Preprint. Full version in Synthese 203 (186), 1-27 (2024. DOI: 10.1007/s11229-024-04638-4

On the probabilistic character of irreducible mental causation

Dejan R. Dimitrijević[1]

Abstract It has recently been remarked that the argument for physicalism from the causal closure of the physical is incomplete. It is only effective against mental causation manifested in the action of putative mental forces that lead to acceleration of particles in the nervous system. Based on consideration of anomalous, physically unaccounted-for correlations of neural events, I argue that irreducible mental causation whose nature is at least prima facie probabilistic is conceivable. The manifestation of such causation should be accompanied by a local violation of the Second Law of thermodynamics. I claim that mental causation can be viewed as the disposition of mental states to alter the state probability distribution within the nervous system, with no violation of the conservation laws. If confirmed by neurophysical research, it would indicate a kind of causal homogeneity of the world. Causation would manifest probabilistically in both quantum mechanical and psychophysical systems, and the dynamics of both would be determined by the temporal evolution of the corresponding system state function. Finally, I contend that a probabilistic account of mental causation can consistently explain the character of the selectional states that ensure uniformity of causal patterns, as well as the fact that different physical realizers of a mental property cause the same physical effects in different contexts.

Keywords Causal closure of the physical · Maxwell's demon · Mental causation · Probability distribution · Second Law of thermodynamics

1 Introduction

The main problem facing all dualistic descriptions of the immaterial mind, be it substance or property dualism, is the explanation of mind-body interaction. It stems from the principled impossibility of direct,

[1] University of Niš, Faculty of Sciences and Mathematics, Department of Physics, Višegradska 33, 18000 Niš. Email: dimke68dd@gmail.com

physical observation of mental properties. All we can observe are physical effects, such as body movements caused by force. Each of these effects can be viewed and studied as a series of instantiations of physical properties, i.e., physical events, and described entirely by means of physics.

Mental properties, as understood by interactionist dualists, on the other hand, can only be perceived from the first-person perspective. In a dualistic worldview, the subject of perception is the immaterial, nonphysical mind. To the extent that they are irreducible, mental properties have no common features with physical properties. All our measuring instruments are constructed on the principles derived from the laws of physics and can only be used to measure physical quantities. We lack the theory of mental causation that would allow us to individuate mental properties based on their causal power and explain their relations with physical properties.

This state of affairs has led most members of the scientific community to support the metaphysical doctrine of physicalism in recent decades. They share a widespread conviction that it is rational to base one's ontological commitments on the methods of natural sciences, especially physics, and that these methods imply a physicalist worldview (Dowell, 2006; Stoljar, 2021), according to which there is nothing "over and above" the physical in the world. The mainstay of physicalism is the argument from causal closure of the physical (CCP), which rests on the metaphysical thesis that all physical events that have causes have sufficient physical causes[2]. If we trace the causal chain from any event back into the past, according to CCP we will never have to leave the physical domain.

However, some serious objections to the validity of the CCP argument have appeared in the literature recently[3]. The most serious among them refer to its incompleteness, i.e., the fact that it contains a hidden premise that begs the question in favor of physicalism. As a result, the attempts of interactionist dualists to formulate a rational, intelligible account of mental causation gain importance. My aim is not to offer a comprehensive account of this type here. The purpose of this paper is to defend the thesis that the interactionist model of mental causation must be at least prima facie probabilistic, as well as to point out some of its general features and consequences.

The paper is organized as follows. In Section 2, I show that CCP, as the crucial argument in favor of physicalism, is question-begging because it contains a hidden premise that reduces every observable manifestation of causation to the physical. Section 3 aims to present the thought experiment with Maxwell's demon and the dilemmas that its consideration imposes from the point of view of mental

[2] Different presentations of the argument from the causal closure principle can be found in Crane (1995), Jackson (1996), Spurrett & Papineau (1999), Loewer (2007), Papineau (2001, 2013), Melnik (2003), and Kim (2005).
[3] Some of the objections to CCP are presented in Lowe (2000, 2006, 2008), Bishop (2010, 2012), Gibb (2010, 2015), Tiehen (2015), Saad (2018), and Dimitrijević (2020).

causation. In Section 4, a basic idea is outlined upon which an intelligible probabilistic, interactionist account of mental causation that does not fall under the constraints of CCP can be constructed. The immaterial mind acts analogous to Maxwell's demon: by causing anomalous, physically non-accounted for correlations of neural events, where no force acts on the components of the system. Section 5 discusses the features of an interactionist account of mental causation, based on the fact that anomalous correlations imply that a change in the probability distribution of states takes place within the brain and nervous system. This causes the redistribution of energy, momentum, and other conserved quantities without altering their total amount in the system. The discussion expands on the argumentation given in Dimitrijević (2020), whose key features are presented in a contracted form. In Section 6, I argue that, if an interactionist, probabilistic account of mental causation turns out to be correct, the analogies between the state descriptions of quantum mechanical and psychophysical systems, as well as between their respective causal mechanisms, would be overwhelming. This would suggest a kind of causal unity of nature, with the probability of the state of the system as the central concept. The section ends with a comparative discussion of the presented dualistic account of mental causation with the analogous physicalist account proposed by Papineau (2013). Section 7 contains a summary of the discussion.

2 Causal closure of the physical

The central premise of the argument from causal closure of the physical is the causal closure principle, which essentially claims that every physical effect that has a cause has a physical cause. Here, the relata of causation are events in Kimean sense – instantiations of properties of particulars such as particles or fields at a time. The most commonly used form of the argument from causal closure of the physical (CCP) is obtained when this principle is supplemented with two more premises[4]. The first of them simply states the fact that mental causes have physical consequences, i.e., that they are causally efficient. The second premise excludes systematic overdetermination, requiring that if a physical effect has a physical cause at time *t*, then no other event can be a cause of the same effect at time *t*. The conclusion of the argument is that mental events that have physical effects are identical to physical events. The argument is extremely important because, if successful, it virtually eliminates the possibility of irreducible mental causation. CCP is the main reason for most supporters of the ontological doctrine of physicalism to believe that there is nothing "over and above" physical properties and events in the world.

[4] This is the form of the argument proposed by its most influential proponents, Papineau (2001) and Kim (2005).

Strong inductive support for the causal closure principle, and consequently for CCP, was provided by two inductive arguments formulated by Papineau (2001). The first one, the argument from fundamental forces, rests on an analogy from physics. The fact that all physical forces can be reduced to a few fundamental conservative forces implies, according to this argument, that this is probably also the case with the supposed special, mental forces: they should be reducible to the composition of fundamental conservative physical forces. Thus, the argument from fundamental forces implies that since every isolated physical system must be conservative, this must also be the case when mental causation is manifested in the system. In effect, this is an application of the law of conservation of energy to living organisms and their nervous systems.

The second Papineau's inductive argument is known as the argument from physiology. It claims that since physiological research conducted in the last two hundred years has revealed no trace of special, non-physical forces, they probably do not exist. If there were non-physical forces, they would be made apparent by causing anomalous accelerations of particles in the system – meaning, the accelerations that were not accounted for by the action of physical forces. The absence of such accelerations implies that all processes within the nervous system can be attributed to the action of known physical forces, so the causal closure principle is true. Since all conservative forces can be expressed as negative gradients of the corresponding potential energy, this argument is equivalent to the claim that non-physical forms of energy probably do not exist because no manifestation of non-physical energy has ever been observed.

There are numerous objections to Papineau's arguments[5], although their influence remains considerable. A particularly serious line of objections shows that these arguments are incomplete: they contain a hidden metaphysical premise that specifies the mechanism of causation. Papineau takes into consideration only those causes and effects that are well known and nomologically described in the *physical* world. He tacitly identifies *all* causes with the action of conservative Newtonian forces and the effects with the acceleration caused by these forces. Consequently, a hypothetical mental causation should manifest itself with the same features that we traditionally observe in physical causation. Thus, our measuring devices should be able to detect some mental force, be it deterministic or indeterministic, acting on the particles of matter in the nervous system and causing them to anomalously accelerate in accordance with Newton's laws of motion. The work of that force must raise the overall energy content of the system. Only if we could not attribute the increase in the system's energy to any of the known physical forces, according to Papineau, would we be forced to conclude that there are hitherto unknown, presumably mental causes of this anomalous acceleration.

[5] Some of the objections to Papineau's arguments can be found in Gillette & Witmer (2001), Montero (2003), Bishop (2012), and Garcia (2014).

The described view of the potential action of the non-physical mind is typical of most physicalist writers. In this regard, I am grateful to an anonymous reviewer for pointing out a characteristic quote by Daniel Dennett from his book "Consciousness Explained" (1991, p.35), in which he, in discussing nonphysical minds, writes:

> No physical energy or mass is associated with them. How, then, do they get to make a difference to what happens in the brain cells they must affect, if the mind is to have any influence over the body? A fundamental principle of physics is that any change in the trajectory of any physical entity is an acceleration requiring the expenditure of energy, and where is this energy to come from? It is this principle of the conservation of energy that accounts for the physical impossibility of "perceptual motion machines," and the same principle is apparently violated by dualism.

