Preprint – Accepted for publication in *European Journal for Philosophy of Science*

Causal closure of the physical, mental causation, and physics

Dejan R. Dimitrijević

Department of Physics

Faculty of Sciences and Mathematics

University of Niš

Abstract

The argument from causal closure of the physical (CCP) is usually considered the most powerful argument in favor of the ontological doctrine of physicalism. Many authors, most notably Papineau, assume that CCP implies that physicalism is supported by physics. I demonstrate, however, that physical science has no bias in the ontological debate between proponents of physicalism and dualism. I show that the arguments offered for CCP are effective only against the accounts of mental causation based on the action of the mental forces of a Newtonian nature, i.e. those which manifest themselves by causing accelerations. However, it is conceivable and possible that mental causation is manifested through the redistribution of energy, momentum and other conserved quantities in the system, brought about by altering the state probability distribution within the living system and leading to anomalous correlations of neural processes. After arguing that a probabilistic, interactionist model of mental causation is conceivable, which renders the argument from causal closure of the physical ineffective, I point to some basic features that such a model must have in order to be intelligible. At the same time, I indicate the way that conclusive testing of CCP can be done within the theoretical framework of physics.

Keywords: Causal closure of the physical; Mental causation; Second law of thermodynamics; Physics; Probability distribution

1 Introduction

At the beginning of his influential paper 'The Rise of Physicalism' (2001), David Papineau, one of the leading proponents of the ontological doctrine of physicalism in our time, made a bold statement that "physical science has come to claim a particular kind of hegemony over other subjects in the second half of [20th] century. This claim to hegemony is generally known by the name of 'physicalism'" (2001, 3). By physicalism we mean, in the shortest, the thesis that the world is ultimately physical: there is nothing over and above physical properties and events in our universe (e.g. Dowel 2006, 25). By proclaiming a radical ontological doctrine as the manifestation of hegemony of the fundamental natural science, Papineau in the strongest possible way expresses his conviction that veracity of this doctrine is supported by modern physics. As a crucial premise in support of this conviction he cites the thesis of causal closure of the physical: since all the physical effects in our experience have physical causes, the perceived mental causes of physical effects must themselves be physical.

The attempt to find support for an ontological thesis in precepts of the physical science is quite understandable. There is an enormous confidence of both general public and scientific community in the ability of physics to explain and predict the course of natural phenomena. This confidence stems from the unprecedented successes of theoretical physics and the rapid rise of ensuing technology, especially in the last couple of centuries. It is a shared belief of the majority of contemporary scientists that physical science will eventually lead to the complete theoretical description of the world.

On the other hand, the doctrine of physicalism is far from capable of arousing similar level of confidence, mainly due to the fact that the existence of apparently non-physical, especially mental properties cannot be disputed. Thus, in order to motivate the unintuitive claim that there is nothing over and above physical, the proponents of physicalism have developed a wide spectrum of reductive and non-reductive doctrines, ranging from type physicalism, to token physicalism, to a variety of nuances of supervenience and realization physicalism (Stoljar 2017).

The original, reductive physicalism, based on the claim that mental properties can be identified with the physical ones, lost its popularity mainly due to the problem of multiple realizability. Numerous variations of non-reductive physicalism are faced with difficulties which for the most part originate from the unclear definition of 'physical' property used in their formulation. In theory-based conceptions the property is usually thought of as physical if it is a part of the vocabulary of a physical theory. The main objection against such formulation is known

as Hempel's dilemma (Hempel 1980). If we interpret the notion 'physical' as something that contemporary physics claims exist, than physicalism must be false because current physics is not complete. On the other hand, if 'physical' is understood as something that hypothetical ideal and complete future physics contains, physicalism reduces to a trivial thesis, because we do not know what kind of entities some future physics will contain. Indeed, it looks probable that considerable widening of the inventory of physics is needed if we hope to include mental properties in the comprehensive description of physical reality. I show in this paper that it is in the least conceivable that some of these novel inclusions may consist in items that could help bridge the gap which traditionally divides the realms of biology, psychology and physics.

The argument from causal closure of the physical, also referred to as the argument from completeness of physics, was devised as a means to strengthen the physicalist cause.[1] The core of the argument is the ontological thesis known as the causal closure principle, which basically claims that the chances of all physical effects are determined by their prior physical causes. The power of this principle is considerable, being drawn from many decades of thorough research conducted by global scientific community – although this surely does not vindicate Papineau's referring to it as 'empirical' (2001).[2] Consequently, many scientists share the impression that causal closure principle is somehow entailed by the laws of physics.

