in the first world, the second world is the nearest world in which only one bullet had pierced the deer's heart. The first world is analogous to $w^{\star}$, and the second world to $w^{\star}_{Z}$, in which the mental and physical property-instantiations are analogous to the first and second bullet, and the behavioral effects are analogous to the deer's death. Clearly, a world full of symmetrically overdetermined deer-shootings is incredibly rare, just as $w^{\star}$ is incredibly rare, on the interactionist model I am proposing. Nevertheless, these rare worlds do not violate any physical laws, nor do they violate plausible constraints on the metaphysics of causation.

[21] For vividness, one may imagine a world that is macroscopically just like the actual world throughout its entire history, but which is just *filled*, slice-by-slice, with these thermodynamic miracles.

[22] Less obvious, perhaps, is that the causal chains in $w^{\star}$ and $w^{\star}_{Z}$ are realized in a microphysically identical way. Might the presence of PAIN(x), when instantiated, still make some difference to the underlying micro-state, despite realizing the same causal chain from a macroscopic perspective? I have in mind, however, an interaction law which is higher-order, i.e. where the nomic connections are between states like PAIN(x) and MOTOR(x), rather than one on which PAIN(x) necessitates any precise changes in the micro-dynamics. And then a natural assumption – arguably in line with scientific practice, is a kind of principle of least action, which intuitively says that a given dynamical transition will keep the change in a system's energy distribution to a minimum, as the system is moved from one point to another. Formally, this means minimizing what is called the action functional, which is the integral $\int(T - V)dt$, where T is kinetic energy and V is potential energy. We might plausibly assume something fairly analogous about our interaction laws, namely that PAIN(x) makes the smallest possible alteration in the particle trajectories (thereby keeping its difference to the energy distribution of a system to a minimum) in order to realize the target macro-physical state, relative to the trajectory that would have occurred in the absence of PAIN(x). But in fact all one really needs is for the interaction law not to be maximally biased against least action, i.e. by not requiring that the phenomenal states *always* make a microphysical difference when the microphysics suffice. As long as they do not always make a microphysical difference in the face of already-sufficient initial conditions, there will be some world pairs our there, like $w^{\star}$ and $w^{\star}_{Z}$, where the microphysics is identical between the two worlds.

The phenomenal properties in w^* satisfy the requirements of **The Minimal Criterion**, as do the micro-physical properties, because *both* are sufficient to bring about the relevant behaviors even in the absence of the other. Hence, w^* hosts widespread overdetermination. Yet, the existence of $\langle w^*, w^*_Z\rangle$ is perfectly consistent with our desiderata for causation, because the extreme rarity of MICRO(x) among the nomically-possible microstates entails that overwhelmingly most possible worlds (and almost certainly the actual world) will not be like w^*, in that instantiations like SCREAM(x) will standardly be counterfactually sensitive to changes in the mental properties. What this means is that **No Widespread Overdetermination** will be satisfied. Yet, even though w^* is very rare, it is nevertheless suitable for a non-supervenience argument for dualism, to be given in Section IV, because all such an argument would require is that there be *some* world-pair in which global mental/physical supervenience is broken.

It is important to emphasize, here, that w^* is nomically possible, despite hosting widespread overdetermination.[23] This is because the laws of our world – which are plausibly of a similar form to CSTM – allow for the existence of the relevant sorts of atypical microstates.[24] In other words, CSTM both explains (i) why worlds like w^* are highly abnormal in the space of nomically possible worlds, and (ii) why worlds like w^* are nevertheless nomically possible. (i) is due to the fact that, in the absence of genuine, binding causal connections between C-FIBER(x) and MOTOR(x), there are many possible microstates consistent with C-FIBER(x) which go onto realize all sorts of different trajectories. *Very few* of those trajectories would then correspond to the right kind of motor activity being realized in the brain. But (ii) is due to the fact that, nevertheless, there is nothing in CSTM (because there is nothing in the laws of motion) which forbid *some* rare microstates where the trajectories so happen to pass through a brain state of C-fiber firing followed by a brain state of pain-producing motor activity! The fact that (i) is satisfied means that w^*, being so rare, is consistent with **No Widespread Overdetermination**. And the fact that (ii) is satisfied means that the

[23] Thanks to an anonymous reviewer for pushing me to clarify this point.

