Synthese (2023) 201:63
https://doi.org/10.1007/s11229-023-04046-0

ORIGINAL RESEARCH

The metaphysical neutrality of cognitive science

Kuei-Chen Chen[1] · **Jeff Yoshimi**[2]

Received: 28 May 2022 / Accepted: 10 January 2023 / Published online: 11 February 2023

Abstract

Progress in psychology and the cognitive sciences is often taken to vindicate physicalism and cast doubt on such extravagant metaphysical theses as dualism and idealism. The goal of this paper is to argue that cognitive science has no such implications—rather, evidence from cognitive science is largely (but not wholly) irrelevant to the mind-body problem. Our argument begins with the observation that data from cognitive science can be modeled by supervenience relations. We then show that supervenience relations are neutral, by showing how they can be coherently interpreted in physicalist, idealist, and dualist terms. We distinguish several types of supervenience relation, and show that each coheres better with some positions on the mind-body problem than the other. Since these variants of supervenience are not empirically equivalent, there is a possibility that data from cognitive science will end up supporting some positions on the mind-body problem more than others. It is in this sense that cognitive science is mostly, but not wholly, neutral.

Keywords Metaphysical neutrality · Mind-body problem · Supervenience · Determination · Cognitive science · Neural correlates of consciousness

K.-C. Chen and J. Yoshimi have contributed equally to this work.

✉ Jeff Yoshimi
jyoshimi@ucmerced.edu

Kuei-Chen Chen
ckueichen@gmail.com

[1] Department of Philosophy, National Chengchi University, No. 64, Sec. 2, Zhinan Rd., Taipei 116, Taiwan

[2] School of Social Sciences, Humanities, and Arts, University of California, Merced, 5200 North Lake Road, Merced, CA 95343, USA

1 Introduction

Progress in psychology and the cognitive sciences is often taken to vindicate physicalism and cast doubt on such extravagant metaphysical theses as dualism, idealism, downward causation, and the existence of an immortal soul. As E. O. Wilson once remarked, "The brain and its satellite glands have now been probed to the point where no particular site remains that can reasonably be supposed to harbor a nonphysical mind... mind-body dualism is being completely abandoned at long last..." (1999, p. 108). When the precise neural dynamics responsible for human cognition—and in particular, for human decision-making—are made visible by future neuro-imaging, that will close the case against physicalism, or so the story goes.

The goal of this paper is to argue against this conception of the metaphysical implications of cognitive science. We argue that evidence from cognitive science is largely (but not wholly) irrelevant to the mind-body problem. No amount of progress in the cognitive sciences will settle the question of what the ontological status of mind and brain are, or how exactly they are related, though it might provide some evidence in the direction of one family of mind-body relations over another.

Here is a simple argument in support of our conclusion. Premise: proponents of every position on the mind-body problem accept that we have sciences like psychology and cognitive science, which tell us things like "if you lose your visual cortex, you become blind." Conclusion: one can accept cognitive science and also be a dualist, idealist, etc. In the course of the paper we give a more detailed argument to this conclusion, and show how the compatibility works in detail. A concept that takes center stage in our argument is supervenience.

Mental-physical supervenience is the claim that the physical states of an object determine its mental states.[1] Once we specify everything physical about an object, we thereby fix everything mental about that same object. Supervenience was once a star in philosophy of mind, but in the recent literature attention has shifted from supervenience to a variety of other psycho-physical relations. These include grounding (Correia & Schnieder, 2012; Raven, 2020; Wilson, 2018), constitution (Jurgens & Kirchhoff, 2019; Kirchhoff, 2015), realization (Melnyk, 2018), and the determinate / determinable relation (Haug, 2010; Schroer, 2011; Wilson, 2021).

This paper involves a bit of a supervenience revival; we pull it off the shelf, dust it off, and put it to new use. Nowadays people are more interested in relations that add things to supervenience, but relative to our goals it is preferable to leave supervenience bare. Its meager commitments are its virtue. Focusing only on the way physical states determine mental states makes it possible to bracket difficult (and perhaps essentially intractable) philosophical questions, while empirical research proceeds. In light of this, we will argue that determination can be usefully thought of as a model of empirical data that is largely neutral between metaphysical positions.

More specifically, we distinguish between two broad determination relations: full determination (DET) and probabilistic determination (PDET) (formal definitions of these relations are given in Sects. 4.1 and 4.3). Full determination is the standard view

[1] As Yoshimi (2007) points out, supervenience can be parsed into two components—determination and dependence—but determination is the more essential of the two, and it is often taken by itself to define supervenience.

that the base states of an object completely fix its supervenient states. Probabilistic determination allows for some "slippage" in the relation, so that base states of an object specify probability distributions over supervenient states. Both (DET) and (PDET) have additional parameters that can be used to further specify them, e.g. modal operators and, in the case of (PDET), parameters describing a probability distribution. Thus determination is a "family of metaphysical relations."

We show, in full philosophical and metaphysical detail, the compatibility between this family of relations and the standard positions on the mind-body problem. In the process we spell out how data from cognitive science might provide inductive evidence for or against full as opposed to probabilistic determination. Though one of these relations might be a better model of empirical data, each of them is ultimately compatible with multiple metaphysical schemes, which jointly cover the entire space of standard positions on the mind-body problem. Therefore, data from cognitive science at most tilt the evidence towards certain positions without showing their truth decisively.[2]

The broad outlines of this argument have a distinguished history. At least since Hume it has been recognized that observed correlations do not imply specific metaphysical relationships. Underdetermination of theories by data is a long-standing issue in philosophy of science (Turnbull, 2018), and our paper can be construed as an attempt to deal with one variant of the issue: how metaphysical theories of mind-body relationships are underdetermined by data from cognitive science. In contemporary philosophy of mind, several others have discussed versions of some of the compatibility claims we argue for here (Metzinger, 2000, pp. 4–5; Kriegel, 2020; Owen, 2019). Our approach has at least four distinctive characteristics. First, we propose a general framework that can be extended to cover multiple forms of compatibility. Second, our argument highlights an important pattern in the empirical data pertinent to the mind-body problem. Third, we address the unique challenges idealism poses to compatibility claims. Fourth, we describe conditions under which empirical data could influence the choice of a metaphysical theory.[3]

2 The main argument

We intend to show that cognitive science is mostly metaphysically neutral, and we motivate this claim by examining empirical data from cognitive science. Within cognitive science we focus on data that are relevant to the mind-body problem. A convenient way to designate these data is as "NCC data," where "NCC" refers to the literature on the neural correlates of consciousness. Research in this area seeks to identify patterns of brain activity that reliably co-vary with different types of mental state (below we describe relevant features of these data in more detail). Though research in cognitive science outside of this area also bears on the mind-body problem, and not all

[2] Another approach to selecting between empirically equivalent (or nearly empirically equivalent) theories is that of appealing to super-empirical virtues, like theoretical parsimony. A person who takes this approach might simply stop reading at this point, and say "Physicalism is the best option for super-empirical reason *R*, and so I endorse physicalism." Following Kriegel (2020), we are doubtful about such an approach.

[3] We thus pursue an approach to metaphysics described by Tahko (2011), who argues that empirical evidence is relevant to metaphysics insofar as it helps choose between metaphysical possibilities.

NCC research has precisely the form we emphasize, "NCC data" still designates the main body of empirical research we rely on, and appropriately emphasizes observable correlations.[4]

To avoid confusion in this terminologically dense space, we will say that a metaphysical relation like determination "models" data, on the one hand, and "coheres" with positions on the mind-body problem, on the other (see Fig. 1).

The concept of "modeling data" is drawn from statistics, where a statistical model of a set of data is a mathematical construct which, if it obtains, makes those data likely (Dekking et al., 2005). When this occurs, we also say the data "support" the model. We will argue that determination models NCC data in roughly this way: assuming brain states determine mental states, observed NCC data are likely, and thus the NCC data support determination.

We use "coherence" in a sense that is inspired by its usage in epistemology (BonJour, 1985), where two claims cohere with one another if a plausible story can be told whereby both claims are true. In the present context, a distinction can be made between *weak* and *strong* coherence. Consider a proposition d, which is either (DET) or (PDET), and let p be another proposition that states a position on the mind-body problem. Then d weakly coheres with p just in case they are logically consistent with each other. By contrast, a form of determination d strongly coheres with a position on the mind-body problem p just in case the truth of this position makes d inductively likely. Take dualism as an example. As Fig. 1 shows, we will argue that it strongly coheres with (PDET) but not with (DET). This means that, provided the truth of dualism, there are inductive grounds for thinking that (PDET) is also likely true, but the same cannot be said of (DET). In a similar way, physicalism and idealism strongly cohere with (DET), but not (PDET). Intuitively, the relation of coherence comes in degrees, with weak coherence a limiting case of strong coherence.

We should also clarify what we mean by "metaphysical neutrality." We don't mean neutrality with respect to any issue in metaphysics whatsoever, but rather compatibility with respect to the major positions on the mind-body problem.[5] We will define a family of relations to be "mostly metaphysically neutral" if every position on the mind-body problem coheres with some member of the family of relations, but each relation in the family coheres better with some positions than others (see Fig. 1). The "neutrality" of the family corresponds to the fact that it coheres with all positions on the mind-body problem. The "mostly" corresponds to the fact that each member of the family coheres better with some positions than others. This in turn facilitates the possibility of empirical work providing evidence for some positions on the mind-body problem over others. If empirical considerations end up supporting some members of the family of determination relations more than others, then by way of the unequal coherence relations, they will end up supporting some positions on the mind-body problem more than others.

We can now state the main argument:

[4] As will become clear, our argument doesn't rely on any assumptions about qualia, so in principle we could also work with correlational data that relate brain states to unconscious mental states. That is, our argument could in principle be extended to defend neutrality with respect to the ontological status of unconscious mental states.

[5] Cf. the way Klein et al. (2020) use the term.