The principal aim of this paper is to show that this impression is not justified. After considering the most convincing arguments used by Papineau in support of his conclusions, I demonstrate that physical science has no bias in the ongoing debate between proponents of physicalism and dualism. In other words, claims that CCP, and therefore physicalism, are supported by physics are unfounded, because so far the arguments in favor of CCP are effective only against the accounts of mental causation based on the action of the mental forces of a Newtonian nature – which means, leading to anomalous accelerations in the brain.[3] I argue, however, that it is conceivable that mental causation can be manifested through the redistribution of energy, momentum and other conserved quantities in the system, caused by altering the state probability distribution within the living system and leading to anomalous correlations of neural

[1] For some notable expositions of the causal closure principle and CCP argument see Crane (1995), Jackson (1996), Spurrett & Papineau (1999), Loewer (2001), Papineau (2001, 2013), Melnik (2003), and Kim (2005).

[2] Bishop (2012, 61), for example, argues that "physics neither proves its own completeness nor needs such a principle", and that CCP "makes no mention of anything but physical properties and laws", so that CCP "clearly must be a metaphysical principle".

[3] By Newtonian forces I will always mean the causes of acceleration of the particles that make up the physical system – not just the forces in classical Newtonian mechanics.

processes. Moreover, I contend that a viable model of mental causation can be formulated from the position of interactive dualism, within the frameworks of physical science, and point to some basic features that such a model must contain. If successful, this inference undermines the argument that is known as the most important pillar on which the doctrine of physicalism rests, but at the same time indicates the way in which it can be tested conclusively by means of contemporary physics.

2 The argument from causal closure of the physical

There are plenty of arguments for physicalism proposed by its proponents in recent decades, but none is more convincing than the argument from the causal closure of the physical, usually presented in the form given to it by Papineau (2001) and Kim (2005). It consists of three premises:

(C 1) *The causal closure principle:* If a physical effect has a cause, then it has a sufficient physical cause.

(C 2) *Mental causation:* Every mental event has a physical effect.

(C 3) *The causal exclusion principle:* If a physical event *x* has a physical cause *y*, than no other event can be the cause of *x* at the same time.

The conclusion of the argument is that all mental events are identical to the physical events.

The relata of causation are understood as *events*, taken in a Kimian sense, i.e. instantiations of properties at a time. The proposed formulation of the causal closure principle (C 1) gives us the opportunity to examine its sustainability under the spotlights of physics, which interprets events as measurable instantiations of properties of interacting particulars, such as particles, waves, fields etc. Effectively, (C 1) comes down to the claim that while construing a causal chain in order to explain a physical effect, we never need to leave the domain of the physical, since any event in the causal chain which leads to a physical effect must also be physical. The second premise of the argument, mental causation (C 2), simply denies the epiphenomenalist claim that mental properties are mere ‘nomological danglers’, with no power to

causally effect our behavior. Instead, it acknowledges the common sense judgement that mental events and states, such as our decisions and inclinations, have physical consequences. Finally, the causal exclusion principle (C 3) requires that physical effects and mental causes are not systematically overdetermined. While particular instances of physical properties may be caused by simultaneous occurrences of mental effects, the possibility that *every* instance of a physical property is overdetermined in the same way can be safely dismissed. Acceptance of these premises implies accepting the physicalist conclusion that mental and physical effects must be identical.

Understandably, not all philosophers have been convinced by this line of defense of physicalism. All of the premises in the argument are being widely contested in the literature. I will leave aside the discussion of objections to premises (C 2) and (C 3), because it is out of the scope of this paper and because I believe them to be well founded.[4] An interactionist is primarily interested in challenging (C 1), as the core claim of the argument.[5] The first objection is based on difficulties in defining 'physical', best expressed by Hempel's dilemma: if one is unable to unequivocally say what 'physical' is, it is hard to see how the principle of causal closure of the physical is to be defended. Some of the proponents of CCP adopted the approach known as the *via negativa* argument, which defines physical in terms of non-mental.[6] However, the oponents of this strategy, such as Gillett & Witmer (2001) and Bishop (2010), argue that the inductive support offered in its favour begs the question for physicalism. Another line of objections to the causal closure of the physical rests on the claim that it is a metaphysical principle lacking an empirical support; I will soon turn to the discussion of an important attempt made by Papineau to answer this claim.

Probably the most serious objection to CCP is the claim that current formulations of the causal closure principle are problematic, because they are either too strong, or too weak. If the formulation is too strong, it cannot be differentiated from the conclusion of the argument, which begs the question for physicalism, and at the same time makes obtaining an empirical support for it virtually impossible. On the other hand, a weaker formulation of the principle makes the argument invalid unless it is supported by a hidden premise with the effect that only physical properties and states are causally efficacious. The former approach reduces the causal closure of the physical to mere typicality condition (Bishop 2010, 2012), while the latter begs the question

[4] *Cf.* Garcia (2014, 97).
[5] For some important objections to causal closure principle and CCP see Lowe (2000, 2006), Gibb (2010, 2015), Tiehen (2015), and Saad (2018).
[6] For detailed elaboration of the *via negativa* argument see Spurrett & Papineau (1999), Montero & Papineau (2005), and Wilson (2006).

for physicalism. In both cases the argument is implausible. Gibb argued that "attempts to address this problem by revising the causal closure principle are problematic, as there is a reason to think that the resulting causal closure principles will lack empirical support" (2015, 629). To make matters even worse for the physicalists, recently Saad (2018) claimed that it is even possible to construct a causal argument for dualism. If correct, this claim would „refute the view that causal considerations prima facie support physicalism but not dualism" (2018, 2475).