[24] As noted in the previous subsection, CSTM is not itself the correct physical theory of our world, but this is only because the first postulate needs to be replaced with some non-classical laws of motion, such as the Schrodinger Equation. Nothing said here would be undermined by this replacement.

world-pairs which will serve as a basis for the modified conceivability argument of section IV are possible with respect to our physical laws.[25]

One final worry that should be addressed is whether, in w^*, MICRO(x) & ~PAIN(x) is possible, and whether this undermines our assessment of counterfactuals with MICRO(x) & ~PAIN(x) in the antecedent.[26] One concern, here, may be that if MICRO(x) & ~PAIN(x) is *metaphysically* impossible, then the relevant counterfactuals will merely be vacuously true counterpossibles (at least on the orthodox assessment of counterpossibles). While it is true that on *physicalist* solutions to the mind-body problem, MICRO(x) & ~PAIN(x) is metaphysically impossible, this will not be true on dualist accounts, according to which MICRO(x) & ~PAIN(x) is a genuine metaphysical possibility, because the two property-instantiations are distinct. On the other hand, though, one may worry that MICRO(x) & ~PAIN(x) is *nomically* impossible in w^*. If so, might this undermine the treatment of $\langle w^*, w^*_Z \rangle$ as being a pair of physically-identical worlds? The answer is that MICRO(x) & ~PAIN(x) is indeed nomically impossible in w^*, but it is not *physically* impossible. Detaching mental properties from physical properties, on the dualist picture, would violate something like a *sui generis* psychophysical law – perhaps an emergence-law, for instance – to the effect that such-and-such physical property-instantiations are always accompanied by such-and-such mental property-instantiations. But it would not violate any *physical* law in w^*, where there are non-physical laws which guarantee that the physical possibilities form a proper subset of the nomic possibilities. Since a global non-supervenience argument for dualism will require absolute sameness of *physical* facts between w^* and w^*_Z, the worlds must not differ in physical laws. But this does not mean that they can't differ in nomic facts more broadly, if they only differ in *non-physical* laws of nature.

IV. An Interactionist Zombie Argument

Without further ado, and with a world-pair like $\langle w^*, w^*_Z \rangle$ in mind, interactionists can offer a modified conceivability argument for dualism.

[25] Strictly speaking, my arguments could do without this: the argument would still work if, say, I had appealed to worlds which are not physically possible with respect to our laws, but which are physically possible with respect to each other. However, the argument is stronger if the relevant worlds are *actually* physically possible, in part because the argument can now be accepted by those who believe that the actual laws of physics are metaphysically necessary.

[26] Thanks again to an anonymous referee for this point.

MIRACULOUS ZOMBIE-WORLDS CONCEIVABILITY ARGUMENT:

(1) **Global Supervenience**: For all A, B: A and B are non-distinct only if no two worlds differ with respect to their A-properties without also differing with respect to their B-properties.

(2) **Miraculous Zombie-Worlds**: There are conceivably two possible worlds, w^* and w^*_Z such that w^* and w^*_Z are physically identical throughout their entire history, but differ with respect to their mental properties.

(3) **Conceivability-Possibility Link**: If something is conceivable, then it is metaphysically possible.

(4) **Lemma**: from (2) and (3), there exist two possible worlds, w^* and w^*_Z such that w^* and w^*_Z are physically identical throughout their entire history, but differ with respect to their mental properties.

(∴) **Dualism**: Mental properties are distinct from physical properties.

This argument is structurally very similar to the original conceivability argument for dualism. All that is changed is that the argument appeals to world-pairs like $\langle w^*, w^*_Z\rangle$ in order to resist the charge of causal redundancy.