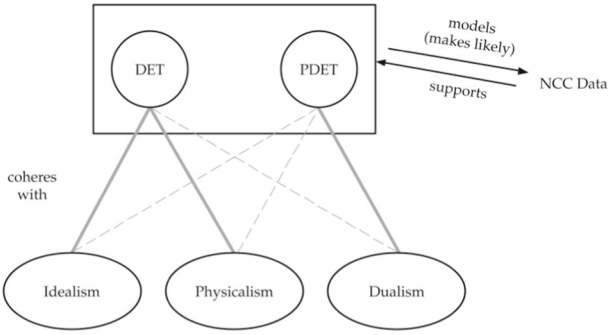

Fig. 1 A conceptual diagram of the main components of the argument. The members of the determination family of relations (top) provide models of NCC data, in the sense of making those data likely. (DET) and (PDET) cohere with all the major positions on the mind-body problem, though each coheres better with some positions than others (solid lines correspond to strong coherence; dotted lines to weak coherence). (DET) coheres best with idealism and physicalism, while (PDET) coheres best with dualism. It is in this sense that the determination family of relations is mostly metaphysically neutral. This near-neutrality carries over to the NCC data that support the relations, and thus to cognitive science

1. If NCC data support a mostly metaphysically neutral family of relations, then cognitive science is mostly metaphysically neutral.
2. NCC data support the determination family of relations. (Sects. 3 and 4)
3. The determination family of relations is mostly metaphysically neutral. (Sect. 5)
4. Therefore, cognitive science is mostly metaphysically neutral. (Sect. 6)

The first premise is meant to be intuitive; the science associated with the data supporting a mostly metaphysically neutral family of relations is itself neutral in the same way (that is, a science inherits the metaphysical neutrality of the data that support it). In other words, even if metaphysical arguments and assumptions are made within a science, to the extent that the data themselves don't uniquely support those arguments and assumptions, the science itself is not committed to them. This may be controversial, but were someone to deny premise 1 our argument would still establish the (only slightly weaker) conclusion that the data of cognitive science are mostly metaphysically neutral.

The argument aims to establish that NCC data don't provide much information about the nature of mind-body relationships, since they are compatible with all major positions. The conclusion is supported by two key ideas. On the one hand, the study of NCCs has produced correlational data that feature patterns of asymmetry,[6] and these can be modeled using determination relations. On the other hand, the determination relations don't uniquely entail any particular view on the mind-body problem. For this reason, NCC data don't force any choice in metaphysics upon us.

The basic formalism of the paper involves quantification over objects and states. On the one hand, the domain of objects (**D**) contains persons (or more broadly "minded entities") at times who are "in" states. Some cases require us to treat the mind and body of a person as separate entities; in those cases, we will speak of the physical domain ($\mathbf{D}_{\mathrm{P}}$) and the mental domain ($\mathbf{D}_{\mathrm{M}}$). On the other hand, we consider two sets of

[6] See Sect. 3 for more discussion.

possible states, **P** and **M**, which denote physical states and conscious states.[7] Physical states capture everything about a person's brain at a time, whereas conscious states capture everything about a person's consciousness at a time. Even though states are highly specific, they are also general in the sense that they do not uniquely identify an object; states are non-individuating.[8]

We will assume that "state" admits both a *property reading* and a *description reading*.[9] Thus there are two ways of interpreting sentences like "person x is in state S". On a *property reading*, "Sx" is true if and only if the property denoted by "S" is instantiated by the object assigned to "x". On a description reading, "Sx" is true if and only if the description that "S" abbreviates is applicable to the object assigned to "x". As an example, consider the predicate "is bored". On a property reading, the sentence "Wilhelm is bored" is true if and only if Wilhelm instantiates the property of boredom. By contrast, on a description reading, the sentence is true if and only if the description "is bored" is applicable to Wilhelm.

A few clarifications relative to this idealized example. First, we focus on maximal states, which fully specify what a person is like mentally or physically (Kim, 1984; Yoshimi, 2007). The predicate "is bored" does not denote a maximal state, since it only specifies part of a person's mental state. Second, we will not assume that the language used to construct maximal states contains such ordinary-language predicates as "is bored". In fact, we will not stipulate what predicates this language contains at all. Our intention is to establish an abstract framework that allows empirical researchers to map their theories to coherent metaphysical positions without worrying too much about metaphysics. Therefore, a user of our framework is free to choose whatever language suits her theoretical needs. In Appendix A we show how well-formed sentences in a user-chosen language can be converted into predicates using lambda abstraction.[10]

3 The cognitive science of consciousness

We now begin to defend the main argument. Since premise 1 is a point we take to be intuitive, we start with premise 2 and examine the nature of NCC data. Our point of departure is the observation that cognitive scientists don't have direct access to brain

[7] In this paper, bold-face letters designate sets of states or objects, upper-case letters designate states, and lower-case letters designate the objects-at-times that can be in states.

[8] Thus, different people can be in the same state at a given time, one person can be in the same state at different times (states can recur), etc. For a discussion of how considerations about individuation might bear on supervenience, see Kim's comments on "individual-specific" and "structure-specific" forms of individuation (1988, p. 135).

[9] See Sect. 5.1 for motivations.

[10] Despite how generic this framework is, it may preclude some metaphysical schemes, e.g. any framework skeptical about the possibility of fully specifying the states of objects. For example, Lowe is doubtful about supervenience, insofar as it is practically impossible to place a system in a particular micro-physical state, down to the last neuron (2012, p. 62). Similar concerns are raised by Cartwright (1999) and Dupré (1993) and other members of the "Stanford School" of philosophy of science, who regard supervenience—and more generally efforts to unify the various special sciences using formal tools—as overly hierarchical and reductionist. We find these forms of skepticism unconvincing. Fruitful idealizations involving precise states are pervasive in science, for example in thermodynamics, where it is useful to assume that two gasses can be in precisely the same state, down to the last molecule.

states or mental states; instead they have access to measurements of states (Tal, 2020). These measurements provide inductive evidence about what state $P \in \mathbf{P}$ or $M \in \mathbf{M}$ a system is in, but there is generally some measurement error, which produces some uncertainty about what the actual state of a system is.

Measurement error can be formalized by treating measurements as random variables.[11] From this perspective, taking a measurement is like using a machine that produces samples from a probability distribution centered at the true value plus or minus some variability associated with the measuring instrument. Measurements of a person's weight, for example, might produce readings centered at 180 pounds, but varying by a few tenths of a pound over multiple measurements.

Random error applies not just to the state being measured, but to the time at which the measurement was taken. A measurement of a person as being 98.5 degrees Fahrenheit at 2.30pm will have some error in terms of temperature, but also in terms of precisely when the person had that temperature. Since the first concerns a state, and the second concerns a time, we can refer to these as "state measurement errors" and "temporal measurement errors" respectively.

Brain states in **P** are typically characterized in terms of neural activity (Dayan & Abbott, 2001). Firing neurons produce electrical currents that can be directly detected using implanted electrodes, and indirectly detected in several ways. Their summed impact on the electromagnetic field can be detected using scalp electrodes via EEG (electro-encephalography). Firing neurons also consume energy, which is associated with oxygenation of blood cells. Blood oxygenation can be detected using functional Magnetic Resonance Imaging (fMRI). There are other techniques as well (Gazzaniga et al., 2019; Kandel et al., 2000). None of these methods can be used to uniquely specify a brain state in **P**; each is associated with some state and temporal measurement error. However, they do complement one another, in that each has a different spatial and temporal resolution. For example, EEG has good temporal resolution but poor spatial resolution; fMRI complements this with good spatial resolution but poor temporal resolution.

The project of investigating the neural correlates of consciousness (NCCs) is extensive and ongoing (Dembski et al., 2021; Koch et al., 2016). Its goal is to identify types of brain activity correlated with types of conscious experience. The project is pursued in multiple ways; we focus on efforts to identify neural correlates for specific types of conscious state. For example, one can dissociate brain states when a tone is heard from brain states when the same tone is registered by the brain but unheard, and similarly for visual experiences. The brain activity specifically associated with conscious perception (but not with unconsciously reacting to the stimulus) is then the NCC for tonal or visual experience.

There is a great deal of work to do making precise what NCCs are, but within the scope of this paper we will only pull out one key implication of this work, an asymmetry in the correlations, whereby (1) varying neural data can be correlated with relatively stable psychological data, but (2) varying psychological data *cannot* be correlated with relatively stable neural data. Although NCC researchers do not, to our knowledge, attempt to support either of these claims directly, existing research is

[11] See, e.g., Dekking et al. (2005, §26.1) for discussion.

consistent with these claims, and there are other lines of research which do support them, and that can be related to (DET) and (PDET) in an intuitive way,[12] In particular, there exist reliable many-to-one functions linking physical states to mental states (in particular psycho-physical functions), but there are no such functions in the reverse direction.[13] This suggests that the brain can change while the mind stays constant, as in Fig. 2, but not conversely. We consider each case separately.

Existence of many-to-one functions from physical to mental states Psychophysical laws show (for example) that a stimulus can be gradually increased in intensity without a participant noticing the change (Murray, 1993). It is generally understood that the changing stimulus produces changing neural activity, which does not produce a mental effect until a threshold is passed. In the 1930s, Wilder Penfield confirmed this idea directly by stimulating cortical tissue, and observing that a minimal intensity of stimulation was required to elicit a clinical response (Mazzola et al., 2012). These results imply that neural activity changes for a time without any measurable change in psychological state. Thus there is evidence for many-to-one functions linking brain states with mental states.

Absence of many-to-one functions from mental states to physical states: There are, to our knowledge, no laws describing many-to-one functions from mental states to brain states. Such a law would allow for small changes in mental activity with no change in brain activity. There are, for example, no laws where people report perception of slowly changing colors, where a corresponding change in the brain can only be detected *after* a threshold of perceived color change is passed. As we just saw, existing psychophysical laws work the other way around. Perhaps such a law could be found in some other domain of cognition or perception, but we are not aware of any.[14]

[12] Claims (1) and (2) are difficult to establish directly, given the fact that both brain and mind are in constant flux. For example, we do not generally record static states of the brain but rather global rhythms like alpha and theta waves. In brain imaging one of the major results of the last few decades was the recognition that when it seems the brain is in its "resting state" it is actually in its "default mode" (Raichle, 2015) producing a dynamical pattern of activations and de-activations correlated with mind-wandering (Mittner et al., 2016). Similarly for conscious experience. Husserl (following Kant) famously argued that temporality is a necessary condition on conscious experience; and it is an obvious fact of direct introspection that our mental states are constantly changing. Even when looking at a blank screen or when submerged in a sensory deprivation tank, our mind wanders in a flow of inner thoughts and spontaneous cognitions (Christoff et al., 2016), and we have some sense of the passing of time. Still, it would be possible to test these claims by establishing a baseline degree of variation in each case and then comparing degrees of *relative* variation, attempting to collect observations where the brain typically varies more relative to its baseline than the mind does relative to its baseline. If we found cases of active brains in people reporting little mental change, but no cases of idling brains in people reporting lots of mental change, then we'd have evidence for asymmetry.

[13] These functions are not actual supervenience functions, like the ones we describe, but rather ways of linking measurements of particular types of neural measurement with particular types of psychological measurement, or vice versa. Such functions, when linking neural measurements to psychological measurements, provide evidence for supervenience functions (as in (DFUN), defined in Sect. 4.1), but are not themselves such functions.