Although the mechanism of forces causing accelerations is perfectly consistent with physical causation, there is no indication that it relates to mental causation. A tacit acceptance of this mechanism is tantamount to including a hidden premise that ensures that causal efficacy is limited to physical properties, thus making Papineau's arguments question-begging in favor of physicalism. The same conclusions can be applied to the causal closure principle, which was formulated with the same causal mechanism in mind. This principle ensures the cogency of CCP only if it is supported by a hidden premise, which limits the causal efficacy to physical properties, thus begging the question in favor of physicalism. This was convincingly demonstrated, for example, by Bishop (2012) and Gibb (2010, 2015). Gibb specified (2010, p. 374) that in fact two further premises are needed to make the argument complete. *Physical affectability* requires that something can affect a physical system either by affecting the amount of energy or momentum, or by redistributing these quantities within a system. *Redistribution* specifies that the redistribution of energy and momentum cannot happen without supplying energy or momentum to the system. These premises describe the effects of a physical force operating in a physical system, which follow from the laws of dynamics.

A physical force can alter the state of a physical system either by causing the particles that make up the system to accelerate, thus changing the total energy and momentum, or by redistributing energy and momentum within the system. In the former case, the failure to attribute the acceleration of particles within a system to the known physical forces would result in a violation of the conservation laws, especially the principle of conservation of energy. In the latter case, the redistribution of energy and momentum within the system without changing their total amount, i.e., without the work of a force being done, would indicate a violation of the Second Law of thermodynamics. Such a force would be able to cause the transfer of heat from a colder to a warmer body without an expenditure of energy, which would

create the perpetuum mobile of the second kind. The inevitability of violating the Second Law of thermodynamics in the case of a violation of the *Redistribution* premise will be discussed in the next section. In the case of purely physical forces, these consequences of the laws of dynamics are straightforward. The transference theory of causation, which equates a physical effect with a change in a conserved quantity of a body or physical system (Dowe, 2000), was posited with these laws in mind. According to this theory, the claim that *x* causes *y* through the mediation of force $\vec{F}$ is equivalent to the claim that energy and momentum are transferred from *x* to *y*. Gibb claims that "*Physical affectability* and *Redistribution* are both entailed by this theory of causation" (2010, p. 376). However, although popular among the philosophers of science, the transference theory is far from generally accepted, not least because its applicability to quantum mechanical systems is doubtful.

Papineau's arguments explicitly equate all forces with causes and acceleration with their effect. They require that the effects of putative mental forces be nomologically equal to the effects of physical forces. Without this specification of the mechanism of causation, the principle of causal closure of the physical would remain an empty claim. According to that mechanism, an observer could verify the manifestation of mental causation only by the presence of anomalous accelerations in the nervous system, especially the brain, which are not accounted for by any of the known physical forces. In other words, mental forces would manifest their presence by adding energy to the system, thereby violating the law of conservation of energy. This is nothing but a generalized mechanism of physical causation; thus, trying to fit mental causation into this picture begs the question in favor of physicalism.

As an illustration of this line of thought, Cucu & Pitts (2019) recently published an indicative analysis of the physical basis of dualistic accounts of mental causation that try to satisfy the requirements of conservation laws[6]. They criticize several dualistic accounts for their inconsistency with the basic laws of physics, such as the first Noether's theorem, its converse, and the locality of field physics. The conclusion of the analysis is that the interactionist dualist is left with only two options. The first one is to accept the "conditionality response", according to which the energy in the brain is not conserved if the mind acts on the system, which means that the laws of conservation are applicable only in the absence of mental causation. For dualists who accept Papineau's arguments according to which this possibility is very unlikely, the authors leave only one chance: accepting one of the quantum-mechanical approaches. The advantage of the latter approaches is – and I emphasize this – that they imply the validity of conservation laws, since they are based not on the effect of force, but on the redistribution of system properties due to one of the proposed quantum-mechanical effects. However, the basis of Cucu & Pitts'

[6] I thank an anonymous reviewer for bringing this work to my attention.

inference is the explicit assumption that the effect of mind on matter can take place exclusively through the action of force, i.e., the exchange of conserved quantities. In their exact words, “minds produce and/or destroy energy and momentum at some times and places” (p. 104). In this way, the authors actually make the hidden premise from the previous consideration explicit. This is the result of their full conviction that no other mechanism of action of mind on matter is conceivable.

There is no reason, however, for an interactionist to accept that the action of mental causes should be manifested by the acceleration of particles in the system. After all, if the mind is an immaterial thing, as the dualists claim, then it is difficult to imagine that physical properties such as energy, momentum or any other, can be attributed to such an entity. As Averil & Keating famously ascertained, “there is no way of specifying the state of a non-physical thing in terms of the variables of physics” (1981, p. 105). It is even more difficult to believe that mental causation can be seen as the transfer of these properties from mental to physical states. If anything, Papineau’s argument from physiology strongly demonstrated that there is no empirical evidence that this transfer occurs. A direct consequence of accepting this argument is that either there is no irreducible mental causation, or some entirely different mechanism is responsible for the causal effect of the immaterial mind on the body.

It is precisely at this point that the key difference between the physicalist and dualistic view of causality emerges. The dualist must explicitly reject the thesis of *Redistribution*, as suggested by Gibb (2010, p. 379), and accept the only remaining option, however exotic it may seem. Namely, she may examine the possibility that mental causation is manifested by the redistribution of energy and momentum without doing work, therefore without changing the total amount of energy and momentum of the system. This approach, as we have seen, is used in virtually all quantum-mechanically based accounts of mental causation. It implies commitment to conservation laws, but at the same time it has a price that Gibb did not foresee: the rejection of the Second Law of thermodynamics. This is not a choice that the majority of physicists and philosophers can easily agree with, so the persuasiveness of such an unusual thesis would be greatly enhanced by citing an example in which causality manifests itself in the proposed way. It turns out that such an example has not only existed for more than a century but has been the subject of heated discussions and controversies among physicists for just as long. It is about Maxwell's famous thought experiment (Maxwell, 1871, pp. 308-309), in which the idea of a demon capable of violating the Second Law of thermodynamics was born. The next two sections are devoted to a discussion of this thought experiment, the possibility of violating the Second Law, and its implications for the problem of mental causation.

3 Second Law of thermodynamics and Maxwell's demon

One of the central concepts of thermodynamics is entropy – the measure of disorder in a physical system. The Second Law of thermodynamics specifies that the total entropy of an isolated physical system never decreases during spontaneous processes, which stems from the tendency of the system towards thermodynamic equilibrium. If entropy is decreased in some part of a physical system, that effect must be compensated in other parts of the system so that the net entropy change of the system is zero during reversible processes or positive during irreversible processes.

Now, the energy and momentum are not being randomly redistributed in the neural systems of living organisms. My decision to raise my hand or to indulge in drafting a scientific paper gives rise to a series of highly ordered, directed, and coordinated neural and muscular events, which seemingly result in increasing order in a physical system. The effect of my decision boils down to the significant redistribution of conserved quantities since the flow of energy, momentum, and charge of many particles takes place in the nervous and muscular systems. This redistribution is governed by the action of forces. If there are irreducible mental forces, this is where we should expect their manifestation: the immaterial mind will use them to achieve the desired physical effect. But, as we have seen, their action will inevitably be accompanied by the occurrence of unaccounted-for energy and other conserved quantities. The fact that it is hard to see how the detection of these suddenly appearing quantities could elude us for so long, despite our best efforts, gives strength to Papineau's argument from physiology. Consequently, the physicalist account of this causation, based on the CCP, effectively assumes that there is nothing in mental causation over and above the action of physical forces and their physical effects. According to physicalism, mental agency is actually part of this causal chain in one way or another, so not only the First Law but also the Second Law of thermodynamics must be preserved. This means that the decrease in entropy that accompanies the intentional actions of a conscious subject must be offset by its increase elsewhere in the system, so that the total entropy increases.

The physicalist explanation of mental causation rests, therefore, on the assumption that the causal action of the hypothetical immaterial mind is excluded because it would have to be manifested by the action of exotic mental forces of a Newtonian character, the existence of which is not consistent with empirical records. However, in Dimitrijević (2020), an interactionist account of mental causation is proposed, based on the idea that redistribution of conserved quantities is accompanied by a local decrease in the entropy of selected subsystems within a neural system, where the selection is performed by the immaterial mind, without the action of force and without any work done. That way, the total amount of energy and momentum in the system remains unchanged during the causal process, i.e., no conservation

laws are violated. The Second Law is violated by the actions of the mind, as predicted by Morowitz (1987), since the actions of a conscious mind are non-random and intentional and therefore go against the general tendency of physical systems towards disorder. The very nature of human creativity boils down to predetermined actions aimed at increasing the regularity of the system and thereby reducing its entropy. In a physicalist view, the mind is seen as either reducible to or realized by the physical properties of the brain, so that the decrease in entropy is more than compensated for by its increase in various heat-producing dissipative processes. If, however, the mind is a non-physical entity, the decrease in entropy is uncompensated, so the Second Law must be violated in the parts of the nervous system where the mental interacts with the physical. In this way, locating the parts of the nervous system where anomalous correlations and the consequent violation of the Second Law occur is the best way to complete the search, started by Descartes, for the elusive interface between mind and body. Simultaneously, such disturbances of the state of equilibrium lead to a spontaneous tendency of the system to return to it. A local decrease in entropy results in gradients of various physical quantities in the system and, consequently, in physical forces that tend to bring the system back into equilibrium. The aforementioned account that I will extrapolate here is based on the idea that the mind creates small-scale correlations of neural processes by inducing small fluctuations in the probability distribution, which in turn leads to a redistribution of physical quantities, to produce significant behavioral effects. This is similar to what Eccles (1980, 1987) and Popper & Eccles (1977) had in mind when they argued that the finely tuned structure of the brain enables small perturbations to have macroscopically significant effects.