Obviously, without a strong support from induction, CCP would have been just one of the strongly contested methaphysical arguments. This is why the following two inductive arguments, put forward by Papineau (2001), are perceived as the main reasons for the broad acceptance of CCP among philosophers:

(F 1) *The argument from fundamental forces*: All forces are reducible to a small number of fundamental forces which conserve energy; therefore, most probably this is also the case with hypothetical special, mental forces.

(F 2) *The argument from physiology*: Despite extensive physiological research in the last couple of centuries, no trace of special (non-physical) forces was found; therefore, special forces probably do not exist.

The law of conservation of energy plays the crucial role in the reasoning behind (F 1). Papineau believes – somewhat unconvincingly – that the fact that all the known physical forces reduce to a small stock of conservative forces gives us the reason to believe that the same must be the case with mental forces. The argument from physiology is broadly considered much more convincing. It rests on evidence that anomalous accelerations in living bodies have not been found yet, despite intense research. This undermines the credibility of the claim that some kind of vital or mental forces operate within them. Since causes of accelerations are forces, at least according to Newtonian mechanics, Papineau takes the combination of these arguments to imply that all observed processes within living organisms can be attributed to the actions of well-known and well-described physical forces. From this, he concludes that there are no irreducible mental forces.

Although Papineau's argument is based on the notion of force, it can be easily and without loss in generality rewritten in terms of the equivalent energy-based formalism, which is more appropriate for contemporary physics. It is well known from mechanics that potential energy of an isolated system depends only on its configuration variables, while the sum of potential and kinetic energy is conserved in time: $U + T = const.$ The forces acting in such

systems can be expressed as negative potential energy gradients,[7] $\vec{F} = -\nabla U$. Therefore, Papineau's conclusion about the lack of special forces comes down to the equivalent claim that there are no non-physical forms of energy in nature. Consequently, as noted by Gibb (2010), Papineau's arguments in favor of the causal closure principle, and thus indirectly of CCP, can be restated in terms of two physically equivalent energy-based premises, which I will express in a slightly modified form:

(E 1) *Every isolated physical system is conservative*. This means that in such systems, including living organisms and their brains to wich these arguments refer, energy, momentum and small number of other physical quantities are conserved.[8]

(E 2) *Non-physical forms of energy probably do not exist*. This is simply another way to say that special, particularly mental forces probably do not exist, i.e. this claim is equivalent to (F 2).

Papineau acknowledges that the validity of his argument can still be somewhat disputed, since the possibility of discovering some non-physical forces in the future cannot be completely excluded. Nevertheless, he believes that the failure of physiological research to detect anomalous accelerations within living organisms, indicating the existence of special forces, i.e. non-physical energy, renders the interactionist position extremely unconvincing.

Papineau's arguments have been arguably the most convincing and widely used arguments in favor of the causal closure principle, and consequently CCP, for nearly two decades, despite the fact that they have been contested from various angles. Gillette & Witmer (2001) argued that the argument from physiology was susceptible to a problem similar to the one

[7] For a discussion of the implications of the equivalence of these two formalisms and the ontological status of force see Wilson (2007).

[8] In this sense, (F 1) and (E 1) imply that living organisms and their brains can be studied as conservative systems, inasmuch as they can be isolated from the environment. An anonymous reviewer correctly objected that there is no good argument supporting the claim that biological organisms or their brains can be viewed as isolated systems. Indeed, the notion of an isolated system is just a useful idealization even when applied to much simpler physical systems, let alone to living organisms and their brains - extremely complex systems, hardly ever in equilibrium, with a constant flux of energy in and out of them. However, all that the defender of (F 2) needs to show is that physiological research has never revealed a violation of the *balance* of energy, or any other conserved quantity in an organic subsystem, which cannot be physically accounted for; the claim that conservation of energy would apply to those subsystems if they were perfectly isolated follows by induction. This is pretty much a standard procedure in physics, in cases where there is no way to isolate a subsystem inside some impenetrable boundary. In force-based terminology, this is equivalent to the claim that no unaccounted for forces have been found. For a recent discussion of the fact that physics regularly uses heavy idealizations of its properties and laws and of limits of such idealizations see Teller (2004).