[14] An area that might seem promising is action. For example, we might ask people to increase their level of concentration on a task, and note that only after a certain threshold of mental effort is passed is a change in the brain observed. But it is well known that when people make decisions this is in fact first registered in the brain (Libet et al., 1993) At the first sign of mental effort a change in the brain indicates that this has occurred, and it is only after a certain amount of time has passed that people report awareness of the decision. Thus, here again the law seems to be many-to-one from neural to psychological measurements.

4 Determination

4.1 Determination (DET)

As noted above, our view is that NCC data can be modeled by members of the determination family of relations. We first discuss full determination, which can be defined as follows:

(DET) $\forall x, y \in \mathbf{D}$, if $x =_{\mathbf{P}} y$ then $x =_{\mathbf{M}} y$

Here "$x =_{\mathbf{P}} y$" means x and y are physically indiscernible, and "$x =_{\mathbf{M}} y$" means they are mentally indiscernible,[15] That is, if mental states determine physical states, then any two people who are physically indiscernible are mentally indiscernible as well. The physical states "fix" or "determine" the mental states. By contraposition, if two people are mentally discernible, they are physically discernible.

The indiscernibility relation is interpreted differently depending on whether we read "states" as properties or descriptions; these interpretations are developed in Sect. 5.1. If two people are physically (or mentally) indiscernible, then either they share all physical (or mental) properties, or the same physical (or mental) descriptions apply to them. However, in a context of states *qua* maximal, all that is required for two people to be indiscernible is that they be in the same state (since states concatenate all relevant properties or descriptions into a single complex property or description). Lemmas expressing these connections are in Appendix B.

There is also a functional formulation of determination (Yoshimi, 2012) equivalent to (DET). Since in some contexts this formulation is more intuitive and easier to reason about than (DET), we present the formulation here and prove the equivalence in Appendix B. The formulation is as follows:

(DFUN) There exists a function $f : \mathbf{P} \rightarrow \mathbf{M}$, where for any $P \in \mathbf{P}$ and any $M \in \mathbf{M}$, $f(P) = M$ iff $\forall x(Px \rightarrow Mx)$

That is, being in a physical state $P \in \mathbf{P}$ entails being in a unique mental state $M \in \mathbf{M}$ that is the value of $f(P)$, such that anyone in that physical state would be in that mental state. We can say that the physical state P "determines" M (thus overloading "determination" so that it applies to single states or state sets) or that M is "realized" in P. We allow that $\mathbf{M}$ contains a null state that involves having no mental properties, e.g. being in a dreamless sleep, to simplify the analysis (the alternative is to partition $\mathbf{P}$ into a subset that maps to $\mathbf{M}$ and a subset that does not).

The function f can be many-to-one. Many brain states can be mapped to the same mental state by f. An inverse can be defined for f that associates outputs of f with their pre-images under f, $f^{-1}(Y) = \{X \mid f(X) = Y\}$. This inverse associates mental

[15] Traditionally, indiscernibility is defined using a variant of Leibniz's law, whereby two objects are indiscernible with respect to a set of properties if they have exactly the same properties in that set (see Yoshimi 2007 for discussion). Given how we have defined maximal physical and mental states, however, "physically indiscernible" just means "in the same P in $\mathbf{P}$" and "mentally indiscernible" means "in the same M in $\mathbf{M}$." See (PS1) and (MS1) in Appendix B. Since the members of $\mathbf{P}$ and $\mathbf{M}$ are non-individuating, numerically distinct entities can be physically and mentally indiscernible.

states with the potentially multiple physical states that realize it. In Yoshimi (2012) these pre-images are referred to as "realization classes," since they contain the sets of physical states that realize or determine a specific mental state.

4.2 Temporal alignment condition

We can use the functional formulation of supervenience, (DFUN), to understand how temporal processes at the supervenient level constrain temporal processes at the base level. This will be key to distinguishing (DET) and (PDET) below.

As shown in Fig. 2, (DET) is compatible with physical processes varying even when the supervenient state stays the same. When the supervenient state remains the same, the base system can vary, so long as it stays in the realization class for that supervenient state. However, when the supervenient state changes, for example from M_1 to M_2, the base state must change from the realization class for M_1 to the realization class for M_2. Thus (DET) involves a kind of temporal alignment. In the figure, locations on the horizontal axis represent time, while locations on the vertical axis represent possible states. If two points at the supervenient level occupy the same vertical position but different horizontal positions, this indicates that the system is in the same supervenient state at multiple times. In this case the system is in state M_1 for a time, then in state M_2. During the first time interval the physical state can change, so long as it stays in the realization class $f^{-1}(M_1)$, and similarly for the second time interval. This is why the temporal evolution of M_1 and M_2 are plotted as straight lines but the temporal evolutions within $f^{-1}(M_1)$ and $f^{-1}(M_2)$ are plotted as curves. The hatch-mark indicates when a change occurs in the supervenient state. When the supervenient state state changes from M_1 to M_2 at the hatchmark, the base state much change from $f^{-1}(M_1)$ to $f^{-1}(M_2)$.

The above observations point to a condition on (DET). As long as any change in the supervenient state is associated with an appropriate change in the base state at the same time, (DET) is not violated (see Fig. 3). The requirement represented by Fig. 3 can be stated as follows:

> (Temporal Alignment Condition) For any pair of successive mental states M_1 and M_2 such that $M_1 \neq M_2$, if M_2 replaces M_1 at t, then there are $P_1 \in f^{-1}(M_1)$ and $P_2 \in f^{-1}(M_2)$ such that P_2 replaces P_1 at t

Any account of mental-physical relations that observes the temporal alignment condition can be compatible with (DET). For example, dualist systems and forms of emergence can observe the condition, so long as when mental states change, physical states change in an appropriate way (of course, this is more or less plausible depending on the details of the metaphysical system). The temporal alignment condition only concerns the timing of state changes, and is thus consistent with many different metaphysical accounts of the mental-physical relationship (cf. Kriegel, 2020).

It is, however, possible to violate temporal alignment. For example, on a dualist system in which mental states causally communicate with physical states, there might be a time lag between a mental intention to (say) raise one's arm, and the corresponding change in the motor cortex of the brain, which occurs *after* the mental intention first

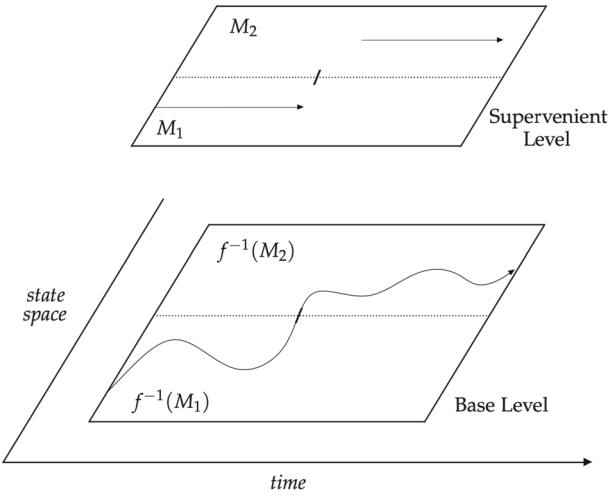

Fig. 2 A space-time diagram illustrating the kind of temporal alignment assumed by (DET). Each of two supervenient states M_1 and M_2 is associated with a realization class consisting of base states that are mapped to them by f. At the time when the supervenient state changes from M_1 to M_2, indicated by a hatch-mark, the base state must move from the realization class for M_1 to the realization class for M_2

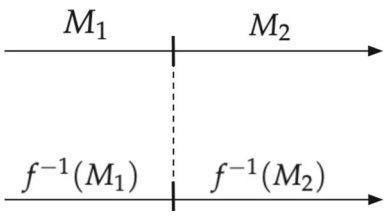

Fig. 3 The temporal alignment condition. When the supervenient state changes from M_1 to M_2, the base state must change as well, from a member of the realization class for M_1, to a member of the realization class for M_2

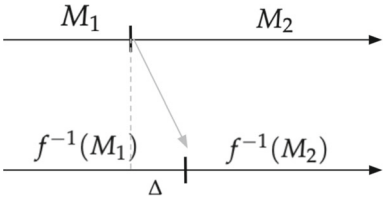

Fig. 4 A violation of temporal alignment, which implies a violation of (DET). The supervenient state changes from M_1 to M_2 *before* the base state moves from the realization class for M_1 to the realization class for M_2. During the resulting time interval Δ the system is in M_2 but not in a base state that realizes M_2, and so (DET) is violated. The gray arrow suggests downward causation has led to this violation of temporal alignment (e.g. an agent makes a conscious decision, but it takes some time for that decision to impact the brain), but this is not the only way the temporal alignment condition can be violated

occurs.[16] This creates a time interval Δ, just after a supervenient state has occurred, during which it is not in one of its associated base states. See Fig. 4.

Here (DET) no longer obtains. The mental state changes from M_1 to M_2 before the brain state changes from the realization class for M_1 to the realization class for M_2.

[16] Note that the condition will have to be slightly modified in a dualist context; see Sect. 5.3.

Thus during Δ the brain can be in a physical state outside of $f^{-1}(M_1)$, which violates (DFUN) and thus (DET). This situation is, however, compatible with probabilistic forms of determination, which we turn to now.

4.3 Probabilistic determination (PDET)

Probabilistic determination is a variant on determination: if two objects are indiscernible at the base level, then they will tend to be "nearly" indiscernible at the supervenient level.[17] We thus assume that it is meaningful to speak of the extent to which mental states are similar; that is, we assume **M** is a metric space. To formally define probabilistic determination, we first define a variant on indiscernibility, which we represent by the symbol "$\approx$". Since we only use it for the mental states in **M**, we specify our definition to that case:

> Define $x \approx_{\mathbf{M}} y$ to mean x and y are in states in **M** drawn from the same $\mathscr{P}_{\mathbf{M}} \in \mathbb{I}$, where $\mathbb{I}$ is the set of all informative unimodal probability distributions over **M**.

The assumption that $\mathscr{P}_{\mathbf{M}}$ is an informative unimodal distribution makes it likely that mental states drawn from it will be located in some specific region of **M**.[18] For example, consider a normal distribution $\mathscr{P}_{\mathbf{M}}$ with a relatively small variance (a "narrow" distribution) centered on a state of hunger in **M**, and sample it twice. The two resulting states M_1 and M_2 can be colloquially described as "probably similar," since they are probably from nearby states (95% of states sampled from a normal distribution will be within two standard deviations of its mean), but might not be. The parameters of $\mathscr{P}_{\mathbf{M}}$ will determine the degree to which two states are "probably similar." For example, if $\mathscr{P}_{\mathbf{M}}$ is a narrow distribution (low variance, as in the case just described) then x and y are more likely to be similar than if the distribution is broad (large variance).