There is a striking analogy between the described effect of the mind on the body and the way Maxwell's demon affects a thermodynamic system. In Maxwell's famous thought experiment, an insulated container full of gas at a uniform temperature is divided into two equal chambers, A and B, by an impenetrable barrier. The demon opens the hole in the barrier only to faster molecules passing from B to A and slower molecules passing from A to B, thus creating a temperature gradient without doing any work – contrary to the Second Law. Maxwell conceived this thought experiment to show that the Second Law is statistical in nature and can be applied only when dealing with masses of matter and not with individual molecules. However, many physicists found the possibility of violating the Second Law quite unsettling, which gave rise to numerous attempts to exorcise the demon by showing that no device can be contrived that would operate in the way Maxwell's demon does, i.e., that *perpetuum mobile* of the second kind cannot be constructed[7]. These attempts to save the Second Law from the demon are thoroughly analyzed by Earman & Norton (1998, 1999), who concluded that all of them presuppose that the demon is a thermodynamic system already governed by the Second Law, which means that the combined

[7] For the most important attempts at exorcising the demon, see Smoluchowski (1912), Szilard (1929), Brillouin (1953), and Landauer (1961).

container-demon system must also be a thermodynamic system governed by the Second Law. So, in effect, no exorcism is needed because the validity of the Second Law is presupposed.

The situation changes decisively if the demon is not a thermodynamic system, as is commonly understood, but an intelligent, immaterial agent. In that case, it does not interact with the gas and can be considered to be outside the physically isolated system of gas in the container. As convincingly demonstrated by Earman & Norton, the Second Law then cannot be applied to the demon, and hence to the combined gas-demon system, unless an independent postulate is found to ensure it. Since the nature of such a postulate is unclear, the actions of the demon can be interpreted by an interactionist as manifestations of the causal power of an immaterial mind, which operates through the redistribution of momentum and energy in the system without altering their total amount. Redistribution comes about without the expenditure of work by the mind and without forces operating on particles and causing accelerations. An outside observer, unaware of the existence of the demon, infers that there are only physical forces in the system but that a physically unexplainable correlation occurs in the molecular dynamics. Such an observer will only be able to detect an anomalous, seemingly spontaneous redistribution of molecules in the container based on their speed. All our instruments and measuring techniques rely on detecting the action of purely physical forces, so the direct study of the mechanism of causation responsible for this anomalous correlation is beyond the reach of the observer. Upon analysis, she will only be able to deduce that an unobservable agent causes the redistribution by controlling boundary conditions in the system – specifically the barrier between the chambers. The agent imposes selection rules, which increase the probability of finding faster molecules in A and slower ones in B, thus decreasing the entropy in the system. Alternatively, the observer may model the gas dynamics by finding a functional dependence of the *a priori* probability of physical states A and B, realized in this simple example by the corresponding chambers, on the parameters of the gas molecules. In other words, the observer could assume the existence of a new probability law expressed by a modified distribution function of molecules by speed, whose unexpected asymmetry would have to be ensured by a hidden parameter. The nature of this hidden parameter would remain a mystery to our observer, as it would not correspond to any observable physical quantity in the system since its origin is in the mere act of choice of the immaterial demon. Regardless of these interpretive difficulties, both approaches would provide equivalent, mathematically correct descriptions of gas dynamics and allow the observer to establish at least approximate nomological relations between the physical parameters of the gas and the parameters of the system set by the demon. It should be noted that these relations can only be probabilistic since the nature of the system and boundary conditions are such that there is no way to know with certainty the

physical state of an individual molecule at a particular moment in time; instead, only the state of the entire ensemble of molecules can be the subject of prediction.

4 Maxwell's demon and interactional dualism

The brain and nervous system are certainly immeasurably more complex than a simple collection of molecules in a container. Still, there are striking analogies between the anomalous correlations caused by the demon in the container and those we observe in the neural systems associated with the process of executing our conscious decisions. These analogies lead us to suppose that the causal effect of the immaterial mind on the components of the nervous system can be modeled after the effect of Maxwell's demon on gas molecules. Lowe, one of the most prominent proponents of substance dualism, mentions anomalous correlations similar to these in his dualistic account of mental causation (2000, 2006, 2008). Lowe rejects the assumption that the causal power of mental properties must be manifested through anomalous accelerations in the system and recognizes the crucial role of correlations in neural processes. He argues that the convergence of neural events is a formal property of causal trees, which explains why apparently independent causal chains of neural events converge upon a particular body movement. However, Lowe's account fails to provide an insight into the causal mechanism responsible for the described correlations. It is based on fact-causation and claims that "what is brought about is not an event but a fact or state of affairs" (2000, p. 582). To construct an intelligible account of mental causation, however, one would have to make certain assumptions regarding the fundamental processes that would explain how a cause brings about its effect.

The convergence of neural events that Lowe pointed to implies the intentional redistribution of the physical properties of the neural system. The curious case of Maxwell's demon that was under our scrutiny in Section 3 suggests that it is conceivable that this redistribution is brought about by the mind without causing anomalous acceleration of the particles within the system – by redistributing its state probability. In Maxwell's thought experiment, the relevant distribution variable is molecular speed, but if the immaterial mind is able to perform this redistribution, then it is equally conceivable that it is able to use any physical or hypothetical mental parameter for this purpose. As we have seen, this is the only physically sound mechanism by which mind can act on matter, avoiding direct action on the individual constituents of the physical system through Newtonian forces, whose presence would lead to the appearance of anomalous accelerations and consequently to the violation of the fundamental laws of conservation. But if this is so, the nature of mental causation cannot be deterministic, because that would mean either the individuation of deterministic forces in the system or the identification of another

mechanism that would enable an unambiguous prediction of the state of the system in the future. In the next section, I will briefly outline the basic features that an account of mental causation constructed along these lines must possess to be intelligible, my main concern being to explain its prima facie probabilistic character.

To form a clearer picture of the proposed mechanism of mental causation, let us consider a convenient generalization of Maxwell's thought experiment. Suppose that many constituents of an isolated physical system are randomly distributed in two equienergetic states, S_1 and S_2, which freely exchange constituents. The constituents of the system are arbitrary, and their physical nature need not be specified. They are identical in all their properties, except for the additional property X, which only half of them have and which in no way alters the mechanism of their interaction with other constituents. In the illustration below, constituents with an additional property are marked with **x** and those without it with **o**. Since the probability of finding both types of constituents in states S_1 and S_2 is equal, the arrangement (a) schematically shows their expected, approximately random spatial distribution at time t_0. In the vocabulary of statistical physics, the distribution (a) corresponds to one of the microstates of the system through which the macrostate with maximum entropy is realized. A physical system striving to reach an equilibrium state, in accordance with the Second Law of thermodynamics, tends to this random distribution. Of course, spontaneous fluctuations in the random spatial distribution are always present, but since the number of constituents is large, they are negligible in our rough scheme.

	S_1	S_2
(a)	**x o x x o x o o x o**	**o x x o x o o x x o**
(b)	**x x x o x o x x o x**	**o x o o x o o x o o**

Fig. 1 A generalization of Maxwell's thought experiment. If there is no net change in energy, the increase in system entropy during the transition from a random distribution of properties of the system constituents (a) to a more ordered state (b), caused by an immaterial agent, indicates a violation of the Second Law of thermodynamics

If the system evolves in such a way that the arrangement of its constituents at time *t* is shown schematically in (b), an outside observer will determine that the probability of finding **x**-constituents is higher in the state S_1 and **o**-constituents in the state S_2. Naturally, she will first assume that there is a force acting on the constituents and causing this redistribution. According to the aforementioned requirement of *Redistribution*, this can only be done by supplying additional energy or momentum to the system. If, however, there is no empirical confirmation of the occurrence of such energy, the explanation becomes less straightforward. The observer must infer that there is some previously unknown selection rule, or

even an unknown law, that controls the transitions between S_1 and S_2 and increases the probability of the observed distribution, with no forces acting and with strict adherence to all conservation laws. She would not be able to know whether the examined law is of a deterministic or probabilistic character, nor whether it refers to the system as a whole or to its constituents. Be that as it may, since the appearance of anomalous correlation is its only measurable manifestation, only statistical analysis methods would be available to express it. Therefore, she would be forced to conclude that the causality that occurs in the system, at least in the first pass, must be treated as probabilistic.