V. Mohammadian's Challenge

Here is a potential worry, inspired by Mohammadian (2021): Do w^* and w^*_Z really share in all physical facts?[27] There is a particular kind of regularity, such as: $\forall x(\text{C-FIBER}(x) \rightarrow \text{SCREAM}(x))$, which obtains in both worlds. This regularity does not appear to be a law in w^*, because the relevant law in w^* is the interaction law $\Box\forall x(\text{PAIN}(x) \rightarrow \text{SCREAM}(x))$.[28] This is supplemented with an emergence law $\Box\forall x(\text{C-FIBER}(x) \rightarrow \text{PAIN}(x))$, and these together explain the regularity $\forall x(\text{C-FIBER}(x) \rightarrow \text{SCREAM}(x))$, which only comes by way of the interaction and emergence laws

[27] Mohammadian's discussion targets Quantum Collapse Interactionism (QCI), the view that consciousness collapses quantum wavefunctions. The following worry does apply, I think, to QCI, where I think Mohammadian is right on point. But, as I attempt to show, it won't apply generally to views which posit a different kind of interaction law for mental properties.

[28] $\Box$, here, is a nomic necessity-operator, rather than a metaphysical necessity-operator.

together. Due to the overdetermination in $w^{\star}$, the macro-regularity in question is *also* explained by the conjunction of the micro-state of the universe plus the fundamental dynamical laws, and perhaps also some metaphysical principles connecting up the microphysics with the macroscopic world.

But have a look at $w^{\star}_{Z}$. Is $\forall x(\text{C-FIBER}(x) \rightarrow \text{SCREAM}(x))$ a law? Mohammadian worries that, in zombie-worlds, regularities between the macrophysics and what were, in the non-zombie world, the effects of phenomenal properties, count as laws of nature. If that is the case, then the zombie world hosts a physical fact that the non-zombie world does not. Hence the worlds are not physically identical, and the zombie argument fails.

Luckily, $\forall x(\text{C-FIBER}(x) \rightarrow \text{SCREAM}(x))$ does not at all fit the bill for a law of nature in $w^{\star}_{Z}$. Mohammadian considers three features of laws, as discussed extensively in the literature on laws of nature.[29] First, laws explain their instances. It is unclear, though, why $\forall x(\text{C-FIBER}(x) \rightarrow \text{SCREAM}(x))$ explains any particular instance of SCREAM(x) in $w^{\star}_{Z}$, since clearly only a very special kind of instantiation of C-fiber firing would instantiate SCREAM(x), namely one realized by precisely the right micro-state. Without also specifying the exact microphysical state, we are left with insufficient conditions for the instantiation of SCREAM(x), and once we do specify the precise micro-state, the macroscopic information is explanatorily redundant.

Second, laws support counterfactuals. The regularity in question, of course, does not support counterfactuals: by stipulation, a different micro-state which realized the same macro-state would not evolve to instantiate SCREAM(x). $w^{\star}_{Z}$ would contain semblances of higher-level causation all over the place, but the relevant counterfactuals would be ultra-sensitive to changes in the underlying microphysics.[30]

Third, the laws support inferential connections between observed and unobserved instances. Consider two miraculous worlds: one in which all instantiations of C-FIBER(x) precede instantiations of SCREAM(x) and one in which only the observed instantiations of C-FIBER(x) precede instantiations of SCREAM(x). Why should the inhabitants of the miraculous world believe they are in the former and not the latter, given that the former is even less likely, according to CSTM, than the latter? Both possibilities would be statistical accidents which are in no way supported by the

[29] See, e.g., Loewer (1996) and Lange (2000) for more on these characteristics of laws.

[30] See Vaassen (2022).

dynamics or by the probability postulate. Of course, agents would be justified in *thinking* that there was a law of the form $\Box\forall x(\text{C-FIBER}(x) \rightarrow \text{SCREAM}(x))$ which had real nomic teeth, which supported counterfactuals, and which played a role in explanation. But they would be wrong. In point of fact, *given* the information that the regularity came entirely from the micro-states, and that the micro-states that were compatible with C-FIBER(x) but incompatible with SCREAM(x) were of vastly higher measure than the ones that they had apparently observed, inferences from observed instances of $\Box\forall x(\text{C-FIBER}(x) \rightarrow \text{SCREAM}(x))$ would be completely unjustified. Suppose, for instance, the agents in this world are local Maxwellian demons; that is, they are capable of observing/reacting to the micro-states, and visualizing all of the other possible micro-states, of *only* the isolated systems they come across. These agents, given sufficient knowledge of how the dynamics map initial regions of phase space to their time-evolved regions, would come to realize that each individual instantiation of C-FIBER(x) & SCREAM(x) that they had come to observe turned out to be an utter fluke: if the micro-physical state had been changed in the slightest, SCREAM(x) would have failed to be instantiated. Consequently, the agents would come to realize that there was no reason to infer from the observed instances of (C-FIBER(x) & SCREAM(x)) that the unobserved instances of C-FIBER(x) would continue to obey this regularity. Hence, the putative law would not support inferences to future or unobserved regularities.