We can now define probabilistic determination of mental states by physical states:

> (PDET) $\forall x, y \in \mathbf{D}$, if $x =_{\mathbf{P}} y$ then $x \approx_{\mathbf{M}} y$

If two people are physically indiscernible, then they are in states drawn from the same informative unimodal probability distribution, and are thus probably in similar mental states.

As with (DET), there is a functional formulation of (PDET). This function associates physical states in **P** with probability distributions over mental states in **M**. This is also known as a "stochastic map," which is a function where "[I]nstead of assigning a

[17] Kim mentions but does not develop the idea of stochastic supervenience (2006, p. 550).

[18] These are important qualifications: if we simply required the probability distribution to be non-uniform, this would be consistent with even the slightest departure from randomness. We will not attempt to formally define "informative distribution" in this paper, though the concept of a weakly vs. highly informative prior exists in the literature on Bayesian inference (Lemoine, 2019). Roughly speaking, a more informative distribution has lower variance than a strongly informative distribution. That the distribution be unimodal rules out cases where the distribution is informative but multi-modal, making it likely that one of a set of potentially very different mental states occurs. It may make sense to relax this restriction but we will not do so here.

unique element of Y to each element of X, it assigns a probability distribution on Y to each element of X".[19]

> (PDFUN) There exists a function $g : \mathbf{P} \rightarrow \mathbb{I}$, where for any $P \in \mathbf{P}$ and any $\mathscr{P}_{\mathbf{M}} \in \mathbb{I}$, $g(P) = \mathscr{P}_{\mathbf{M}}$ iff for every x such that Px, there is some M sampled from $\mathscr{P}_M$ such that Mx

The functional formulation makes the distinction between (PDET) and (DET) especially clear. While the function associated with (DET) maps physical states to unique mental states, the function associated with (PDET) maps physical states to unique probability distributions over mental states.

(PDFUN) implies that there can be some "slippage" in the link between physical and mental states. Consider a person's brain state as they drive their car, listening to the radio, with mountains in the distance, thoughts of an upcoming meeting in their head, and the smell of fast food lingering in the car. This corresponds to a specific complex mental state $M \in \mathbf{M}$. According to probabilistic determination, the brain state of this person does not completely specify this mental states. It is, rather, associated with *some version* of this complex mental state, drawn from a probability distribution centered at M. This is compatible with the slippage described in Fig. 4. Suppose the person decides to the turn the dial on the radio. This desire takes some time to register in the brain, and so even if their brain state is constant for a very short period of time, their mental state varies slightly, consistently with (PDET) but not (DET).

4.4 Empirical evidence might distinguish (DET) and (PDET)

(PDET) is not empirically equivalent to (DET), and so experimental evidence could in principle provide inductive support for one over the other.

To see this, consider a case of what we will call "observed indeterminism," where relatively constant brain states known to be associated with a particular type of mental state are measured alongside varying mental states of that type.[20] Suppose, for example, we place someone in a quiet room and ask them to report what they hear, and receive repeated reports of silence but occasional reports of hearing a tone, despite an absence of tonal stimuli or relevant activity in their auditory cortex.

If (DET) is true, any such such observed indeterminism between mind and brain should be due to measurement error (recall the discussion in Sect. 3). The occasional reports of hearing a tone in the example just described would then be instances of state measurement error. Note, moreover, that the average measurement error associated with a device can be estimated. A standard ruler is pretty good down to the millimeter level of resolution but not so good after that. Its average error rate is around $\pm$ a millimeter. In a similar way it is possible to estimate average error rates for behavioral reports, e.g. the standard amount people will lie, or get bored and confabulate, etc., in a given experimental situation. Thus, it should be possible to study whether the

[19] The quote is from John Baez's well-known blog, *Azimuth* ("Relative Entropy - Part 2", 7/2/2013). For a more formal discussion see Fullwood and Parzygnat (2021).

[20] There may be other ways to empirically distinguish (DET) and (PDET), but this is a relatively easy case to reason about.

variance in behavioral measurement in cases of observed indeterminism is what we'd expect relative to these baseline error rates.

If (PDET) holds, on the other hand, then any observed indeterminism is due both to measurement errors (which are inevitable), and to the probabilistic nature of the determination relation itself.

Thus, the empirical commitments of the two positions can be distinguished by considering the source of observed indeterminism in cases where behavioral measures vary despite relatively constant activity in relevant brain areas. We can distinguish two cases

(DET) $\longrightarrow$ measurement error only
(PDET) $\longrightarrow$ measurement error plus additional error arising from the probabilistic determination relation

Observe which patterns of errors occur, and then by a kind of *modus tollens*, take that pattern to support one relation over the other. If we only observe known measurement errors, then we have inductive support for (DET) as against (PDET). If additional errors are observed, then we have inductive support for (PDET) as against (DET).

Even if we could empirically distinguish (DET) and (PDET) in principle, it's not clear how viable this would be in practice. The experimental situation we have described is quite complex and would involve many sources of noise. Moreover, the smaller the variance associated with (PDET) is, the more difficult it would be to detect. But science has a way of surprising us, and so there might be a clever way to pursue a study of this kind. In fact, the data needed to pursue this kind of work may already exist, in which case a careful re-analysis of existing data would be enough to get started.

5 The argument for neutrality

So far we have considered the nature of NCC data and discussed how the data can be modeled by relations in the determination family, i.e. (DET) and (PDET). We now turn to premise 3. Our proposal is that we can think of (DET) as a kind of schema, which can be further specified in several ways, and thereby be shown to be compatible with multiple ontological positions. Similarly for (PDET). More specifically, to interpret (DET) and (PDET) in accordance with a position on the mind-body problem, one must specify:

(a) A domain **D** of objects
(b) An interpretation of "$=_{\mathbf{P}}$" and "$=_{\mathbf{M}}$"

Filling in (a) tells us what the basic entities of a position on the mind-body problem are; filling in (b) tells us what it means to consider those entities to be indiscernible. By identifying reasonable ways of filling in (a) and (b), we aim to show the plausibility of a threefold compatibility thesis: that the determination family of relations is compatible with physicalism (*P-Compatibilism*), with idealism (*I-Compatibilism*) and with dualism (*D-Compatibilism*).[21]

[21] Compatibility theses for other positions, such as neutral monism, could be developed in this framework, but for reasons of space we will not pursue them here.

There are two main lines of reasoning in our argument. First, we argue that (DET) is compatible with most forms of *substance monism*, i.e. "the doctrine that all concreta are of a common type" (Schaffer, 2018, §1.2 9). This captures the idea that the physical and the mental are just the same system understood in two ways. Some forms of substance monism (e.g. emergence) also allow for more complicated mental-physical relations, including relations that allow mental changes to occur without corresponding physical changes, so that the temporal alignment condition is no longer met. In light of this possibility, our view is more aptly described as follows: any form of substance monism is compatible with (DET) *insofar as it respects the temporal alignment condition*. For those forms of substance monism that fail to meet this requirement, we will show that they are still compatible with (PDET).

Second, we argue that *substance dualism* coheres best with (PDET), though it is logically compatible with both (DET) and (PDET). The reason it coheres better with (PDET) is that when we have two categories of substance, there are more opportunities for temporal alignment to be violated. If for example communication between the two types of substance is unreliable or noisy—or if agents can make decisions which occur immediately in the mental realm but take some time to manifest in the physical—we'd expect the temporal alignment condition to be violated at times, so that (DET) would not obtain. Instead we will argue that (PDET) obtains in these cases, insofar as indiscernible physical states determine mental states that are probably similar, but not necessarily identical.

Our discussion of neutrality is focused on the question of how many categories of being there are in relation to mind and body, which is arguably the heart of the mind-body problem. However, the neutrality we defend is even broader than this, encompassing other dimensions of the mind-body problem that will not be our central focus.

First, we won't describe modal variations on (DET) or (PDET) in any detail. Supervenience claims can be taken to be metaphysically or nomologically necessary (McLaughlin and Bennett, 2021, §3.1), and the indiscernibility relations can be defined as holding between persons in the same world or in different worlds (Kim, 1988, p. 130). These choices don't impact our argument; in fact, they strengthen it. They provide more ways to adapt the determination family of relations to specific metaphysical positions, without conflicting with empirical data. For example, depending on whether one wants to allow for zombies, they can set the modal operator on (DET) to be nomologically or metaphysically necessary (this case is discussed below).

Second, in the case of (PDET), there is an additional free parameter. Different pairs of physical twins are associated with different members $\mathscr{P}_{\mathbf{M}} \in \mathbb{I}$, that is, different informative probability distributions over mental states. These distributions will have different means (they are centered on different mental states in $\mathbf{M}$), but we assume that they all have the same variance. Thus variance (like modal force) is a free parameter of (PDET), which intuitively says *how* informative the $\mathscr{P}_{\mathbf{M}} \in \mathbb{I}$ are. Depending on what that variance is, mental states sampled using $\mathscr{P}_{\mathbf{M}} \in \mathbb{I}$ will tend to vary a lot or a little.

Third, there is the nature of the mind-body relation itself, e.g. whether it is a relationship of constitution or grounding or causation. This is sometimes referred to as a question of "metaphysical explanation" (Brenner et al., 2021). These discussions usu-

ally take (DET) as an implicit background assumption. After all, (DET) is a fairly weak relation, a largely neutral way of tracking states and state transitions. The task in discussions of metaphysical explanation is to explain *why* these state correlations occur. Metaphysical explanation spells out various possibilities for what explains them. Since we argue that the empirical data support (DET), the empirical data are compatible with whatever accounts of metaphysical explanation support (DET). That is, metaphysical explanation, like modality and category of being, are free parameters relative to what the empirical data support. Arguments along more or less the lines we have in mind are developed by Kriegel (2020). We discuss some forms of this dimension of the mind-body problem where relevant below, but do not make it a central topic here.

If we're right, the upshot of these points is that the determination family of relations—which is sufficient to capture all the varied data of cognitive science—is compatible not just with physicalism, idealism, and dualism, but also with a wide range of variants of those positions.

5.1 Physicalism

We begin our discussion of the neutrality of (DET) with respect to categories of being by defending P-Compatibilism, which is not controversial. Indeed, much of the enthusiasm associated with supervenience in the 1990s was associated with efforts to use it to define physicalism (though these efforts largely failed, in light of its neutrality; see McLaughlin and Bennett, 2021). Our main task in this section is to show that our framework straightforwardly accommodates the widely accepted thesis of P-Compatibilism.