In the described thought experiment, it is easy to recognize that it reduces to the effect of Maxwell's demon on gas molecules as constituents of a system in which the equienergetic states S_1 and S_2 correspond to the finding of molecules in the container chambers, and the selection property X is the speed of the molecules. Here, the demon plays the role of an agent that causes an anomalous redistribution of gas molecules. It can do this either by establishing boundary conditions that control the transitions of molecules between two available states or by changing the *a priori* probability of states S_1 and S_2 in such a way that the probability of finding faster molecules is higher in one of them. The direct consequence of this redistribution of state probabilities is the redistribution of physical properties in the system: at the end of the observed process, in one of the chambers, the concentration of faster molecules becomes too much above the statistical average to be explained by random fluctuations, while in the other, the situation is reversed. In other words, the gas is hotter in the first chamber than in the second.

From our earlier discussion of Maxwell’s demon and Earman & Norton’s analysis of numerous attempts at its exorcism, we can conclude that the demon must be an immaterial agent, not part of a physical system, to produce such effects. Its actions violate the Second Law of thermodynamics while preserving the laws of conservation. An interactional dualist can go one step further and claim that it is reasonable to assume that an immaterial mind could exert its causal effect on the physical world in an analogous way. The analogy can be extended by assuming that the role of the constituent parts of the system in which mental causation is realized can hypothetically be played not only by individual particles, such as molecules or ions, but also by more complex subsystems: ion channels, synapses, neurons, or neural circuits. The intentional nature of the mind's actions implies that they are not random but are characterized by nomological regularity. It can be assumed that they are governed by still unknown psychophysical laws, not unlike those proposed by Chalmers (1996). If we take Papineau's argument from physiology seriously by adopting the conclusion that there are no mental forces of the Newtonian type, I believe that this is the only direction in which a dualist can continue to look for a satisfactory account of mental causation.

Maxwell used his thought experiment with the demon to illustrate the statistical nature of the Second Law of thermodynamics. He wanted to emphasize that this law can only be applied to large masses of matter, i.e., to systems made up of a huge number of molecules, and not to individual molecules (Maxwell, 1871, 1878; Myrvold, 2011). Similarly, we have seen that the modified thought experiment with the immaterial mind as an agent indicates that mental causation must, in the first instance, be considered probabilistic. Like Maxwell's demon, it gives clues to its presence through anomalous correlations, from which it can only be indirectly inferred. Granted, it cannot be excluded that psychophysical laws will turn out to be deterministic at some deeper level of reality. Nevertheless, I argue that foregoing conclusions about the nature of mental causation are generally valid, since at the phenomenal level manifestations of mental causation will always be probabilistic. As we shall see, this may be a fortunate circumstance, as we already have at our disposal powerful tools of statistical analysis, developed for the purposes of statistical mechanics. It may turn out that a similar methodology can be applied to the study of causal phenomena in the psychophysical domain.

From the point of view of an external observer, who perceives the processes from a third-person perspective, anomalous correlations appear in the system, the cause of which cannot be traced by physical means. For the mind of the subject who perceives them from the first-person perspective, however, there are no anomalies, and the causal chain, completed by the contribution of the immaterial mind, is unbroken. There is no difference in the degree of knowledge of distributions (a) and (b) in **Fig. 1** from the perspective of the immaterial, immanent mind because it has immediate knowledge of the positions and properties of the system constituents. In the next section, I will offer an explanation of how the immaterial mind can acquire this kind of knowledge, i.e., what is meant by the statement that it possesses such knowledge. At any rate, although for an external observer there is a big difference in the entropies of microstates (a) and (b), they will not differ for the immaterial mind. This implies that the violation of the Second Law of thermodynamics will occur only from the point of view of an external observer and not from the point of view of the mind itself. In other words, the Second Law is inapplicable from the perspective of the immaterial mind.

If this interactionist point of view turns out to be correct, it will mean that mental causation not only violates both the Second Law of thermodynamics and CCP but also introduces a subjective element into the theory of causation. At this point, a significant analogy with the interpretation of Maxwell's demon by its creator must be pointed out. Maxwell famously argued that the notion of heat, as the measure of the energy dissipation, loses its meaning if all the initial and boundary conditions in the system, viz., the positions and momenta of all the molecules in a system, are known (see Myrvold, 2011). In fact, according to Maxwell, the distinction between heat and work is relative and comes from our

inability to trace the motion of every molecule, which makes not only the notion of heat but also of work and entropy means-relative. In Maxwell's own words, "confusion, like the correlative term order, is not a property of material things in themselves but only in relation to the mind that perceives them" (Maxwell, 1878, p. 221). Hence,

> from the perspective of a Maxwell demon, there would be no distinction between heat and work, and the very concepts needed to formulate the second law would break down, and thus the second law would be inapplicable, because the very concepts needed to formulate it would fail to apply. Its machinations would, however, look like a violation of the second law, as formulated using any distinction between heat and work (Myrvold, 2011, p. 240).

Therefore, Maxwell's statistical view of the Second Law introduces a subjective element into the very heart of thermodynamics. According to this outlook, the violation of the Second Law during the operation of the demon is present only from the perspective of the outside observer and not from the perspective of the demon itself. The presented analogy shows that an account of causation based on the assumption that the active subject is immaterial leads to the conclusion that our interpretation of concepts and laws that rule our world is essentially perspective-relative. This is an immediate generalization of Maxwell's conclusion that the interpretation of the terms of thermodynamics is means-relative.

Let me summarize the discussion so far. Recognizing the hidden premises in Papineau's arguments for physicalism and in the CCP led us to the conclusion that it is possible that mental causation manifests itself differently from physical causation: through anomalous correlations of processes – not through anomalous accelerations of particles – in the nervous system. The analogy with Maxwell's demon demonstrates that it is perfectly conceivable that the actions of an immaterial mind could influence the state of the physical system in a way that contradicts both the Second Law of thermodynamics and the causal closure principle. To require the mind to comply with the Second Law implies that the mind is nothing but a physical system within a broader physical system of the body, which presupposes that there is no irreducible mental causation. In order to produce significant behavioral effects, the immaterial mind could bring about small-scale correlations of neural processes by inducing small fluctuations in the probability distribution, which would lead to a redistribution of physical quantities. Our actions do not happen randomly but depend on our conscious choices in a highly ordered and intentional way, which implies that these actions are nomologically determined by mental facts. A dualist can try to explain this regularity through the action of still-unknown psychophysical laws. The very fact that mental causation is manifested through anomalous correlations and the redistribution of the properties of the system suggests that the nature of psychophysical laws is prima facie probabilistic. It is difficult to see how the mind could

act on individual particles, which would indicate the existence of deeper, deterministic laws, while avoiding the transfer of energy to these particles and the violation of conservation laws. It might be conceivable that some kind of deterministic dependence exists at a deeper level of mind-body interaction, but it seems that at the first pass the putative psychophysical laws are probabilistic. They can be formulated and studied by generalizing the mathematical apparatus well known from statistical mechanics. In the next section, I will outline the basic features that an intelligible probabilistic account of mental causation should have, drawing on the ideas laid out in Dimitrijević (2020). In Section 6, some important implications of this account will be highlighted.

5 Elements of the probabilistic account of mental causation

The basic premise of the interactionist worldview is that two essentially different classes of properties exist in the world, physical and mental, and that there are causal relations between instantiations of these properties at time *t*, i.e., between physical and mental events. It is important to understand why this point of view is unacceptable for the majority of naturalistically oriented authors. No matter how different the reasons may be among individual researchers, I believe that they can be reduced to the simple fact that physical properties are observable while mental properties are unobservable, which means that they cannot be measured. Measurement is a physical process of quantifying a property by comparing it with a reference quality of the same kind. It includes an extensive usage of physical laws, both in the construction of the measuring instruments, with which the measured object or system interacts, and in the measuring procedure. Thus, the whole problem with mental causation stems from the fact that mental properties do not conform to physical laws and do not participate in physical interactions. Therefore, the presence of mental properties can only be perceived through subjective experience and their causal powers evaluated through functionalization – by using their causal roles to express them in the function of their physical realizers.