VI. Conclusion

In this paper, I have considered the argument that the possibility of zombie-worlds delivers epiphenomenalism about the mental, and shown that interactionist dualism is perfectly consistent with the possibility of zombies. Hence, insofar as conceivability arguments work for epiphenomenal dualists, sufficiently similar arguments will work for interactionists. Of course, an interactionist who makes such an argument will inevitably have to give up on the causal closure of the physical.[31] For many, this is too much to stomach. But many interactionists are already committed to denying causal closure. Much worse, for the interactionist, would be to learn that their favorite argument against physicalism is unavailable. It is this conclusion that I believe they can avoid.

[31] Though, see other strategies from Vaassen (2019, 2021), Kroedel (2015, 2020), and Bealer (2007).

Acknowledgments: Many thanks to two anonymous referees at *Synthese*, Brian McLaughlin, and especially Bram Vaassen, for many helpful comments on earlier drafts of this paper.

The author expresses no conflicts of interest.

References

Albert, D. (2000). *Time and Chance.* Cambridge: Harvard University Press.

Albert, D. (2015). *After Physics.* Cambridge: Harvard University Press.

Bailey, A. (2006). Zombies, Epiphenomenalism, and Physicalist Theories of Consciousness. *Canadian Journal of Philosophy, 36*(4), 481-509.

Bailey, A. (2009). Zombies and Epiphenomenalism. *Dialogue, 48*(1), 129-144.

Balog, K. (1999). Conceivability, Possibility, and the Mind-Body Problem. *Philosophical Review, 108*(4), 497-528.

Balog, K. (2012). In Defense of the Phenomenal Concept Strategy. *Philosophy and Phenomenological Research, 84*(1), 1-23.

Baysman, U., and Wildman, N. (forthcoming). Physicalism or Anti-Physicalism: A Disjunctive Account. *Erkenntnis*.

Bealer, G. (2007). Mental Causation. *Philosophical Perspectives,* 21(1), 23–54.

Bennett, K. (2003). Why the Exclusion Problem Seems Intractable and How, Just Maybe, to Tract It. *Noûs*, 37(3), 471-97.

Bennett, K. (2004). Global Supervenience and Dependence. *Philosophy and Phenomenological Research, 68*(3), 501-529.

Bennett, K. (2008). Exclusion Again. In J. Hohwy and J. Kallestrup, (eds.), *Being Reduced*. Oxford: Oxford University Press.

Boltzmann, L. On Certain Questions of the Theory of Gases. *Nature* (51): 413-415.

Brown, H. and Uffink, J. (2001). The Origins of Time-Asymmetry in Thermodynamics: The Minus First Law. *Studies in the History and Philosophy of Modern Physics*, 32(4): 525–538.

Brown, H., Myrvold, W. and Uffink, J. (2009). Boltzmann's H -Theorem, Its Discontents, and the Birth of Statistical Mechanics. *Studies in the History and Philosophy of Science*, 40(2): 174–191.

Callender, Craig. (1997). What is 'The Problem of the Direction of Time'? *Philosophy of Science* (64), 223-234.

Callender, C. (2004a). There is No Puzzle about the Low-Entropy Past. In Christopher Hitchcock (ed.), *Contemporary Debates in the Philosophy of Science*. Oxford: Blackwell, 240-257.

Callender, C. (2004b). Measures, Explanations and the Past: should "special" initial conditions be explained? *British Journal for the Philosophy of Science*, 55: 195-217.