As Crane and Mellor put it, "Physicalists believe that everything is physical..." (1990, p. 185). The position can be defined in our framework as follows:

(Physicalism) $\forall x(x \in \mathbf{D}_{\mathrm{P}})$

Since (Physicalism) only specifies a domain for the first-order quantifiers, it is logically consistent with (DET). What remains is to show that $=_{\mathbf{P}}$ and $=_{\mathbf{M}}$ can be reasonably interpreted in a physicalist manner. There are at least two ways to do so, depending on whether one opts for a description or a property reading of states.

Early uses of supervenience in the philosophy of mind (Davidson, 2001; Hellman & Thompson, 1977) do not assume the existence of properties. On these views, physical reality can be described in different ways relative to different theoretical vocabularies, and supervenience specifies a relationship between these different types of description. This suggests an interpretation of indiscernibility relations based on a description reading of states:

$x =_{\mathbf{P}} y$: x and y have identical descriptions in a physical vocabulary
$x =_{\mathbf{M}} y$: x and y have identical descriptions in a psychological vocabulary

Given this interpretation of indiscernibility, (DET) says that any two entities that have identical physical descriptions will have identical psychological descriptions. Since this version of (DET) does not assume an ontology of properties, it might be characterized as a "nominalist" or "nominalist-friendly" version.

The indiscernibility relations in (DET) can also be interpreted in terms of a property reading of states. On this approach, the indiscernibility relations are understood in terms of objects having all the same physical or mental properties[22]:

$x =_{\mathbf{P}} y$: the same members of a set **P** of physical properties are instantiated by x and y
$x =_{\mathbf{M}} y$: the same members of a set **M** of mental properties are instantiated by x and y

The version of (DET) based on this interpretation of indiscernibility corresponds to the standard definition of mental-physical supervenience in the literature.

Several objections can be made at this point. One is to try to define (Physicalism) in terms of (DET), in which case I-Compatibilism and D-Compatiblism are definitionally ruled out. This was the basis of efforts to replace "token physicalism" (Fodor, 1974) with "supervenience physicalism" (Kim, 2012), which roughly correspond to our (Physicalism) and (DET), respectively. The arguments of the next few sections can be understood as responses to that line of argument (which, as noted above, mostly failed, in light of this type of consideration: supervenience does not uniquely specify physicalism).

So we don't want to define token physicalism, or (Physicalism), in terms of supervenience, or (DET). But this leads to the second worry: is token physicalism even compatible with supervenience? In particular, (DET) mentions mental states, but it is not clear whether one can really speak of mental states if one takes physical particulars to be all there are (Schneider, 2012b). There are several ways to respond to this worry. One is to emphasize the nominalist reading of (DET), which in this context claims that entities that have identical physical descriptions have identical psychological descriptions, without reference to properties. Another approach is to flesh out the combination of (Physicalism) and (DET) in functionalist terms: the mental states in **M** are taken to be functional states, which are *realized* in the physical states in **P**. This position is often described as token identity together with a harmless type dualism or property dualism. The idea is that any token instantiation of a mental property at a time will be identical to a token instantiation of a physical property at that same time. However, the properties instantiated will not be identical: my hunger now is identical to my being in brain state 25 now, but hunger is not the same property as brain state 25.[23] This is also described as "non-reductive physicalism," insofar as type dualism precludes any simple form of reduction, e.g. classic Nagel style reduction via bi-conditional bridge-laws, which seems to require type-identity. Proposals along these lines are accommodated by the property reading of (DET), and are thus consistent with this framework.

So far we have covered the standard varieties of physicalism. We now turn to two influential views that are consistent with token physicalism but are not always classified as physicalist views. First, consider Chalmers' *naturalistic dualism* (1996, p. 123–8). This is the view that, in addition to physical properties, conscious creatures also instantiate non-physical properties. A prominent reason for accepting naturalistic dualism is the metaphysical possibility of *zombies*: creatures whose physical aspects

[22] Also see footnote 15.

[23] For critical analysis of these ideas see Latham (2003) and Schneider (2012a).

are exactly like those of humans but from which phenomenal consciousness is absent (1996, p. 94). Zombies can be accommodated by using the aforementioned strategy of "tuning the parameters" of the determination family: we simply let (DET) be nomologically rather than metaphysically necessary. With the modal parameter set in this way, (DET) is neutral about the mind-body relationships typical of the possible worlds to which zombies belong. One is then free to adopt naturalistic dualism. Note that nonreductive physcalists who are unhappy with zombies can make a similar move: they can set the modal parameter of (DET) to metaphysical necessity, thereby disallowing zombies. The choice makes no difference to the empirical data of cognitive science. Either way we have the same measurements of brain states and the same measurements of conscious states. Therefore, whether one allows or disallows zombies, one can coherently accept some member of the (DET) family of relations, thus making one's view consistent with cognitive science.

Second, consider emergentism, which construes mental states as emergent states (O'Connor & Wong, 2020). Some might argue that (DET) is incompatible with emergence, because emergentists tend to deny that mental states can be "read right off" physical states in the way (DET) suggests they can. In response to the worry, we will now argue that emergence in all its varieties can be shown to be compatible with (DET).

We can begin with epistemic varieties of emergence, which say that the higher-level properties of a system are emergent if they are difficult or impossible to predict on the basis of a thorough knowledge of its lower-level parts. We won't spend much time on this, because we are interested in the ontology of mental-physical relationships. However, it is clear that epistemic emergence is compatible with (DET), since an inability to predict what will happen at the mental level is compatible with actual determination relations.[24] Chaos provides a well known example of determinism without predictibility, since chaotic systems are mathematically deterministic, but in practice produce values that are difficult to predict (Devaney, 2018).

What about the metaphysical varieties of emergence? According to Wilson (2016), metaphysical emergence requires causal autonomy. That is, metaphysical emergence requires that mental states have causal powers distinct from those inherited from physical states. Some varieties of emergence, what Wilson calls "strong emergence" relations, require that mental states have *more* powers than are provided by their physical base, while with "weak emergence" mental states have *fewer* powers than are provided by their physical base. Wilson argues that all varieties of metaphysical emergence can ultimately be reduced to these two, which we thus focus on here.

For our purposes, it is sufficient to point out that our framework can accommodate both cases. To see this, note that the supervenient causal powers associated with emergence either are, or are not, consistent with the temporal alignment condition. If they are consistent, (DET) obtains. This is clear from Figs. 2 and 3 above. In those figures we could imagine mental states emerge from physical states but have unique causal powers (the desire to eat, read Hegel, doomscroll, etc.). These causal powers are consistent with (DET) so long as temporal alignment obtains: any time these powers

[24] In fact, according to Kim, supervenience is part and parcel of the very idea of emergence (2006, p. 550). If so, it might not even be necessary for us to spend time on metaphysical varieties of emergence, either.

give rise to a change in mental state from M to to M' there must also be a change in brain state from $f^{-1}(M)$ to $f^{-1}(M')$.

On the other hand, if emergent causal powers lead to violations of temporal alignment, (DET) no longer obtains, but (PDET) does. A case like this is described by O'Connor and Wong (2005), who argue that emergent properties E "arise from and are sustained by underlying microstructures" (p. 664), which can have genuine causal powers that only express themselves after a time delay. This violates temporal alignment and thus (DET). However, (PDET) still applies, since the violations of alignment only occur in a specific emergent property, like an act of will. Visual experience is still determined by visual cortex; auditory experience by auditory cortex, etc.[25] So a given brain state will still mostly determine a mental state, except for the part that corresponds to a person's state of will. On different occasions that same brain state would determine maximal mental states differing only in their volitional component, which would thus all be probably similar to each other in the sense defined in Sect. 4.3. So sameness of physical states entails probable similarity of mental states, in accordance with (PDET).

5.2 Idealism

We now argue for I-Compatibilism. According to idealism, "all things are mental" (van Inwagen, 2015, p. 55). Formally:

(Idealism) $\forall x(x \in \mathbf{D}_{\mathrm{M}})$

As with (Physicalism), it is easy to show that the conjunction of (Idealism) and (DET) doesn't lead to any logical contradiction. Consequently, insofar as one is able to identify sensible idealist interpretations for both formulae, one succeeds in establishing I-Compatibilism. That is our goal in this section.

We follow the distinction made by Chalmers (2020) between *anti-realist* and *realist* forms of metaphysical idealism. Anti-realist idealism is familiar from the history of philosophy, where Berkeley is the representative case. Its key claim is that "for any nonmental fact p about concrete reality, what it is for p to obtain is for appearances that p (or closely related appearances) to obtain" (Chalmers, 2020, p. 354). For a tree to exist it must appear. There is nothing to the tree beyond that. By contrast, realist idealists do not focus on appearances, but rather on a fundamental reality whose ontology is mental. So the perennial tree falling in the forest is a mental phenomenon, even if it is not "heard" by a person in the everyday sense. We can interpret (Idealism) both ways and show that (DET) is consistent with each.

Let's begin with realist idealism and focus on the version of it that takes the building blocks of reality to be microsubjects.[26] On this view, qualia are a fundamental feature of reality similar to mass and charge; the bearers of such fundamental qualia are microsubjects, which are (roughly) the mental counterparts of fundamental par-

[25] Husserl entertains something like this view. Cf. Yoshimi (2010), where the idea is described using the concept of "partial superveninece."

[26] This is what Chalmers describes as *micro-idealism* For another version of realist idealism, see footnote 28.

ticles.[27] They can be distinguished from macrosubjects, i.e. ordinary subjects like human beings. According to this view, the world is a massive configuration of qualia (mental states had by microsubjects), which grounds all remaining facts about the world, including those about macrosubjects (Chalmers, 2020, p. 359).

At first glance, it might seem difficult to interpret $\mathbf{D}_{\mathrm{M}}$, $=_{\mathbf{P}}$ and $=_{\mathbf{M}}$ in the spirit of this form of realist idealism. The natural choice of $\mathbf{D}_{\mathrm{M}}$ seems to be the set of all microsubjects. But since the facts about microsubjects' mental states ground everything, these facts also ground facts about microsubjects' physical states, which suggests that physical states supervene on mental states, rather than vice-versa. The problem disappears, however, if we focus on the domain relevant to discussions of mind-body correlations in cognitive science:

$\mathbf{D}_{\mathrm{M}}$: the set of all macrosubjects

Given this domain, **P** and **M** should be interpreted as containing the physical and mental states of macrosubjects, which on this view are ultimately grounded in micro-qualia at the fundamental level. The details of the micro-qualia are irrelevant to the macro-level. Whatever is going on at the level of fundamental physics, we still have asymmetric correlations between macro-physical states and macro-mental states, that are well explained by (DET). Thus we can understand $=_{\mathbf{P}}$ and $=_{\mathbf{M}}$ in the same way as we did in Sect. 5.1, as shared properties or descriptions. For example, if two macrosubjects are physically indiscernible and there is heightened activity in the visual cortex of one of the macrosubjects, there will be heightened activity in the other's visual cortex. (DET) then says that they will have indiscernible macro-level mental states. This is all consistent with a background ontology of micro-qualia.[28]

What about anti-realist idealism, or phenomenalism?[29] Recall that for anti-realist idealists the existence of a tree is simply a matter of someone experiencing a tree, and nothing more. A generic way of understanding this kind of view, which can be adapted to more specific versions, is to posit a set of all possible experiences as fundamental reality, and then to ground all ontology – including our ontology of mind and brain—on that set.