The observability of a property is closely related to its individuation, which in science is done causally, on an empirical basis. The scientific judgment that the physical object O has the property P is made in such a way that we empirically establish that the causal relations in which O enters with other elements of physical reality can, in accordance with the physical law L, be explained only by the presence of the property P. Events in the physical world, as instantiations of physical properties on an object at a time, can be registered by our senses and measuring instruments because they enter into physical causal relations with them. In Section 2, we saw that causation in the physical world is manifested by the effect of the Newtonian force, which produces the acceleration of the constituents of the system, i.e., by the

transfer of energy and other conserved properties. I have argued that Papineau's argument from physiology presents a good indication that mental causation is unlikely to manifest itself in an analogous way. A thought experiment with Maxwell's demon showed that the interaction of an immaterial mind with a material system is at least conceivable; moreover, it demonstrated how such an interaction can manifest itself.[8] The discussion of generalized Maxwell's thought experiment goes a step further by establishing that it is conceivable that the relations between instantiations of physical and mental properties take the form of probabilistic psychophysical laws. In what follows, I will try to show that this presents the interactionist with an opportunity to construct an intelligible account of mental causation using the methods of statistical mechanics generalized to systems with additional degrees of freedom – the unobservable mental parameters. To preserve the generality of the discussion, I will not make assumptions regarding the specific nature of the subsystems in which mental causation can manifest; it is a matter of experimental research to determine whether it occurs at the level of ion channels, synapses, neurons, neural circuits, or the whole brain. The aim of the consideration is more modest: to indicate the expected manifestations of mental causation, how to distinguish them from physical causation and what are its general characteristics. A successful model of this kind could significantly blunt the edge of those criticisms of dualism that proceed from the premise that dualistic solutions, as a rule, are in conflict with the demands of naturalism.

The state of a complex psychophysical system, such as the human neural system or its subsystem, is a function of a set of mutually independent fundamental physical $\{q_i | i = 1, 2, \dots, k\}$ and mental $\{m_j | j = 1, 2, \dots, l\}$ state variables, or degrees of freedom[9]. In physics, the choice of the physical parameters that are considered state variables is dictated by the type of physical system and context. Mental state variables are impossible to specify precisely because they cannot be observed, i. e. measured; one can only hope that some future, more complete theory of mind and mental causation will be able to do so. All that can be done at this stage is to introduce them in an indirect way by functionally relating them to physical properties. As in the case of complex, many-particle physical systems, the state of the psychophysical system at time t can be represented by a phase point in the generalized, $(k + l)$ – dimensional phase space whose coordinates are (q_i, m_j), and its dynamic evolution by the trajectory of the phase point in this space. The probability of a macrostate realized by many microstates filling a phase

[8] If there was any doubt that Maxwell's demon should be taken seriously, it was certainly dispelled by Zhang & Zhang (1992), who gave a specific example of a time-reversal invariant and conservative dynamical system that does not preserve phase volume, in which the violation of the Second Law occurs. Their system was constructed with a membrane, which divides chambers in the experiment and acts as a force field, so that no information theory argument can be invoked to save the Second Law from the demon.

[9] An interesting discussion of the relative autonomy of the properties of a system on the dispositions of the particles or other system constituents can be found in Chakravartty (2019, p. 13).

volume can be determined, as in conventional statistical mechanics, by calculating the number of elementary cells in this phase volume. Therefore, bearing in mind our earlier inferences concerning the *modus operandi* of Maxwell's demon, I believe that an intelligible interactionist account of mental causation can be summarized in the following way:

> *Mental causation* is instantiated as the disposition of mental properties to alter the state probability distribution within the nervous system or its subsystem, thus leading to the redistribution of energy, momentum, and other conserved quantities without altering the overall energy and momentum content of the system.

It is important to note that altering the state probability within a physical system, which leads to the redistribution of the conserved quantities, is not an exclusive property of mental causation. Physical causation is also treated as probabilistic in systems with many degrees of freedom, where a deterministic description would be too complicated, or in quantum mechanical systems, where such description is completely impossible. Even when the dynamics of a physical system are governed by deterministic laws, they can be formally reinterpreted probabilistically as realizations of dispositions. From a probabilistic point of view, therefore, there is no essential difference between the descriptions of physical and mental causation in a physical system. The key feature that distinguishes physical causation from mental causation is that physical effects are mediated by the Newtonian forces, which necessarily alter the overall energy and momentum content of the system, as specified by Gibb's *Redistribution* premise; mental causation, on the other hand, can be recognized by the fact that physical effects occur without the perceptible mediation of any physical forces. So to provide an intelligible account of mental causation, an interactionist must indicate the way in which an immaterial mind may achieve redistribution without the help of acceleration-causing forces.

In order to show how this can be achieved, we continue our generalization from statistical mechanics. We can surmise that the probability $W(E_i, m_j)$ that a psychophysical system belongs to canonical ensemble, in a state characterized by energy E_i and the set of mental state variables $\{m_j | j = 1, 2, \dots, l\}$, can be obtained from the expression $W(E_i, m_j) = G(E_i, m_j)p_m(E_i, m_j)$. Here, G is the statistical weight of the level determined by (E_i, m_j), i.e., the number of microstates corresponding to this set of psychophysical variables, and p_m the probability of a microstate corresponding to these variables. We will analyze how the modification of each of these two factors could lead to a change in the probability of the state of the system, W.

The introduction of additional degrees of freedom in the form of mental state variables necessarily changes the probability of the state corresponding to a particular energy. Moreover, it increases the number of configurations corresponding to a microstate of the system, thus changing the statistical weight of the level G in a way that resembles the splitting of energy levels in a physical system into sublevels. This is a direct generalization from statistical mechanics, only in this case the levels are not differentiated by their energy but by their mental state variables. We can understand this better if we represent the state of the system as a phase point in a generalized phase space with as many dimensions as there are physical and mental state variables, i.e., $k + l$. Adding each new degree of freedom introduces a new dimension to this generalized phase space and consequently increases the number of phase cells in the phase volume. This procedure is equivalent to increasing G. Generalized phase space can be a useful tool for studying the dynamics of a psychophysical system. Each phase cell represents a specific psychophysical state of the system, and each phase trajectory represents a specific system dynamic governed by psychophysical laws, the analysis of which can enable us to better understand mental causation. In the case of complex psychophysical systems, the phase trajectory corresponds to nothing less than the way the mind chooses to act in the physical world.

How the *a priori* probability of microstates p_m depends on mental variables is determined by the psychophysical laws, which are at present unknown to us. As we saw in the example of Maxwell's demon, to control that probability, the mind must control either the state probability distribution or boundary conditions in the system. The first option means that the probability of the state is determined by a function in which both physical and mental parameters figure as system variables instead of only physical parameters as in purely physical systems, such as those described by Maxwell-Boltzman or Fermi-Dirac distributions. In this scenario, the immaterial mind is able to assign specific values to the mental parameters, thereby controlling the state probability. This manifests itself in such a way that some states of the subsystem become more or less probable than if this probability were controlled only by the aforementioned physical distribution functions. The transition probabilities between the mentioned states are simultaneously changed. From that moment on, physical laws take over the role of effective causes of all processes. Depending on which class of subsystems of the nervous system it would take place, the result of this redistribution of system state probability could be, for example, an increased permeability of an ion channel or activation of a neuron or a neural circuit; within the established framework of this work, I will refrain from all speculations of this kind. The second possibility of influence of mental variables on p_m consists in the setting of boundary conditions by the immaterial mind by establishing selection rules that constrain the transitions between states, analogous to Hund's rules or the Pauli's Exclusion Principle in atomic physics. These nomological constraints "narrow down the set of physical possibilities […] –

that is, the constraints act directly on the space of possibilities." (Adlam, 2022). As before, in the absence of knowledge about the psychophysical laws that this process obeys, we can only speculate about exactly in which subsystems within the nervous system this change in boundary conditions occurs. A potential candidate is any subsystem that can be found in two or more states, with distinctive physical consequences of transitions between these states – whether on a microscopic, mesoscopic, or macroscopic spatial scale. Our lack of knowledge of even the spatial scale at which the putative action of the immaterial mind may occur makes the problem rather complex. Most accounts of irreducible mental causation based on quantum-mechanical considerations can be attributed to this class of possibilities.

Whichever of the proposed mechanisms actually works in the nervous system, the result of the mind's intervention is a redistribution of energy between subsystems, resulting in bodily action without the expenditure of work and thus without changing the energy content of the system but leading to a local decrease in entropy. Such redistribution represents the physical realization of an underlying causal process during which the constituents of the system show a disposition to find themselves in the chosen state. It is a state that becomes favored by the very act of deciding to carry out some action by the immaterial mind. This would explain how the very act of making a choice makes a certain state more likely than other physically equivalent possible states, whose probability is otherwise equal in purely physical systems.