expressed by Hempel's dilemma: current physiology is not complete, and we do not know what kind of entities future physiology will bring; it is conceivable that some of them will even be mental. Montero (2003) questioned whether the lack of evidence for the existence of fundamentally mental properties warranted the inference that such properties do not exist. Garcia (2014) agrees with the appraisals of cited authors that Papineau's conclusions are premature and question-begging, and rejects the strong version of the causal closure principle for both 'level' and 'domain' versions of that principle. All of these these remarks have not been convincing enough to notably sway the physicalists' belief, mostly because of the stubborn fact that deep into the 21st century no manifestation of non-mental forces has been found. However, the most important objection to Papineau's arguments, in my view, pertains to their incompleteness. In order to support CCP convincingly they need to be supplemented by a premise that Papineau left implicit and whose validity would prove to be questionable: that "the only efficacious states and causes are physical ones" (Bishop 2012, 62). This means that Papineau's arguments did nothing to amend the already mentioned deficiency regarding the weak form of the causal closure of the physical.

3 Premises about the nature of causation made explicit

The most explicit formulation of this deficiency was given by Gibb (2010), who argued that Papineau's argument, which leads from the premises (E 1) and (E 2) to the conclusion that there are no non-physical causes of physical effects, i.e. to the thesis (C 1), is not valid. The reason is that there is no reference to the notion of causation in the premises, although it is central in the conclusion of the argument. This omission implies that there are some implicit metaphysical assumptions regarding the mechanism of causation, the nature of which has been clarified by Gibb. She successfully demonstrated (2010, 374) that the argument becomes valid only if complemented by two further causal claims which I convey without alterations:

(G 1) *Physical affectability*: The only way that something non-physical could affect a physical system is by (1) affecting the amount of energy or momentum within it, or (2) redistributing the energy and momentum within it.

(G 2) *Redistribution*: Redistribution of energy and momentum cannot be brought about without supplying energy or momentum.

It is important to note that (G 1) and (G 2) are simply statements of the basic laws of dynamics, presented in a form which explicates the possible effects of the known physical, Newtonian forces, generalized to the non-physical domain.[9] Papineau limits the possibility of existence of hypothetical mental forces to only two categories: indeterministic and deterministic *Newtonian* special forces. Therefore, he explicitly denies the conceivability of special forces that could manifest their causal power in a non-Newtonian way, i.e. in a way which does not involve *immediate* acceleration production in living organisms, inexplicable by the action of any known physical force. In terms of energy-based formalism, this is equivalent to the requirement of the transference theory of causation that the mechanism of causal relations comes down to the transference of the conserved quantities – primarily energy and momentum – from causes to the effect. This is due to the fact that we identified forces with causes of accelerations of physical entities. It follows that Papineau's argument is implicitly based on the transference theory of causation,[10] which is far from generally accepted. It is debatable, for example, whether this theory can always be satisfactorily applied in the context of quantum mechanics.[11] Additionally, as will be discussed in the next section, it is hard to see how energy or any other physical property can be attributed to an immaterial and non-spatial thing such as mind, and even harder to imagine its transfer from mental to physical states.[12] Hence, an interactionist has no reason to accept the transference theory of causation as an explanation of the link between mental and physical properties and events.

---

[9] The claims (G 1) and (G 2) are true for *any* physical force, both in classical and in quantum domain. Otherwise, a force would violate either the laws of conservation of energy and/or momentum (by affecting a physical system without altering its energy or momentum), or the Second Law of Thermodynamics (by redistributing the energy and momentum without supplying them to the system, which means without doing work).

[10] This claim is valid weather the Papineau's argument is exhibited in the force-based or the energy-based formalism, because these formalisms are – as demonstrated – physically equivalent. The very notion of a conservative force implies, in order to be defined, the introduction of the idealization of an isolated system, as opposed to the open systems with boundaries through which there is a flux of energy and other conserved quantities. The work of the force is numerically equal to the increase of energy of the system on which the force acts, which is necessarily accompanied by the equal decrease of the energy of the system which performs the action. Consequently, saying that A causes an acceleration of B through a force is equivalent to saying that causation is manifested by the transference of energy from A to B. Therefore, by choosing to identify forces with causes of acceleration Papineau committed himself to the transference theory of causation. I thank the anonymous reviewer for requiring me to clarify this point.

[11] Consideration of various interpretations of specific elements of quantum theory designed in order to challenge CCP and construct an interactionist model of causation is beyond the scope of this paper. For some of the interesting ideas of this kind see e.g. Penrose (1994), Beck & Eccles (1992), Stapp (1993), and Schwartz et al. (2004).

[12] Averil & Keating expressed in no uncertain terms that "changes in the energy of non-physical things are undefined, i.e. there is no way of specifying the state of a non-physical thing in terms of the variables of physics" (1981, 105).