We can develop such an approach by first introducing a set of possible experiences **E**, which are something like Carnap's elementary experiences or *Elementarerlebnisse*.[30] **E** contains experiences of looking at other people, hearing what they say, observing measurements of their brains, etc. It also contains experiences directed at

[27] Chalmers takes a photon that has mental states to be an example of a microsubject. On Chalmers' view, micro-idealism is equivalent to *grounding micropsychism* which holds, among other things, that entities like photons have no fundamental properties other than mental ones (2020, pp. 359–360).

[28] For similar reasons this view is consistent with what Chalmers calls "cosmic idealism," where the cosmos as a whole is a mental entity which grounds the various constituent parts of reality. All the cognitive science stays the same, even if our background ontology changes. In fact these forms of idealism might even be considered forms of physicalism, if physicalism is understood as "whatever is consistent with final physics in its idealized form", a problem first raised by Hempel and sometimes referred to as "Hempel's dilemma" (Ney, 2008).

[29] Here we ignore the complicated logical relations that might hold between idealism and phenomenalism; see Chalmers (2020, pp. 355–356) for discussion.

[30] See, e.g., Carnap (1937, p. 12) for the term. The approach we describe is motivated largely by Husserl's account of the "constitution" of the objects of psychology and biology in the flux of conscious experience Gurwitsch, 1964, p. 2; quoted in (Yoshimi, 2010).

multiple people, as well as experiences of comparing them. We thus define persons as experienced macrosubjects, i.e. macrosubjects appearing in the *Elementarerlebnisse* that are members of **E**:

$$\mathbf{D}_{\mathrm{M}} = \{x |\ x \text{ is a macrosubject appearing in some member of } \mathbf{E}\}$$

The definition should *not* be read as presupposing that there are mind-independent macrosubjects, where some of these appear in members of **E** but others don't. Rather, to oversimplify things, for a macrosubject to exist just is for there to be an *Elementarerlebnis* in which the macrosubject seems to be experienced.[31] There can be mental states in which a person seems to be experienced even if the person turns out to be a figment of imagination—for this reason, nothing about mind-independence is presupposed by our definition.

There is a complex and interesting story one can tell about how macrosubjects—experienced organisms that appear as persisting bearers of physical and mental states—are "constituted" in rule-governed time-evolutions in **E**, but we will restrict ourselves to noting that *pairs* of macrosubjects can be given in experiences and perceived as indiscernible or not. Indiscernibility relations can then be spelled out as follows:

$x =_{\mathbf{P}} y$: x and y are experienced as being physically the same
$x =_{\mathbf{M}} y$: x and y are experienced as being mentally the same

Given this interpretation, (DET) becomes a thesis about patterns of experiences of indiscernibility.[32] The experiences of a scientist observing the relationship between brain states and mental states then becomes explicitly phenomenological, what Husserl called an "experience of psycho-physical conditionality" (Husserl, 1989, p. 78; also see Yoshimi, 2010). According to Husserl, if two organisms are experienced as having the same physical properties (as being physically indiscernible), they will also be experienced as having the same "sensibility" (as being sensorially indiscernible), for example.

Thus we have a phenomenology of determination, which can capture all the data of cognitive science in a framework that posits **E** as fundamental reality. Imagine a neuroscientist, Edith, reading questions to a participant in an fMRI machine. On the basis of a verbal report, Edith experiences the participant as being in the mental state of feeling tranquil. On the basis of her visual experience of the fMRI machine, Edith

Carnap's *Elementarerlebnisse* are closely related to the Gurwitsch's concept of a field of
Footnote 30 continued
consciousness, a "totality of co-present [phenomenal] data" (Yoshimi & Vinson, 2015), p. 104). Both Gurwitsch and Carnap were influenced by Husserl; see Rosado Haddock (2008) for discussion with respect to Carnap. Note that Husserl was not (at least on standard readings) either an anti-realist or a phenomenalist, but that his texts contain the resources for developing such a view.

[31] See Pelczar (2015) for a sophisticated account of how idealists can accommodate the existence of spatiotemporal objects.

[32] Husserl develops an account of the experience of indiscernibility along these lines in his early phenomenology of logic (his analysis of how universals and other formal and mathematical structures are constituted in experience). According to Husserl, we come to experience universals by noting that a group of particulars are indiscernible with respect to a series of pairwise comparisons (Kasmier, 2003) In the cases of interest here (comparisons of individuals for physical and psychological indiscernibility) the process is the same, though the comparisons are more complex, insofar as they are mediated by various kinds of instrumentation.

experiences the participant as being in a particular brain state. She also experiences the reported tranquility as being the result of ("conditioned by") the brain state, via an experience of psycho-physical conditionality. According to (DET), if there were to be another participant alongside the first one—such that Edith experiences the second participant as being in the same brain state as the first—she would expect similar behavioral reports of tranquility, consistently with her overall experience of psycho-physical conditionality. If she consistently had experiences like this she would have evidence for (DET).

Given that $\mathbf{D}_{\mathbf{M}}$, $=_{\mathbf{P}}$ and $=_{\mathbf{M}}$ can be coherently interpreted in the way just described, (DET) is compatible with anti-realist idealism.

Not everyone would be happy with our argument for I-Compatibilism. Some might think that there should be a clause against supervenience built into the definition of idealism. For example, they might follow Pelczar and take idealism to be "the view that facts about spacetime and its contents supervene on broadly phenomenological facts, but not vice versa" (Pelczar, 2015, p. 114). But such an anti-supervenience clause is consistent with our view, because the kind of supervenience that Pelczar has in mind is not the kind of supervenience we have in mind. Pelczar's formulation makes it clear that he is concerned with how mental facts relate to physical facts, where facts are construed non-idealistically, i.e. as facts about what things are rather than about what appearances things have in experience. But we are concerned with *idealistically construed* facts; more precisely, our attention is on how *experienced* mental facts relate to *experienced* physical facts. This is precisely why we chose to illustrate (DET) with the example of the neuroscientist operating the fMRI machine. Given this, the kind of supervenience that Pelczar rejects is not the kind of supervenience we affirm—one is free to reject the former while accepting the latter.

5.3 Dualism with (DET)

We now defend D-Compatibilism, which states that the determination family of relations is compatible with dualism.[33] Dualism is the view that mental substances are distinct from, but sometimes linked to, physical substances. Dualism has enjoyed a resurgence in recent years. It is said to be natural in the sense that it is widespread in human history and "seems correct to most people who have not steeped themselves in philosophical reflection" (Meixner, 2012, p. 38). Moreover, alternative positions face various *aporia*, and dualism itself can be defended by philosophical arguments. Swinburne, for example, argues that even when everything about a certain body is known, it is still conceivable for the person in possession of the body to remain unknown. This suggests that the body and the person are different objects (1997, pp. 148–150). Other arguments draw on emergentist and evolutionary considerations, describing the mind as an emergent substance that appears in evolutionary history as a way to guide organisms in non-deterministic ways (Hasker, 1982; Meixner, 2012).[34]

[33] "Dualism" in this section means "substance dualism." It is standard to treat property dualism separately (our discussion is in Sect. 5.1).

[34] For other contemporary approaches to dualism, see, e.g., Swinburne (2009), Zimmerman (2010), Lowe (2012) and Owen (2020). For surveys, see Robinson (2020) and Lycan (2009).

As noted earlier, dualism generally coheres better with (PDET), but it is still compatible with (DET). We thus examine (DET) in Sect. 5.3 and turn to (PDET) in Sect. 5.4.

Capturing dualism in our framework requires some additional machinery. First, whereas physicalism and idealism can be formalized using a single domain, we now need to distinguish $\mathbf{D}_{\mathrm{P}}$ and $\mathbf{D}_{\mathrm{M}}$, which are the distinct domains of physical and mental substances. Second, we must distinguish those physical entities associated with a mental life from those that are not. We refer to those physical entities with a mental life as $\mathbf{D}_{\mathrm{P}}^{*} \subset \mathbf{D}_{\mathrm{P}}$. Third, to capture the idea that when minds are embodied the two kinds of substances are linked, we need a function $\Phi : \mathbf{D}_{\mathrm{P}}^{*} \rightarrow \mathbf{D}_{\mathrm{M}}$ that associates each physical substance $x \in \mathbf{D}_{\mathrm{P}}^{*}$ with a mental substance $\Phi(x) \in \mathbf{D}_{\mathrm{M}}$.[35] For any $x \in \mathbf{D}_{\mathrm{P}}^{*}$ and $y \in \mathbf{D}_{\mathrm{M}}$ such that $\Phi(x) = y$, we say that x and y are the *accompanying substances* of each other. Notice that we do not specify that Φ be a surjective function; we do not require that every mental substance has an accompanying physical substance. This allows for the possibility of purely disembodied beings or for a mental substance to persist after its accompanying physical substance perishes. (Of course, cognitive science is restricted to the study of embodied beings, for the obvious reason that its measurements are physical and require a physical embodiment.) Using this apparatus, we can define dualism as follows:

(Dualism) $(\forall x \in \mathbf{D}_{\mathrm{P}}^{*})(\exists y \in \mathbf{D}_{\mathrm{M}})\ \Phi(x) = y$

The idea that physical substances determine embodied mental substances can be formulated for (Dualism) as follows[36]:

(DET′) $\forall x, y \in \mathbf{D}_{\mathrm{P}}^{*},\ \text{if } x =_{\mathbf{P}} y,\ \text{then}\ \ \Phi(x) =_{\mathbf{M}} \Phi(y)$

This says that for any two physical substances that are physically indiscernible, their accompanying mental substances are mentally indiscernible.[37] Since (DET′) is logically consistent with (Dualism), D-Compatibilism is true. This result makes sense, since (DET′) only tracks the relationship between pairs of indiscernible physical substances and pairs of indiscernible mental substances. It makes no commitments regarding the ontological basis of those correlations, so it is orthogonal to one's view about the ontological basis of those correlations, be it causal interaction, pre-established harmony or something else. It is also completely silent about disembodied beings. Such beings can exist or not (that is, we can require Φ to be surjective or not), and it makes no difference to cognitive science, which, in a dualist framework, only requires (DET′).