The generalized method of statistical mechanics gives us, among other things, the means to analyze the functional dependence between physical and mental properties, with the prospect of eventually establishing their nomological connection. We have seen that functionalizing unobservable mental parameters is of utmost importance since there is no way of registering their very existence directly, from a third person perspective. Their causal roles must be determined indirectly, through the study of the causal roles of the physical properties with which they are correlated. To illustrate this more clearly, let us get back to our generalization of Maxwell's thought experiment, displayed in Section 4. If the arrangement (b) shown in **Fig. 1** temporally follows the arrangement (a) as a result of an act of an immaterial agent with no force acting on the constituents of the system, it is clear that the order in the system has increased, and its entropy has decreased. That means that the Second Law is violated, as the probability W of the arrangement (b) is much less than that of the arrangement (a). In Section 4, I argued that in order to change the probability of the state of the system in this way, without any physical interaction with its constituents, an immaterial agent could either establish specific boundary conditions between states S_1 and S_2 or adjust *a priory* probability of these states in order to increase the probability of finding the constituents with property X in one of them. Suppose that X is a mental property. It is not relevant for this discussion whether it is immanent to a given constituent or type of constituents of the

system, attributed to it by the mind, or acquired at some point in accordance with psychophysical laws. The answer to that question must be offered by a future, concrete model of mental causation if the basic principles presented in this paper turn out to be correct. Mental property X is unobservable to the outside observer but is, by assumption, directly accessible to the immaterial mind. An outside observer, although unable to perceive property X, distinguishes its bearers as constituents of the system who perform the empirically established causal role R. In the eyes of this observer, the distribution (b) of system constituents shown in **Fig. 1** can be formally represented in the manner shown in **Fig. 2**. In order to statistically analyze the system dynamics, the observer can consider its states S_1 and S_2 as split into sublevels L_1 and L_2, populated respectively by those constituents that perform a causal role and those that are causally idle. In every other sense, the constituents are identical. This is analogous to the splitting of energy levels in atomic and other physical systems, which brings about a change in the statistical weight of the levels and changes the way state probabilities are calculated. Here, the splitting is done according to some mental parameter, not energy, but the conclusion about changing the way of calculating the state probability remains. At the end of the analysis, the observer will be able to functionally define the unobservable property X as a property of the constituents that populate the level L_1 of the state S_1 and that perform the physical causal role R. So the main advantage of this type of analysis is that it gives researchers the means to individualize mental parameters by relating their specific causal roles to corresponding changes in the state probability of some part of the neural system. Alternatively, if it turns out that adjusting the *a priori* state probabilities, instead of establishing boundary conditions, is actually the more probable *modus operandi* of the immaterial mind, we could use statistical analysis to understand the nature of the psychophysical laws responsible for the system's manifested probabilities.

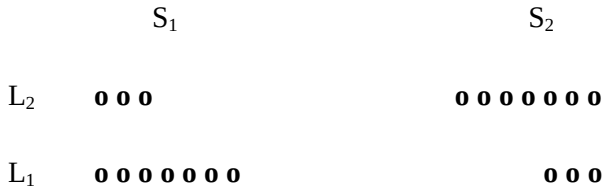

Fig. 2 The distribution (b) from Fig. 1 from the point of view of an outside observer. States S_1 and S_2 are split into sublevels L_1 and L_2, depending on the possession of the unobservable, functionalized mental selection parameter X by the system constituents. The spatial arrangement of the constituents is disregarded

The essence of the probabilistic interpretation of irreducible mental causation is that the constituents of the system, without an obvious physical cause, show a disposition to find themselves in a certain state. It is precisely that state that corresponds to the subject's choice and leads to a physical action

in accordance with that choice. In Section 4, I mentioned that it was this feature of mental causation that led Lowe to conclude that the role of correlates of neural processes is crucial to understanding mental causation. The discussion so far allows for a dualistic answer to Lowe's question of how independent causal chains of neural events converge toward a particular bodily movement. An interactionist dualist may argue that the immaterial mind, in some of the ways outlined in this section, changes the probability of the state of the system and thus makes some processes in the nervous system more likely than they would be if there were only physical causes in the system. It is up to empirical research in the domain of neural sciences to determine exactly in which part of the nervous system this redistribution of state probability occurs. In any case, its direct physical consequence is the corresponding redistribution of energy, momentum, and possibly other conserved physical properties in the system.

An anonymous reviewer drew my attention to the fact that, based on the mathematical theory of natural extensions or dilations of stochastic processes, it is possible to transform stochastic processes into deterministic ones and vice versa. “A system that is originally described stochastically, e.g. due to uncontrollable interactions with its environment, is successfully extended into the environment as long as all interactions are integrated in the behavior of the system itself. This leads to an increasing number of degrees of freedom, enabling an integration of all previously stochastic behavior into an overall deterministic dynamics” (Atmanspacher & Rotter, 2008, p. 304). The described mechanism of mental causation, as we have seen, implies a similar increase in the number of degrees of freedom, which is achieved by the successive inclusion of the functionalized mental parameters of the system. Naturally, the question arises as to whether such a procedure would at some level lead to the integration of the stochastic behavior of the observed system into deterministic dynamics, and weather this could explain the high degree of determinism and intentionality in the individual's behavior. In any case, this possibility deserves attention and further research.

The described mechanism of mental causation leads us to the conclusion that the dynamics of a complex system in which mental causation is manifested can best be modeled by following the time evolution of the state probability distribution function $f(q_i, m_j)$ in the generalized phase space of the system. In perspective, this should enable us to establish the probabilistic nomological relations between physical and mental state variables, i.e., psychophysical laws. The function $f(q_i, m_j)$ can be defined as the number of system constituents per unit volume of the generalized phase space of the system. It is important to note that this function can be theoretically introduced as a mathematical tool even in the case that mental causation is deterministic in nature, to facilitate evaluations of the outcome of events during mind-body interaction. However, its role becomes crucial if mental causation turns out to be inherently probabilistic, as I believe it is. In that case, $f\left(q_i, m_j\right)$ contains the most complete available information

about the state of a complex psychophysical system. Combined with the psychophysical laws responsible for its temporal evolution, it becomes a measure of the disposition of the psychophysical system to evolve in a certain way. We are now in a position to offer an interactionist answer to the question of what it actually means to claim that the immaterial mind has direct knowledge of the state of the subject as a psychophysical system. In the terminology of the considered probabilistic account of mental causation, it means that the mind directly perceives the distribution of physical and mental parameters, which enables it to redistribute the probability of the state of the system, i.e., to control the dynamics of the state distribution function $f(q_i, m_j)$ in accordance with the decisions it makes. Only after gaining knowledge about psychophysical laws will it be possible to say something more about the mechanism of this action of the mind.

Towards the end of this section, I will briefly address the possible objection that appeal to psychophysical laws is intuitively unclear, and therefore insufficient for explanation of the mechanism by which mental parameters lead to the redistribution of the state probability of the system, and that more should be said to understand this mechanism properly. Now, to explain a physical phenomenon it is sufficient to point to the relevant physical laws that cause it, primarily the laws of acting forces, and to the specific conditions in which their effect is manifested in a given system. The impression that such an explanation of a physical process is intuitively clear comes from the closeness of physical phenomena to our sensory experience and acquired knowledge, as well as from familiarity with the way in which the effect of each physical force is manifested – by causing an acceleration. This intuition is partially lost already when explaining more abstract phenomena, or phenomena inaccessible to the senses in the domain of theoretical physics – for example, quantum mechanics, whose laws give predictions of astonishing accuracy, but whose interpretations have been a stumbling stone for almost a century. Even then, it is tacitly understood that *modus operandi* of physical causation is clear. In comparison, the claim that mental causation manifests through correlation, instead of through acceleration, although conceivable and possible, may be unintuitive. But the parallel with physical explanations indicates that eventual discovery of psychophysical laws would make the offered explanation, based on invoking these laws, as complete as the explanations of, for example, quantum mechanical phenomena.

A way to gain knowledge about the probability distribution function and psychophysical laws is to study anomalous neural correlations and their relationship to the decisions of the conscious mind. This would be a good start, certainly within the means of current or near-future science. It is in this direction that the proponents of Integrated Information Theory (IIT) are looking for interconnected units of neurons in the form of neuronal coalitions, or networks of neurons, which would represent physical substrates of

consciousness.[10] It is reasonable to expect that in such subsystems of the nervous system, the dispositional nature of the corresponding probability distribution function could be manifested in the most obvious way. In any case, only empirical evidence derived from neurological research can provide the final pieces to the puzzle of mental causation and help us decide whether it is truly irreducible, or whether there is nothing "over and above" the physical after all.

6 Some physical analogies and their upshots

In this section, I will briefly consider a couple of cases where there are significant physical analogies to the account of mental causation proposed in this work, because they will provide a better insight into some of the more far-reaching consequences of the offered account. First, the mechanism of mental causation, with the function $f(q_i, m_j)$ at its core, is very similar to that of quantum mechanical causation, which is invoked in causal explanations of microphysical events. This ostensibly formal analogy will be briefly discussed because it actually implies that a higher degree of causal unity of nature can be achieved. I will then compare my probabilistic interactionist account of mental causation with its closest physicalist analogue known to me. Proposed by Papineau (2013), it rests on statistical physics and thermodynamics, which in itself speaks of the fact that the inherently probabilistic nature of mental causation did not escape its author. A comparison of these two accounts will allow us to get a better picture of the similarities and differences between the approaches to this problem that proceed from two diametrically opposed ontological premises.