Callender, Craig. (2011). The Past Histories of Molecules. In Claus Beisbart & Stephan Hartmann (eds.), *Probabilities in Physics*, Oxford: Oxford University Press, 83–113.

Campbell, D., Copeland, J., & Deng, Z. (2017). The Inconceivable Popularity of Conceivability Arguments. *Philosophical Quarterly, 6*7(267), 223-240.

Carroll, S. (2021). Consciousness and the Laws of Physics. *Journal of Consciousness Studies, 28*, 16-31.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory.* New York and Oxford: Oxford University Press.

Chalmers, D. (2002). Does Conceivability Entail Possibility? In Tamar Szabo Gendler & John Hawthorne (eds.), *Conceivability and Possibility*. Oxford: Oxford University Press, 145-200.

Chalmers, D. (2004). Imagination, Indexicality, and Intensions. *Philosophy and Phenomenological Research*, 68 (1), 182-90.

Chalmers, D. (2009). The Two-Dimensional Argument Against Materialism. In B. McLaughlin, & S. Walter, *Oxford Handbook to the Philosophy of Mind.* Oxford: Oxford University Press.

Chen, E. and Tumulka, R. (2022). Uniform Probability Distribution over All Density Matrices. *Quantum Studies: Mathematics and Foundations*.

Chen, E. (2021). Quantum Mechanics in a Time-Asymmetric Universe: On the Nature of the Initial Quantum State. *British Journal for the Philosophy of Science* 72 (4):1155–1183.

Cohen, J. and Callender, C. (2010). Special Sciences, Conspiracy and the Better Best System Account of Lawhood. *Erkenntnis* 73(3), 427–447.

Davies, P. (1974). *The Physics of Time Asymmetry*. Berkeley: University of California Press.

Fenton-Glynn, L. (2021). *Causation*. Cambridge: Cambridge University Press.

Feynman, R. (1967). *The Character of Physical Law*. Cambridge: MIT University Press.

Frankish, K. (2007). The Anti-Zombie Argument. *Philosophical Quarterly, 57*(229), 650-666.

Frigg, R. (2009a). Typicality and the Approach to Equilibrium in Boltzmannian Statistical Mechanics. *Philosophy of Science*, 76(5): 997-1008.

Frigg, R. (2009b). Probability in Boltzmannian Statistical Mechanics. In Gerhard Ernst & Andreas Huttemann, (eds.), *Time, Chance and Reduction: Philosophical Aspects of Statistical Mechanics*. Cambridge: Cambridge University Press.

Frisch, M. (2010). Does a Low-Entropy Constraint Prevent Us from Influencing the Past? in Andreas Hüttemann and Gerhard Ernst (eds.), *Time, Chance, and Reduction: Philosophical Aspects of Statistical Mechanics*, Cambridge: Cambridge University Press, 13–33.

Goldstein, S. (2001). Boltzmann's Approach to Statistical Mechanics. In J. Bricmont; D. Durr; M. Galavotti; F. Petruccione & N. Zanghi, (eds.), *Chance in Physics: Foundations and Perspectives*. Berlin: Springer.

Halpern, J. Y. and J. Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science* 56, 843–87.

Hill, C., & McLaughlin, B. (1999). There are Fewer Things in Reality Than Are Dreamt of in Chalmers's Philosophy. *Philosophy and Phenomenological Research, 59*, 446-454.

Hill, C. (1997). Imaginability, Conceivability, Possibility and the Mind-Body Problem. *Philosophical Studies, 8*7, 61-85.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy* 98, 194–202.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review* 116, 495–532.

Kim, J. (1989). Mechanism, Purpose, and Explanatory Exclusion. *Philosophical Perspectives, 3*, 77-108.

Kim, J. (1993). The Non-Reductivist's Troubles with Mental Causation. In J. Heil, & A. Mele, *Mental Causation* (pp. 189-210). Oxford: Clarendon Press.

Kim, J. (1998). *Mind in a Physical World.* Cambridge: MIT Press.

Kim, J. (2005). *Physicalism, or Something Near Enough.* Princeton: Princeton University Press.