[35] For factors worth taking into account when defining such a function, see Kim (1988, §4). Also cf. the "causal pairing problem" in O'Connor and Wong (2005), which is used as an argument against dualism: "It is exceedingly odd that *particular* minds and brains form a lifelong 'monogamy' despite the absence of any apparent relational framework" (2005, p. 660).

[36] Such a dualist conception of supervenience was developed by Kim (1988, pp. 143–144), though Kim's discussion wasn't accompanied by a formal definition of substance dualism.

[37] Are we changing the topic by switching from (DET) to (DET′)? Not really. Although we described (DET′) as being different from (DET), we could focus on (DET′) alone and dispense with (DET). That is, we could require that all forms of supervenience specify two domains and a function Φ between them, and then for all but the substance dualist case assume that those domains are identical and that Φ is the identity function (cf. Kim 1988, pp. 144–145).

So dualism is consistent with (DET′). Despite this, their coherence is weak. Dualism must somehow guarantee that a modified temporal alignment condition is observed. The modified condition is this:

> (Modified Temporal Alignment Condition) For any physical substance x and any pair of mental states M_1 and M_2 such that (i) M_2 succeeds M_1 and (ii) $M_1 \neq M_2$, if $\Phi(x)$ goes from M_1 into M_2 at t, then there are $P_1 \in f^{-1}(M_1)$ and $P_2 \in f^{-1}(M_2)$ such that x goes from P_1 into P_2 at t.

However, most forms of dualism fail to provide this guarantee. A noisy channel could run without glitches for a time, but not forever. A free agent could in principle go for years without making any free choices, but in the long run free choices will be made. In fact if a necessity operator is prefixed, it seems (DET′) is not even consistent with most forms of dualism. Still, one could imagine forms of dualism that guarantee the condition is met, perhaps via *a priori* pre-established harmony, or in some other way. It may be, for example, that Owen's hylomorphic dualism (Owen, 2019) is an example that fits this case (if not, it falls under (PDET), discussed below).

5.4 Dualism with (PDET)

What happens when the temporal alignment condition is *not* observed? We argue that even if (DET) is violated in this scenario, (PDET) holds. There are several ways temporal alignment can be violated in a dualist framework, which generally involve unreliable relationships between physical and non-physical substances.

First, causal interactions involve probabilistic determination. Errors inevitably creep into any electrical signal, for example, so that the image shown on a television will often contain glitches and missing pixels. Similarly for Cartesian dualism: on Descartes' model, glitches would presumably occur at the pineal junction from time to time, altering some components of, say, your visual experience. Thus a given brain state would give rise to a *more or less specific* mental state: for example, a perception of a red apple with bits of color altered here or there where the mind-body transduction failed. That is, a given brain state would determine an informative probability distribution over mental states, in accordance with (PDFUN) and thus (PDET).

Though causation is the natural case to consider here, (PDET) could obtain for similar reasons in a parallelist or occasionalist framework, if for example a divine being mediating parallel correlations wished to introduce delays or noise for some reason.

These are fairly outlandish schemes, but existing positions on the mind-body problem can also be understood as forms of (PDET). We saw that O'Connor and Wong's version of strong emergence (which comes close to being a form of property dualism) explicitly allows causal powers at the supervenient level to express themselves at the physical level after a time delay. Owen's hylomorphic substance dualism (Owen, 2019) may also fit this case. Again, our goal is simply to show that exiting positions, whatever their details, are either compatible with (DET) or (PDET).

6 The metaphysical implications of cognitive science

As we've seen, (DET) coheres best with positions in the mind-body problem that meet the temporal alignment condition, in particular substance monism. If mental and physical states are states of the same system understood in different ways, it's natural that temporal alignment is observed and that (DET) obtains. (PDET) coheres better with cases where temporal alignment is not observed, e.g. cases where brain and mind communicate over a probabilistic channel, or where certain features of the mind can operate independently of the brain. Summing up:

(DET) ⟶ Substance monism (or any view consistent with TA)
(PDET) ⟶ Substance dualism (or any view inconsistent with TA)

Here, "TA" stands for the temporal alignment condition, and the arrows indicate coherence in the sense of Sect. 2.

We can combine these coherence relations with the fact that empirical evidence could provide some inductive support for (DET) over (PDET), or conversely (see Sect. 4.4). If the experiments or re-analyses described above were pursued, we could combine those results with the coherence relations described in Sect. 5, to provide evidence for or against different positions on the mind-body problem. More specifically:

Measurement error only ⟶ (DET) ⟶ Substance monism (or any view consistent with TA)
Measurement error + additional error ⟶ (PDET) ⟶ Substance dualism (or any view inconsistent with TA)

Thus, cognitive science is mostly metaphysically neutral. NCC data might tilt things towards (DET) and away from (PDET), or conversely, which in turn means scientists might be able to provide evidence for some positions on the mind-body problem and against others, though it will never—if our analysis is sound—settle the question.

7 Conclusion

Kriegel, developing his own arguments for the empirical equivalence of multiple positions on the mind-body problem, admits that the resulting neutrality is "in some ways disappointing" but goes on to argue that it should be "liberating" (2020, p. 275). Instead of saddling the scientific study of consciousness with the impossible task of resolving once and for all what consciousness is—unraveling the explanatory gap, dissolving the hard problem, etc.—we can simply admit our human limitations. The approach Kriegel prefers is therefore a "humbler approach" that no longer holds onto "the Enlightenment notion that science can account for every aspect of reality" (2020, p. 275).

We share Kriegel's diagnosis and attitude: it is liberating and humbling to face the deep mysteries of nature and admit that some of these mysteries are beyond the

reach of science. But we also depart slightly from Kriegel, arguing that some patterns of data could nudge us a bit, spicing our humility with a dash of Enlightened optimism. Science won't solve every mystery of mind, but it might narrow the playing field.

Acknowledgements We are grateful to participants in John Heil's NEH summer seminar on metaphysics of mind in 2009, who provided valuable feedback at the time this project was conceived. We also thank Jessica Wilson, and several anonymous referees for their constructive and helpful comments.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

A Maximal states

Consider a user-chosen language L and classify the predicates in L into mental and physical ones. We construct $\mathbf{M}$, the set of *maximal mental states*, in three steps:

(1) Take any sentence m^* that meets the following conditions[38]:

   (a) m^* is well-formed
   (b) m^* contains all the mental predicates in L
   (c) m^* is not a contradiction
   (d) m^* is not a negation
   (e) m^* either contains the free variable x or no free variable at all

(2) Define $\mathbf{M}^*$ to be the set of all such m^*
(3) Define $\mathbf{M} = \{\lambda x.m^* | \ m^* \in \mathbf{M}^*\}$[39]

$\mathbf{M}^*$ can be thought of as the set of all possible maximal statements about a person's mental state, written in the language of L. A lambda operator takes a statement and converts it into a predicate.[40] For example, suppose "H" is a mental predicate in L, whose intended interpretation is "is hungry". Then "$\lambda x.Hx$" converts the statement "Hx" into a predicate, to be interpreted as *being an x such that x is hungry*. $\mathbf{M}$ is the set of predicates formed this way from the maximal statements in $\mathbf{M}^*$.

[38] For steps (b) and (c), see Kim's treatment of "maximal properties" (1984, p. 158). For step (d), see Yoshimi (2007, p. 116). Here are some examples of m^*. Suppose there are only two mental predicates in L, namely F and G. Then m^* might be $Fx\&Gx$ or $Fx\&(\forall y(y \neq x \rightarrow \neg Gy) \rightarrow Gx)$, but not Fx or $\neg Fx\&\neg Gx$.

[39] To use an example in footnote 38, a possible member of $\mathbf{M}$ is $\lambda x.(Fx\&(\forall y(y \neq x \rightarrow \neg Gy) \rightarrow Gx))$.

[40] For an overview of lambda calculus, see Carpenter (1997).

The set of *maximal physical states*, **P**, can be constructed in an analogous manner. Once that is done, we take **M** and **P** to be the sets of states described in Sect. 2.

B Functional formulations of supervenience

In this appendix we show that (DET) is equivalent to (DFUN) and that (PDET) is equivalent to (PDFUN). We will use the following lemmas, which follow straightforwardly from the fact that maximal states fully specify what a person is like mentally or physically:

(PS1) $(\forall P \in \mathbf{P})(\forall x)(\forall y)(Px \ \& \ Py \rightarrow x =_{\mathbf{P}} y)$

(PS2) $\forall x \forall y(x =_{\mathbf{P}} y \rightarrow (\exists! P \in \mathbf{P})(Px \ \& \ Py))$

(MS1) $(\forall M \in \mathbf{M})(\forall x)(\forall y)(Mx \ \& \ My \rightarrow x =_{\mathbf{M}} y)$

(MS2) $\forall x \forall y(x =_{\mathbf{M}} y \rightarrow (\exists! M \in \mathbf{M})(Mx \ \& \ My))$

(DET) $\rightarrow$ (DFUN): We want to show that a function f associates any arbitrary $P \in \mathbf{P}$ with a unique $M \in \mathbf{M}$ such that for any x, if Px then Mx. That is, we want to show that $f = \{(P, M) \in P \times M | \ \forall x(Px \rightarrow Mx)\}$ is a function. Take an arbitrary $P \in \mathbf{P}$ and an arbitrary x such that Px. Consider another y such that Py. By lemma (PS1), x and y are physical twins, and so by (DET) they are mental twins. It follows by lemma (MS2) that there is a unique $M \in \mathbf{M}$ such that Mx. Since x is arbitrary, $\forall x(Px \rightarrow Mx)$.

(DFUN) $\rightarrow$ (DET): Take x and y that are physical twins. By lemma (PS2), there is some $P \in \mathbf{P}$ such that Px and Py. By the definition of f, there is some $M \in \mathbf{M}$ such that $f(P) = M$, Mx, and My. It follows by (MS1) that x and y are mental twins.

(PDET) $\rightarrow$ (PDFUN): We want to show that a function g associates any arbitrary $P \in \mathbf{P}$ with a unique informative unimodal probability distribution over **M**, that is, a unique $\mathscr{P}_{\mathbf{M}} \in \mathbb{I}$. Take an arbitrary $P \in \mathbf{P}$ and an arbitrary x such that Px. Consider another y such that Py. By lemma (PS1), x and y are physical twins, and so by (PDET) they are in mental states drawn from the same probability distribution $\mathscr{P}_{\mathbf{M}} \in \mathbb{I}$. Therefore, for any $P \in \mathbf{P}$ and x, there is a unique $\mathscr{P}_{\mathbf{M}} \in \mathbb{I}$ such that x is in a state M drawn from $\mathscr{P}_{\mathbf{M}}$.