It is well known that the state of a physical system at the deepest, quantum mechanical level cannot be described as in classical physics, using the coordinates and momenta of system particles as state variables. Heisenberg's uncertainty principle, which is a consequence of the dual, particle-wave nature of quantum objects, prevents simultaneous knowledge of position and momentum, as well as other pairs of non-commuting variables. Instead, according to the First postulate of quantum mechanics, the complete state of the system is specified by its state function ψ. It is a function of the positions of all the particles in the system and time and has no physical meaning in itself. Only the square of the modulus of this function has an immediate physical meaning and gives the probability density of finding a particle at a certain location. The evolution of the physical system in time is described by the Schrödinger equation, whose argument is the state function ψ. Its role in quantum mechanics is similar to that of Newton's laws in

[10] Some of the recent developments in IIT, particularly its latest version 4.0, which claims that consciousness is metaphysically primary while the physical domain is just operational, are presented in Albantakis et al. (2023), Cea et al. (2023), and Marshal et al. (2023).

classical mechanics. Solving it allows us to predict the probability of the state of the system at some point in the future.

The circumstances that led us to the idea of the probabilistic nature of mental causation are similar to those that made physicists reluctantly accept the probabilistic nature of the laws of quantum mechanics nearly a century ago. It turned out that the state variables that describe physical systems and figure in physical laws do not provide the possibility of a complete description of the psychophysical system and its temporal evolution. Instead, it can be assumed that psychophysical systems are dual in nature, i.e., they possess both physical and mental properties, just as quantum objects are dual in nature because they have both wave and particle properties. The particle and wave properties of quantum objects cannot be measured simultaneously but only in separate experiments, and they manifest in radically different ways. Physical and mental properties also manifest differently: physical properties are observable and interact with the measuring apparatus by causing acceleration of parts of the system, whereas mental properties are unobservable and manifest their presence by causing correlations of neural events. The central notion used for the description of the state of the psychophysical system is a probability distribution function $f(q_i, m_j)$; its role is analogous to that of the state function ψ in quantum mechanics. The temporal evolution of the quantum system is described by the Schrödinger equation, whose solution enables the prediction of the future state of the system; the temporal evolution of the psychophysical system is a consequence of hypothetical psychophysical laws, which would in principle enable the prediction of the probability of the outcome of neural events. Such striking parallels justify the assumption that, at the deepest level, mental and physical processes proceed in accordance with similar principles of a probabilistic nature.

All of this indicates that the *modus operandi* of both physical and mental causes is fundamentally the same: they influence the state probability distribution. This results in the appearance of physical effects that always act in such a way as to enable the evolution of the system in the direction of the most probable state. The probability of the state of the system, as not only a description of the state but also a measure of the disposition of the system to evolve in a certain way, appears as a central concept in both models. Admittedly, the interpretation of quantum mechanics is still a matter of debate. There are approaches, such as the pilot-wave model proposed by de Broglie and Bohm, that postulate the existence of a deeper level of reality and warrant the essentially deterministic nature of basic physical laws[11]. Likewise, there is no way of knowing at the present time if there is a deeper level of deterministic

[11] For the exposition of this model, see Bohm (1952). Note, however, that deterministic interpretations of quantum theory take nothing from the fact that it remains non-deterministic in the most important, practical sense: that only the probabilities of individual quantum events can be calculated.

causation underlying the prima facie probabilistic psychophysical laws. However, both physical and mental causation seem to be at least phenomenally probabilistic. Therefore, this account suggests a kind of causal unity of nature, which could indicate that the gap between physical and mental properties is not so insurmountable after all.

Let me turn to a comparison of the probabilistic interactionist account of mental causation considered here with the rather influential physicalist account presented by Papineau (2013). He argues that causation is essentially a macroscopic physical phenomenon, with a clear probabilistic signature, similar to thermodynamic processes[12]. Causation is asymmetric in time, as causes always precede their effects, even though the basic laws of dynamics are invariant to the direction of time. According to Papineau, this suggests that causation is "constituted by the nature of past facts together with probability distributions over the maximally specific microstates that can realize given macrostates" (2013, p. 129). In this, both probabilistic accounts are in full agreement. Papineau further argues that if the physical conditions are fully known, causation ceases to be an asymmetric relation and is lost. Insisting on the analogy of causation and thermodynamic effects and quantities, especially heat, makes his inference analogous to the statistical view of the nature of the Second Law, discussed in Section 4. The same outlook led Maxwell to the conclusion that the very notions of heat, work, and entropy lose their meaning if the dynamical conditions in the system are fully specified. Although Papineau's account of mental causation is built from the position of a reductive physicalism, it bears striking similarities to the probabilistic interactionist account outlined in this work, in that causal relations are inferred from the probabilistic facts concerning the way in which specific microstates are realized. The key difference is that in Papineau's account, the state probability distribution is a function of only physical variables, while in the interactionist account it must necessarily include mental state variables. Their inclusion would completely invalidate Papineau's claim that there is no place for causation if all physical facts about the system are determined. It is probably at this point that we should start looking for empirical facts, primarily anomalous correlations in the nervous system, which would point to the predominance of one of the offered explanations of mental causation.

The main problem with every reductive physicalist explanation of mental causation is multiple realizability. It does not suffice to point to the specific physical realizers of a mentally caused bodily action, because they lack causal uniformity; in addition, one must find a common feature at the level of physical realizations of mental events that would explain the co-variance of a supervening mental cause and a physical effect. To do this, Papineau posits brain states picked by the phylogenetic and ontogenetic

[12] The suggestion that causation is essentially a macroscopic phenomenon is upheld by a number of other prominent writers; see Albert (2000), Loewer (2007), Woodward (2007), and Haug (2019).

selection processes as generic selectional states corresponding to mental states. They do not cause physical effects immediately, but can be used to explain them, because these selectional states are picked out by mental states or are type-identical with them. Selectional states are variably realized, so that different physical realizers can fulfill the role of immediate, effective causes.

To better understand the similarities and differences between the two approaches to the problem of mental causation, let us return to **Fig. 1**. In order to perform some useful work in a physical system, there must be gradients of appropriate physical quantities that will lead to the action of physical forces. As we saw in Section 2, these forces and the corresponding transfers of conserved quantities act as means to restore the system to a state of thermodynamic equilibrium. So it is necessary for our system to have a gradient of the property X, possessed by the constituents marked with **x,** in adjacent parts or states of the system, S_1 and S_2. This corresponds to the transition of the system from the equilibrium state (a) to the selectional state (b), in which the desired effect spontaneously follows through a purely physical mechanism due to the difference in concentration of **x** in S_1 and S_2. Any selectional state that represents the cause of some physical effect is realized in an analogous way. Some complicated arrangement of physical constituents – ions on cell membranes, molecules of neurotransmitters, neurons, or neural circuits – would represent the selectional state that is the physical realization of my decision to hail a taxi. Spontaneous physical processes that, on the basis of physical laws, follow from that arrangement would result in the realization of the decision – for example, arm waving. The described mechanism does not differ in any way between the two accounts that we are analyzing. The difference appears only at the moment when the question is raised as to how the selectional state corresponding to the mental cause is established. This is the link in the explanatory chain in which the contrast between dualistic and all – not only Papineau's – physicalist explanations of mental causation is sharpened. Papineau's account finds the answer in phylogenetic and ontogenetic selection processes, while interactionists find it in the redistribution of the probability of the state of the system caused by the effect of the immaterial mind.

In essence, the main difference between the two accounts is the explanation of the nature of the generic states that ensure uniformity of causal patterns, as well as the fact that different physical realizers of a mental property cause the same physical effects in different contexts. Physicalists have no choice but to claim that these states as purely physical, and that their origin is a combination of phylogenetic and ontogenetic factors. In the interactionist account presented in this paper, generic states are seen as dispositions of the system and are represented by a probability distribution function realized by a specific configuration of the physical and mental parameters of the system. The uniformity of the causal process is ensured by the fact that the concept of state probability applies equally to the laws governing physical and mental events. The disposition of a system to evolve in a particular way is controlled by laws containing

both types of state variables. In both accounts, the role of immediate causes of physical effects belongs to different physical realizers, such as forces whose laws are the subject of physics; however, in the interactionist view, the probability that the realizer will lead to a specific physical effect can, at least in principle, be uniquely derived from $f(q_i, m_j)$. Different physical consequences resulting from the same physical, but different mental states, would become, at least in principle, explainable. Also, in the probabilistic interactionist account of mental causation, the problem of multiple realizability – so awkward for type physicalism – simply does not arise. A mental state can be realized by different combinations of physical and mental parameters, which gives an external observer the false impression that a mental state is realized by different purely physical microstates. Of course, the offered interactionist explanation of mental causation can, like other dualistic models, be criticized for multiplying entities, contrary to the requirement of explanatory parsimony. The dualist, however, can defend the thesis that its plausibility is greater than that of the corresponding physicalist models, since it contains fewer weak points, such as the problem of multiple realizability, while its potential unifying power makes it an interesting metaphysical prospect.