In accordance to the arguments presented above, the claim (G 1) is equivalent to the request that mental causation is manifested through the action of special forces which are strictly Newtonian. That means that first, these forces can reveal themselves only by causing anomalous accelerations in the system, and second, they must be conservative. The claim (G 2) additionally requires that redistribution of energy and momentum in the system can only be done under the influence of a Newtonian force, which means that work is done on the system. By definition, this results in a change of the total energy and momentum of the system. It follows that implicit claims in Papineau's defense of CCP require that nomological features of special forces are essentially identical to the features of physical forces. A dualist could reasonably ask how it is possible that the manifestations of mental and physical properties are so dramatically different if their instantiations are ultimately driven by the forces between which there are no nomological differences!

If this basic Papineau's requirement was accepted, the position of the interactive dualist would indeed become difficult to defend. She would face a nearly impossible task to answer the question as to how physics, despite all the successes it has achieved in the last couple of centuries, including the unprecedented development of its measuring techniques, has failed to register the slightest indication of accelerations in living organisms that are not explained by the action of some of the known physical forces. However, the dualist is in no obligation to agree with such a narrowing of the range of possible manifestations of mental causation.

Which options, then, are open to interactionism? Obviously, the premise of physical affectability (G 1) allows two conceivable ways that the effect of non-physical causes could be manifested on physical entities. The first possibility comes down to the action of the non-physical forces of the Newtonian character, whether they are deterministic or indeterministic, on a physical system, where the work of these forces results in changes in the total energy and/or momentum of the system. This is precisely the type of action that Papineau discusses in his works and that is implicitly based on the transference theory of causation. Interactionists who accept this challenge must demonstrate in a convincing way the presence of non-physical forces, i.e. non-physical energy, which cause anomalous accelerations within living organisms, and thereby change the energy content of the system. This would of course mean that the laws of conservation of fundamental physical quantities in systems such as human brains would no longer be valid. An additional energy of non-physical character would appear so that we could no longer think of such systems as isolated in a purely physical sense. It may be expected that in research of the last decades such a violation of the conservation laws in living organisms would be relatively easily

demonstrated. Papineau's argument from physiology has however shown that persistence of researchers has so far failed to produce any such results. Of course, Papineau himself emphasizes that his argument cannot be considered completely conclusive. We cannot yet exclude the possibility of detecting some manifestation of mental energy in the brain or nerve system. For the purposes of this paper, however, I will accept the validity of premise (E 1), assuming that the search for violations of the First Law of Thermodynamics and other fundamental conservation laws in the context of the explanation of mental causation is not a promising path for dualists.

The other way in which non-physical causation could possibly manifest, given the premises of physical affectability (G 1) and redistribution (G 2), is much more promising. Namely, an interactionist may try to explain causal power of the mental properties through redistribution of energy and momentum in the physical system without altering the total energy and momentum of the system. In other words, although an interactive dualist must accept (G 1), she can reject (G 2). In this way, she can remain committed to the First Law, but at the same time claim that non-physical causation can be manifested by the redistribution of energy and momentum without doing work, and thus without causing any anomalous acceleration in the observed physical system. In the spirit of such an approach, Broad (1925, 103-109) suggested the possibility that the soul manifests its causal power by redistributing energy in the brain, without changing its total amount.

This approach, however, has its price. Although Gibb (2010, 378), believes that by rejection of physical affectability and redistribution no physical laws would be violated, as the matter of fact nothing can be further from the truth because the rejection of (G 2) in living organisms inevitably leads to a local violation of one of the fundamental principles of physical science: the Second Law of Thermodynamics. This is a nightmare scenario for most physicists, because redistribution of the conserved quantities such as energy and momentum in a physical system without the expenditure of energy and momentum could mean, for example, transferring heat from a cold to a hot body, which is the realization of the perpetuum mobile of the second kind! Therefore, I need to turn to the discussion of the Second Law and the way in which the non-physical mind could bring about its violation.

4 The second law of thermodynamics and mental causation

The Second Law of Thermodynamics claims, in shortest, that total entropy S of an isolated physical system never decreases during spontaneous processes. Entropy, as a measure of system disorder, is growing in irreversible, and is constant during reversible processes in an ideal, closed system: $dS \geq 0$. This corresponds to the tendency of each physical system toward thermodynamic equilibrium. The Gibbs-Shannon entropy is associated with the probability of a microstate p_m by the expression

$$S = -k_B \sum_m p_m \ln p_m \tag{1}$$

where sum is taken over all microstates of the system.[13] Microstate represents the specific configuration of individual subsystems constituting the system, as opposed to macrostate, which is a description of the state of the system expressed as a function of its macroscopic properties: pressure, temperature, volume, etc. If all microstates corresponding to the same energy are equally probable, Equation (1) is reduced to the well-known Boltzman principle

$$S = k_B \ln \Omega \tag{2}$$

where $\Omega = 1/p_m$ is the number of microstates that make up the macrostate of the system, and k_B is Boltzmann's constant. The Second Law is therefore responsible for the direction of physical processes: spontaneous processes in nature, which are never reversible, will always be accompanied by an increase in entropy. This is an expression of the system's tendency toward equilibrium, i.e. the state that can be realized in the largest possible number of ways.