Kirk, R. (2008). The Inconceivability of Zombies. *Philosophical Studies, 139*(1), 73-89.

Kroedel, T. (2008). Mental causation as multiple causation. *Philosophical Studies* 139, 125–143.

Kroedel, T. (2015). Dualist Mental Causation and the Exclusion Problem. *Noûs*, 49(2), 357-375.

Kroedel, T. (2020). *Mental Causation: A Counterfactual Theory*. Cambridge: Cambridge University Press.

Lange, M. (2000). *Natural Laws in Scientific Practice.* Oxford: Oxford University Press.

Leeds, S. (2003). Foundations of statistical mechanics: Two approaches. *Philosophy of Science* 70(1), 126–144.

Lewis, D. (1973). Causation. *Journal of Philosophy* 70(17), 556-567.

Loewer, B. (1996). Humean Supervenience. *Philosophical Topics, 24*(1), 101-127.

Loewer, B. (2001). Determinism and Chance. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 32*(4), 609-620.

Loewer, B. (2012). Two Accounts of Laws and Time. *Philosophical Studies*, 160, 115-137.

Lowe, E. (2003). Physical Causal Closure and the Invisibility of Mental Causation. In Sven Walter & Heinz-Dieter Heckmann (eds.), *Physicalism and Mental Causation*. Imprint Academic.

McDermott, M. (2002). Causation: Influence versus sufficiency. *Journal of Philosophy* 99, 84–101.

Mohammadian, M. (2021). If Consciousness Causes Collapse, the Zombie Argument Fails. *Synthese, 199*, 1599–1615.

Moore, M. (2009). *Causation and Responsibility: An Essay in Laws, Morals, and Metaphysics*. Oxford: Oxford University Press.

Moore, D. (2012). On Robinson's Response to the Self-Stultifying Objection. *The Review of Philosophy and Psychology, 3*, 627-641.

North, J. (2011). Time in Thermodynamics. In Craig Callender (ed.), *The Oxford Handbook of Philosophy of Time*. Oxford University Press. 312-350.

Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Oxford University Press.

Parker, D. (2005). Thermodynamic Irreversibility: Does the Big Bang Explain What It Purports to Explain? *Philosophy of Science* 72(5): 751-763.

Penrose, R. (1994). On the Second Law of Thermodynamics. *Journal of Statistical Physics*, Vol. 77.

Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. MIT Press.

Perry, J. (2012). Return of the Zombies? In C. Hill, & S. Gozzano, *New Perspectives on Type Identity: The Mental and the Physical* (pp. 251-263). Cambridge: Cambridge University Press.

Sider, T. (2003). What's So Bad About Overdetermination? *Philosophy and Phenomenological Research* 67(3): 719-726.

Sklar, L. (1993). *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*, Cambridge: Cambridge University Press.

Stoljar, D. (2001). The Conceivability Argument and Two Conceptions of the Physical. *Philosophical Perspectives*, 393-413.

Uffink, J. (2001). Bluff Your Way in the Second Law of Thermodynamics. *Studies in the History and Philosophy of Modern Physics*, 32(3): 305–394.

Vaassen, B. (2019). Causal After All: a Model of Mental Causation for Dualists. Dissertation, Umeå University.
http://umu.diva-portal.org/smash/record.jsf?pid=diva2%3A1343629&dswid=-9748

Vaassen, B. (2021). Dualism and Exclusion. *Erkenntnis*, 86(3), 543-552.

Vaassen, B. (2022). Halfway Proportionality. *Philosophical Studies*, 1-21.

Watkins, M. (1989). The Knowledge Argument Against "The Knowledge Argument". *Analysis, 49*(3), 158-160.

Weslake, B. (2014). Statistical Mechanical Imperialism. In Alastair Wilson (ed.), *Chance and Temporal Asymmetry*. Oxford: Oxford University Press, 241-257.

Winsberg, E. (2004). Can Conditioning on the ‘Past Hypothesis’ Militate Against the Reversibility Objections? *Philosophy of Science*, 71(4): 489–504.

Won, C. (2021). Mental Causation as Joint Causation. *Synthese*, 198(5), 4917-4937.