(PDFUN) $\rightarrow$ (PDET): Take x and y that are physical twins. By lemma (PS2), there is some $P \in \mathbf{P}$ such that Px and Py. By the definition of the stochastic map g, there is a unique $\mathscr{P}_{\mathbf{M}} \in \mathbb{I}$ over **M** such that $g(P) = \mathscr{P}_{\mathbf{M}}$ and such that x and y are both in a mental state drawn from $\mathscr{P}_{\mathbf{M}}$. It follows by our definition of $\approx_{\mathbf{M}}$ that $x \approx_{\mathbf{M}} y$.

References

BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.

Brenner, A., Maurin, A.-S., Skiles, A., Stenwall, R., & Thompson, N. (2021). Metaphysical explanation. In Zalta, E. N. (Ed.), *Stanford encyclopedia of philosophy (winter 2021 ed.)*. https://plato.stanford.edu/archives/win2021/entries/metaphysicalexplanation/

Carnap, R. (1937). Testability and meaning–continued. *Philosophy of Science, 4*(1), 1–40.

Carpenter, B. (1997). *Type-logical* semantics. The MIT Press.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Chalmers, D. J. (2020). Idealism and the mind-body problem. In W. E. Seager (Ed.), *The Routledge handbook of panpsychism* (pp. 353–373). Routledge.

Christoff, K., Irving, Z. C., Fox, K. C. R., Spreng, R. N., & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: A dynamic framework. *Nature Reviews Neuroscience, 17*(11), 718–731.
Correia, F., & Schnieder, B. (Eds.) (2012). *Metaphysical grounding: Understanding the structure of reality.* Cambridge University Press.
Crane, T., & Mellor, D. H. (1990). There is no question of physicalism. *Mind, 99*(394), 185–206.
Davidson, D. (2001). Mental events. In *Essays on actions and events* (2nd ed., pp. 207–227). Oxford University Press.
Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. The MIT Press.
Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). *A modern introduction to probability and statistics: Understanding why and how.* Springer.
Dembski, C., Koch, C., & Pitts, M. (2021). Perceptual awareness negativity: A physiological correlate of sensory consciousness. *Trends in Cognitive Sciences, 25*(8), 660–670.
Devaney, R. L. (2018). *An introduction to chaotic dynamical systems*. CRC Press.
Dupré, J. (1993). *The disorder of things: Metaphysical foundations of the disunity of science*. Harvard University Press.
Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97–115.
Fullwood, J., & Parzygnat, A. J. (2021). The information loss of a stochastic map. *Entropy, 23*(8), 1021.
Gurwitsch, A. (1964). Field of consciousness. Duquesne University Press.
Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2019). *Cognitive neuroscience*: The biology of the mind (5th ed.). W. W. Norton.
Hasker, W. (1982). Emergentism. *Religious Studies, 18*(4), 473–88.
Haug, M. C. (2010). Realization, determination, and mechanisms. *Philosophical Studies, 150*(3), 313–330.
Hellman, G., & Thompson, F. W. (1977). Physicalist materialism. *Noûs, 11*(4), 309–345.
Husserl, E. (1989). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy. Second book: Studies in the phenomenology of constitution.* Kluwer.
Jurgens, A., & Kirchhoff, M. D. (2019). Enactive social cognition: Diachronic constitution & coupled anticipation. *Consciousness and Cognition, 70*, 1–10.
Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., & Mack, S. (2000). *Principles of neural science* (4th ed.). McGraw-Hill.
Kasmier, D. (2003). *Husserl's theory of a priori knowledge: A response to the failure of contemporary rationalism*. Doctoral dissertation, University of Southern California.
Kim, J. (1984). Concepts of supervenience. *Philosophy and Phenomenological Research, 45*(2), 153–176.
Kim, J. (1988). Supervenience for multiple domains. *Philosophical Topics, 16*(1), 129–150.
Kim, J. (2006). Emergence: Core ideas and issues. *Synthese, 151*(3), 547–559.
Kim, J. (2012). The very idea of token physicalism. In S. Gozzano & C. S. Hill (Eds.), *New perspectives on type identity: The mental and the physical* (pp. 167–185). Cambridge University Press.
Kirchhoff, M. D. (2015). Extended cognition & the causal-constitutive fallacy: In search for a diachronic and dynamical conception of constitution. *Philosophy and Phenomenological Research, 90*(2), 320–360.
Klein, C., Hohwy, J., & Bayne, T. (2020). Explanation in the science of consciousness: From the neural correlates of consciousness (NCCs) to the difference makers of consciousness (DMCs). *Philosophy and the Mind Sciences,1*(II).
Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience, 17*(5), 307–321.
Kriegel, U. (2020). Beyond the neural correlates of consciousness. In U. Kriegel (Ed.), *The Oxford handbook of the philosophy of consciousness* (pp. 261–276). Oxford University Press.
Latham, N. (2003). What is token physicalism? *Pacific Philosophical Quarterly, 84*(3), 270–290.
Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in bayesian analyses. *Oikos, 128*(7), 912–928.
Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1993). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). In *Neurophysiology of consciousness* (pp. 249–268). Birkhäuser.
Lowe, E. J. (2012). Non-Cartesian substance dualism. In B. P. Göcke (Ed.), *After physicalism* (pp. 48–71). University of Notre Dame Press.
Lycan, W. G. (2009). Giving dualism its due. *Australasian Journal of Philosophy, 87*(4), 551–563.

Mazzola, L., Isnard, J., Peyron, R., & Mauguière, F. (2012). Stimulation of the human cortex and the experience of pain: Wilder Penfield's observations revisited. *Brain, 135*(2), 631–640.
McLaughlin, B. P., & Bennett, K. (2021). Supervenience. In Zalta, E. N. (Ed.), *Stanford encyclopedia of philosophy* (summer 2021 ed.). https://plato.stanford.edu/archives/sum2021/entries/supervenience/
Meixner, U. (2012). The naturalness of dualism. In B. P. Göcke (Ed.), *After physicalism* (pp. 25–47). University of Notre Dame Press.
Melnyk, A. (2018). In defense of a realization formulation of physicalism. *Topoi, 37*(3), 483–493.
Metzinger, T. K. (2000). Introduction: Consciousness research at the end of the twentieth century. In T. K. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual questions* (pp. 1–12). The MIT Press.
Mittner, M., Hawkins, G. E., Boekel, W., & Forstmann, B. U. (2016). A neural model of mind wandering. *Trends in Cognitive Sciences, 20*(8), 570–578.
Murray, D. J. (1993). A perspective for viewing the history of psychophysics. *Behavioral and Brain Sciences, 16*(1), 115–137.
Ney, A. (2008). Defining physicalism. *Philosophy Compass, 3*(5), 1033–1048.
O'Connor, T., & Wong, H. Y. (2005). The metaphysics of emergence. *Noûs, 39*(4), 658–678.
O'Connor, T., & Wong, H. Y. (2020). Emergent properties. In Zalta, E. N. (Ed.), Stanford encyclopedia of philosophy (summer 2020 ed.). https://plato.stanford.edu/archives/sum2020/entries/properties-emergent/
Owen, M. (2019). Neural correlates of consciousness and the nature of the mind. In M. P. Guta (Ed.), *Consciousness and the ontology of properties* (pp. 241–260). Routledge.
Owen, M. (2020). Aristotelian causation and neural correlates of consciousness. *Topoi, 39*(5), 1113–1124.
Pelczar, M. (2015). *Sensorama: A phenomenalist analysis of spacetime and its contents*. Oxford University Press.
Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience, 38*, 433–447.
Raven, M. J., (Ed.), (2020). *The Routledge handbook of metaphysical grounding.* Routledge.
Rosado Haddock, G. E. (2008). The young Carnap's unknown master: Husserl's influence on *Der Raum and Der logische Aufbau der Welt*. Ashgate.
Robinson, H. (2020). Dualism. In Zalta, E. N. (Ed.), *Stanford encyclopedia of philosophy* (fall 2020 ed.). https://plato.stanford.edu/archives/fall2020/entries/dualism/
Schaffer, J. (2018). Monism. In Zalta, E. N. (Ed.), *Stanford Encyclopedia of Philosophy*. Winter 2018 edition. https://plato.stanford.edu/archives/win2018/entries/monism/
Schneider, S. (2012a). Non-reductive physicalism cannot appeal to token identity. *Philosophy and Phenomenological Research, 85*(3), 719–728.
Schneider, S. (2012b). Why property dualists must reject substance physicalism. *Philosophical Studies, 157*(1), 61–76.
Schroer, R. (2011). Can determinable properties earn their keep? *Synthese, 183*(2), 229–247.
Swinburne, R. (1997). *The evolution of the soul* (revised ed.). Oxford University Press.
Swinburne, R. (2009). Substance dualism. *Faith and Philosophy, 26*(5), 501–513.
Tahko, T. E. (2011). In defence of Aristotelian metaphysics. In T. E. Tahko (Ed.), *Contemporary Aristotelian metaphysics* (pp. 26–44). Cambridge University Press.
Tal, E. (2020). Measurement in science. In Zalta, E. N. (Ed.), *Stanford encyclopedia of philosophy* (fall 2020 ed.). https://plato.stanford.edu/archives/fall2020/entries/measurement-science/
Turnbull, M. G. (2018). Underdetermination in science: What it is and why we should care. *Philosophy Compass, 13*(2), e12475.
van Inwagen, P. (2015). *Metaphysics*. Westview Press.
Wilson, E. O. (1999). *Consilience: The unity of knowledge*. Vintage Books.
Wilson, J. (2016). Metaphysical emergence: Weak and strong. In T. Bigaj & C. Wüthrich (Eds.), *Metaphysics in contemporary physics* (pp. 345–402). Brill Rodopi.
Wilson, J. (2018). Grounding-based formulations of physicalism. *Topoi, 37*(3), 495–512.
Wilson, J. (2021). Determinables and determinates. In Zalta, E. N. (Ed.), *Stanford encyclopedia of philosophy* (spring 2021 ed.). https://plato.stanford.edu/archives/spr2021/entries/determinate-determinables/
Yoshimi, J. (2007). Supervenience, determination, and dependence. *Pacific Philosophical Quarterly, 88*(1), 114–133.
Yoshimi, J. (2010). Husserl on psycho-physical laws. *New Yearbook for Phenomenology and Phenomenological Philosophy, 10*, 25–42.

Yoshimi, J. (2012). Supervenience, dynamical systems theory, and non-reductive physicalism. *The British Journal for the Philosophy of Science, 63*(2), 373–398.
Yoshimi, J., & Vinson, D. W. (2015). Extending Gurwitsch's field theory of consciousness. *Consciousness and Cognition, 34*, 104–123.
Zimmerman, D. (2010). From property dualism to substance dualism. *Aristotelian Society Supplementary Volume, 84*(1), 119–150.