Now, mental states differ from physical ones by a variety of characteristics that clearly indicate their non-randomness. Their intentionality, directedness and aboutness are in sharp contrast with the spontaneity of the purely physical processes. In this sense, it is to be expected that causal actions of the mind oppose the tendency of the physical system to spontaneously transition to a state of maximum entropy, as required by the Second Law. Hence, the immediate effect of the presumed mental causation is not just a random redistribution of energy and momentum, but such that is accompanied by local decrease in entropy of particular subsystems within a living organism. The only way that the mind can produce this effect at some location in the living body without violating the conservation laws, and without altering the total amount of

[13] In quantum mechanics the entropy is given by Von Neumann's expression $S = -k_B \mathrm{Tr}(\rho \ln \rho)$ where Tr is a trace operator, and ρ is density matrix.

energy and momentum of the system, is by performing the redistribution of energy and momentum in the system with no force acting and no work being done – contrary to (G 2).

It follows that an interactive dualist, in order to reject CCP, accepting at the same time universal validity of the First Law of Thermodynamics in living, as in inanimate systems, must be committed to the following thesis:

(D) *Mental causation is manifested* by an intentional, directed, local decrease in entropy, realized by redistributing the energy and momentum within the living system without altering the total amount of energy or momentum of the system.

This local disturbance of the thermodynamic equilibrium results in a tendency of the system to spontaneously return to it, in accordance with the Second Law. Reduction of entropy in any part of the system results in gradients of a number of physical quantities, and thus in the physical forces that tend to bring the system into balance. Such physical processes take place continuously, at each physical system, at the level of fluctuations. All that the mind would be required to do is to bring about small-scale, local correlations of neural processes, hardly distinguishable from fluctuations of physical quantities, in order to produce large-scale physical effects. Mental causation is, in this way, regularly accompanied by the phenomenon of correlated fluctuations with extremely small likelihood of spontaneous occurrence. Recognizing this, Eccles (1980, 1987) and Popper & Eccles (1977) argued that the structure of the brain is finely tuned and that small, mentally caused perturbations – or correlations, as suggested by this analysis – could have macroscopically significant effects.

The mind operating in this way would represent the almost literal realization of Maxwell's demon – an imagined creature that Maxwell introduced in his famous thought experiment (Maxwell, 1871, 308-309). The demon, without doing work, separates the gas molecules at speeds in a container divided by a barrier into two chambers, thus creating the temperature gradient. In this way, in contrast to the Second Law, it reduces the entropy of the system. Maxwell conceived of his thought experiment as an illustration of the statistical nature of the Second Law, implying that the validity of this law is only probable and that, if we are ready to wait long enough, its spontaneous violation is not only possible, but certain. The demon action is, on the contrary, intentional and directed, so the violation of the Second Law is its inevitable result. Clearly, if we accept Papineau's arguments (F 1) and (F 2), with the effect that there are no non-physically caused accelerations in living systems, the only remaining option left to the proponents of irreducible mental causation is to claim that (D) is true, which means that mind

operates in a way analogous to Maxwell's demon. That is not to say that brains should be seen as anything as simple as collections of molecules in thermal equilibrium; such an idealization was used by Maxwell just in order to make a point whose significance is much wider. What Maxwell's thought experiment has shown is that it is conceivable that an intelligent agent can, at least in principle, aspire to influence the state of a physical system in a way which directly contradicts the Second Law.

No wonder, then, that a large number of physicists after Maxwell devoted a lot of energy to finding ways to exorcize the demon and 'save' the Second Law. The earliest such attempts, starting with Smoluchowski (1912), consisted in the naturalization of the demon – its representation as a physical system in thermodynamic equilibrium with gas in the container. Smoluchowski and his like-minded successors contrived a multitude of imagined mechanical constructs that simulated the actions of Maxwell's demon and came to the conclusion that it was not possible to realize a perpetuum mobile of the second kind – a machine that would be able to transform heat into work without energy loss – using molecular fluctuation phenomena. An entirely different approach was used by Szilard (1929). He believed that any device simulating the actions of the demon produces entropy during the act of measuring molecular positions and velocities. The entropy produced is more than sufficient, according to Szilard's calculations, to compensate its decrease during the operation of the device. Brillouin (1953) showed, starting from Szilard's conclusions, that the entropy produced when obtaining information allowing the choice between n equally probable states is at least $k_B \ln n$. This means that while acquiring or modifying one bit of information, a minimal amount of energy $k_B T \ln 2$ is released into the environment of temperature T. Finally, thanks to Bennett's work, (e.g. in Bennett, 1987), an approach based on the idea launched a couple of decades earlier by Landauer (1961) gained popularity, according to which a minimum of $k_B \ln n$ of entropy is generated not during acquisition, but during the *erasure* of information from demon's memory. This variant of exorcism based on the information theory is known as Landauer's principle.