7 Conclusion

Most interactionist accounts of mental causation suffer from an inability to withstand the challenge of the argument from causal closure of the physical. If mental causation manifests itself in the same way as physical causation, which is by producing anomalous accelerations in the nervous system that are inexplicable by the action of known physical forces, then it is not easy to explain the inability of modern science to empirically register such accelerations. Also, the influx of energy accompanying this kind of action of the mind would violate the Law of conservation of energy, since energy transfer is exactly what the action of force comes down to. I have argued that the only remaining, physically plausible option for the interactionist is to assume that mental causation manifests itself by effecting the state probability distribution within the nervous system, without changing the overall energy and momentum content of the system, analogously to the strategy applied by Maxwell's demon. If this is true, the immaterial mind is able to redistribute energy, momentum, and other conserved quantities, which leads to observable physical effects. Since the decrease in entropy of the system is not compensated for, due to the supposed immaterial nature of the mind, mental causation is accompanied by a local violation of the Second Law of thermodynamics. In the absence of anomalous accelerations, mental causation is instantiated in anomalous correlations of neural events. Complete information about the state of the psychophysical system is contained in its state probability distribution function $f(q_i, m_j)$. System dynamics can be

studied by examining the temporal evolution of this function, the knowledge of which, by assumption, leads to the discovery of psychophysical laws that connect the physical and mental parameters of the psychophysical system.

I argued that an intelligible interactionist account of mental causation is conceivable, that it should be at least prima facie probabilistic, and that a world in which it is true is possible. It is only rational to seek an answer to whether this possible world is our world. The main features of this interactionist account are outlined, without pretension to completeness. The answer to the question of the veracity of the account can be obtained through extensive physiological research, supported by statistical analysis of neural correlations.

Declarations

Interests The author has no relevant financial or non-financial interests to disclose.

Acknowledgement This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Contract No. 451-03-47/2023-01/200124

References

Adlam, E. (2022). Determinism beyond time evolution, *European Journal for Philosophy of Science*, 12 (4), 1-36.

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai. S., Fujii, K., David, I., Hendren, J., Lang, J. P. and Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms, *PLoS Computational Biology*, 19 (10), 1-45.

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

Atmanspacher, H. & Rotter, S. (2008). Interpreting neurodynamics: concepts and facts, *Cognitive Neurodynamics*, 2 (4), 297-318.

Averil, E. & Keating, B. F. (1981). Does Interactionism Violate a Law of Classical Physics?, *Mind*, 90, 102-107.

Bishop, R. (2010). The *Via Negativa*: Not the Way to Physicalism, *Mind and Matter*, 8 (2), 203-214.

Bishop, R. (2012). Excluding the Causal Exclusion Argument against Non-reductive Physicalism, *Journal of Consciousness Studies*, 19 (5-6), 57-74.

Bohm, D. (1952). A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II, *Physical Review*, 85 (2), 166–193.

Brillouin, L. (1953). The Negentropy Principle of Information, *Journal of Applied Physics*, 24, 1152-1163.

Cea, I., Negro, N. and Signorelli, C. (2023). The Fundamental Tension in Integrated Information Theory 4.0's Realist Idealism, *Entropy*, 25 (10), 1453.

Chakravartty, A. (2019). Physics, metaphysics, dispositions, and symmetries – À la French, *Studies in the History and Philosophy of Science*, 74, 10-15.

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

Crane, T. (1995). The Mental Causation Debate, *Aristotelian Society Supplementary*, 69, 211-236.

Cucu, A. C. & Pitts, J. B. (2019). How Dualists Should (Not) Respond to the Objection from Energy Conservation, *Mind and Matter*, 17 (1), 95-121.

Dennett, D. C. (1991). Consciousness Explained, New York: Back Bay Books.

Dimitrijević, D. R. (2020). Causal Closure of the Physical, Mental Causation, and Physics. *European Journal for Philosophy of Science*, 10 (1), 1-22.

Dowe, P. (2000). *Physical Causation*, Cambridge: Cambridge University Press.

Dowell, J. (2006). The Physical: Empirical, not Metaphysical, *Philosophical studies*, 131, 25-60.

Earman, J. & Norton, J. D. (1998). EXORCIST XIV: The Wrath of Maxwell's Demon. Part I. From Maxwell to Szilard, *Studies in History and Philosophy of Modern Physics*, 29 (4), 435-471.

Earman, J. & Norton, J. D. (1999). EXORCIST XIV: The Wrath of Maxwell's Demon. Part II. From Szilard to Landauer and Beyond, *History and Philosophy of Modern Physics*, 30 (1), 1-40.

Eccles, J. (1980). *The Human Psyche*, New York: Springer.

Eccles, J. (1987). Brain and Mind: two or one?, in C. Blakemore & S. Green fields, (eds.) *Mindwaves*, Oxford: Blackwell.

Garcia, R. K. (2014). Closing in on Causal Closure, *Journal of Consciousness Studies*, 21(1-2), 96-109.

Gibb, S. (2010). Closure Principles and the Laws of Conservation of Energy and Momentum, *Dialectica*, 64, 363-384.

Gibb, S. (2015). The Causal Closure Principle, *The Philosophical Quarterly*, 65 (261), 626-647.

Gillette, C. & Witmer, D. G. (2001). A 'physical' need: Physicalism and the *via negativa*, *Analysis*, 61 (4), 302-309.

Haug, M. C. (2019). No microphysical causation? No problem: selective causal skepticism and the structure of completeness-based arguments for physicalism, *Synthese*, 196, 1187-1208.

Jackson, F. (1996). Mental Causation, *Mind*, 105 (419), 377-413.

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

Landauer, R. (1961). Irreversibility and Heat Generation in the Computing Process, *IBM Journal of Research and Development*, 5, 183-191.

Lowe, E. J. (2000). Causal Closure Principles and Emergentism, *Philosophy*, 75**,** 571-585.

Lowe, E. J. (2006). Non-Cartesian substance dualism and the problem of mental causation, *Erkenntnis* 65 (1), 5-23.

Lowe, E. J. (2008). *Personal Agency: The Metaphysics of Mind and Action*, Oxford: Oxford University Press.

Loewer, B. (2007). Counterfactuals and the second law. In Price, H. & Corry, R. (eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited* (pp. 293-326). Oxford: Oxford University Press.

Marshal, W., Grasso, M., Mayner, W. G. P., Zaeemzadeh, A., Barbosa, L. S., Chastain, E., Findlay, G., Sasai, S., Albantakis, L. and Tononi, G. (2023). System Integrated Information, *Entropy*, 25 (2), 334.

Maxwell, J. C. (1871). *Theory of Heat*. London: Longmans, Green, and Co.

Maxwell, J. C. (1878). Difussion. In *Encyclopedia Britannica* (9^{th} ed.), Vol. 7, 214-221.

Melnik, A. (2003). A Physicalist Manifesto: Thoroughly Modern Materialism, Cambridge: Cambridge University Press.

Montero, B. (2003). Varieties of causal closure, in Walter S. & Heckman, S. (eds.), *Physicalism and mental causation: The metaphysics of mind and action*, Charlottesville, VA: Imprint Academic, 173-187.

Morowitz, H. J. (1987). The Mind Body Problem and The Second Law of Thermodynamics, *Biology and Philosophy*, 2 (3), 271-275.

Myrvold, W. (2011). Statistical mechanics and thermodynamics: A Maxwellian view, *Studies in History and Philosophy of Modern Physics*, 42, 237-243.

Papineau, D. (2001). Rise of Physicalism, in Gillett, C. & Loewer, B. (eds.), *Physicalism and its Discontents*, Cambridge, MA: Cambridge University Press, 3-36.

Papineau, D. (2013). Causation is Macroscopic but Not Irreducible, in Gibb, S., Lowe, E. J., and Ingthorsson, R. D. (eds.), *Mental Causation and Ontology*, 126-152, Oxford: Oxford University Press.

Popper, K. & Eccles, J. (1977). *The Self and its Brain*, New York: Springer.

Saad, B. (2018). A causal argument for dualism, *Philosophical Studies*, 175, 2475-2506.

Smoluchowski, M. (1912). Experimentell nashweisbare, der üblichen Thermodynamik widersprechende Molekularphänomene, *Physikalische Zeitschrift* 13, 1069-1080.

Spurrett, D. & Papineau, D. (1999). A note on the completeness of 'physics', *Analysis* 59, 25-29.

Stoljar, D. (2021, May 25). *Physicalism*. The Stanford Encyclopedia of Philosophy (Summer 2023 Edition), Edward N. Zalta (ed.). Retrieved December 12, 2023, from https://plato.stanford.edu/entries/physicalism/

Szilard, L. (1929). On the Decrease of Entropy in a Thermodynamic System by the Intervention of Intelligent Beings, in *The Collected Works of Leo Szilard: Scientific Papers* (Boston, MA: MIT Press), 120-129.

Tiehen, J. (2015). Explaining causal closure, *Philosophical Studies*, 172, 2405-2425.

Woodward, J. (2007). Causation with a human face. In Price, H. & Corry, R. (eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited* (pp. 66-105). Oxford: Oxford University Press.

Zhang, K. & Zhang, K. (1992). Mechanical Models of Maxwell's Demon with Non-invariant Phase Volume, *Physical Review A*, 46, 4598-4605.