However, Earman & Norton (1998, 1999) powerfully demonstrated, in their detailed analysis of the history of efforts to exorcize the demon, that all such attempts are ineffective. These efforts miss the point, that is, the very motive for Maxwell's thought experiment: the demonstration of immanent restrictions of the Second Law, which originate from its statistical nature. The authors summarize their conclusions by introducing a dilemma that will prove important for our analysis of mental causation:

> In so far as the Demon is a thermodynamic system already governed by the Second Law, no further supposition about information and entropy is needed to save the Second Law. In so far as the Demon fails to be such system, no supposition about the entropy cost of information acquisition and processing can save the Second Law from the Demon (Earman & Norton 1999, 1).

Acceptance of the first horn of this dilemma ('sound' horn, in the terminology of the authors), by considering the demon as a physical system, *postulates* that such a system is subject to the Second Law. The same must apply to a combined system constituted by gas in container and the demon. It follows that any reduction in entropy in the gas must be accompanied by the same or greater increase in entropy in the demon. Whether this entropy price of gas molecules separation will be paid in the process of data acquisition, their erasure or in some other way, is ultimately only a technical issue. The authors therefore rightly conclude that in the event of taking the 'sound' horn of the dilemma Maxwell's demon has no chance of success, so that all efforts to defeat it have only a heuristic value. That value consists, among other things, in a clear demonstration of the futility of all attempts to construct a perpetuum mobile of the second kind. With this conclusion even Bennett agreed to the greatest extent, noting that

> [...] although in a sense it is indeed a straightforward consequence or restatement of the Second Law, it still has considerable pedagogic and explanatory power (Bennett 2011, 501).

If, however, we take the second ('profound') horn of Earman's and Norton's dilemma, we have to accept that the demon is not a part of the observed physical system and that its intelligence cannot be described by the known physical laws. Accordingly, the Second Law can be applied neither on the demon, nor on the combined system gas-demon. In the authors' words,

> […] we need a new physical postulate to ensure that the Second Law holds for the combined system. Any such postulate, either a general one or one specifically relating entropy and information, requires independent justification. We do not believe that the literature has succeded in providing such justification (Earman & Norton 1999, 2).

Clearly, the 'profound' horn of the dilemma has a special significance for our consideration, because if Maxwell's demon is viewed as an intelligence that interacts in an unspecified way with a physical system, its actions explicitly represent the manifestation of the causal power of the mind! The container can be replaced with any isolated system consisting of two or more initially equienergetic states, and the molecules with its component subsystems. The analysis of Earman and Norton points to the conclusion that the causal effect of the mind on the

physical system, which is not accompanied by a change in energy and momentum, can only occur if the mind is not viewed as part of the system, but acts as an external agent that redistributes the energy and momentum within it in a non-physical way, which is in line with the formulation of the interactionist thesis on the manifestation of mental causation (D). All that an external observer can detect is an inexplicable correlation in the redistribution of energy of the molecules in the container. This correlation is not caused by the effects of any new physical forces, but by the selection rules established by the demon which, by altering the boundary conditions in the system, increase the probability of finding a fast molecule in one chamber, and a slow molecule in another. Correlation results in a decrease in entropy in the system, for which it is not possible to determine whether it was accompanied by an increase in the entropy of the mind, as in the cases discussed by the authors in the context of the 'sound' horn of their dilemma. If dualism is right and the mind cannot be identified with the brain, the nature of the mind is not known to us. If it is not composed of subsystems that can be analyzed by statistical methods, then it is not clear how the entropy – or for that matter, any other property we associate with physical entities – could be attributed to it. The quoted conclusion of Earman & Norton, that we need "a new physical postulate to ensure that the Second Law holds for the combined system" is valid even in the case of the interaction of the non-material mind with the body, the difference being that the new postulate in that case cannot come from current physics. If mental properties are irreducible and if there is a nomic dependence between physical and mental properties, as interactionism claims, than this dependence is contained in as of yet unknown psychophysical laws, not unlike those advocated by Chalmers ([1996])[14] from the position of his naturalistic dualism. I will turn to the discussion of the character of these laws in the next section.

As far as I know, the first elaborate attempt to link mental causation to the violation of the Second Law was made by Morowitz (1987), who claimed that "it is (…) impossible to dissociate the mind's information from the body's entropy. Knowledge of that state of the system without an energetically significant measurement would lead to a violation of the second law of thermodynamics." (1987, 271). The analysis of the problem of Maxwell's demon presented here shows not only that mental causation is not bound to comply to the Second Law, but that the insistence on this would mean nothing less than the *a priori* postulation of physicalism. The

[14] Chalmers, though, built his naturalistic dualism on the assumption that only phenomenal, but not intentional properties are irreducible, so that under psychophysical laws he means those that connect the fundamental phenomenal properties with the physical. In this work I proceed from the standpoint that the dualist model should be consistent and include *all* mental properties in a unique causal mechanism. On this point I concur with Bishop who, commenting on causation at work in human behavior, stated that "focusing on intentional states is likely too narrow for us to gain a fuller insight into meaningful human action". (2012, 59).

requirement for the mind to conform to the Second Law is equal to the claim that the mind is only a material subsystem within the wider physical system, and that any entropy decrease of the system that is caused by its action must be compensated by at least the same increase in the entropy of the mind, which in this case is identified with brain. Of course, a dualist does not have to accept this claim. The explanation of the mental causation based on the thesis (D), contrary to CCP, is perfectly conceivable. The intelligibility of this dualistic explanation depends mainly on its ability to account for the deeper nature of the relation between mental and physical properties. Gibb is right to ascertain:

> The demand for a causal mechanism, and hence the demand for an answer to the question of how, for example, mental events render physical events non-coincidental, is acceptable only if one advances a theory of causation that analyses causation in terms of underlying non-causal processes associated with causation – processes that can be appealed to in order to explain how a cause brings about its *direct* effect (Gibb 2010, 381).

Since the effect of psychophysical causation manifests itself in a local decrease in entropy, it follows from the definition of entropy that this is only possible through the redistribution of the probability of states of the observed system. The mechanism of causation, therefore, comes down to an explanation of the way in which the mind carries out such probability redistribution, which in the macroscopic domain manifests itself by the redistribution of energy, momentum and other physical parameters of the system. I believe that no dualistic model, after the necessary abandonment of the transference theory of causation, can be convincing without being able to explain the details of this mechanism. I will next point to some basic general features that I believe such a model must contain in order to be intelligible.

5 Features of the conceivable naturalistic account of mental causation

The core belief of interactive dualism demands that

> (…) there are two essentially different kinds of *property* out in the world. (…) Genuine property dualism occurs when, even at the individual level, the ontology of physics is not sufficient to constitute what is there (Robinson 2017).

An interactive dualist must, in giving the explanation of mental causation, commence with the assumption that there is a class of mental properties that cannot be reduced to physical, but that interact causally with physical properties through psychophysical laws. It is rational to assume that the hierarchy of mental properties is similar to the hierarchy of physical properties in that there is a finite number of mutually independent, fundamental mental properties from which the mental properties of a higher order can be derived, just as all physical properties can in principle be deduced from a small number of basic physical parameters such as mass, charge, spin, etc. Here, I treat properties in a way common for natural sciences and natural ontology, with higher-order properties understood as conjunctive. Some property elementarists might still question this approach, but as Francesco & Swoyer ascertained, "it is now widely acknowledged, even by minimalists, that at the very least some higher-order relations are needed to confer structure on first-order properties." (2017, Sec. 7.1). Properties of an atom are much more complex than those of an electron, because its state is determined by much greater number of mutually independent, but nomically related parameters, or degrees of freedom, than the state of an electron. Hence, the properties of an atom are of a higher order in the hierarchy of physical properties than the properties of an electron, while the properties of a table of which this atom is a part are even more complex – and so on. The same can be said about the mental states, whose complexity in the observable realm is also hierarchical in nature. It is only rational to assume that this hierarchy continues downward. The following analysis stands even if we manage to identify a set of basic psychological properties as prima facie fundamental.

Hence, the claim that mental causation is manifested in a living organism reduces to the thesis that the state of this complex psychophysical system can be described by means of an extended set of mutually independent state variables, composed of a finite number of physical $\{q_i | i = 1, 2, \dots, k\}$, and mental $\{m_j | j = 1, 2, \dots, l\}$ state variables. Which physical parameters will be considered as state variables depends on the type of physical system and context.

In order to get as close as possible to the level of the organization of matter at which the supposed causal effects of mental to physical properties take place, we confine ourselves to mechanical systems. Here the state variables are the so-called canonical or phase variables – the coordinates and components of the momentum of all the particles that make up the system. The dynamical laws in classical mechanical systems are deterministic in the Laplacian sense, which means that from the known values of variables at one point of time one can uniquely, in principle, determine the state of the system – that is, the values of all variables – at any other moment.

Mathematically, this means that there are functions of state variables of the form $f_i(q_1, \dots, q_k)$, where $i = 1, 2, \dots, k$, such that

$$dq_i/dt = f_i(q_1, \dots, q_k). \quad (3)$$

These functions express physical laws that describe the interdependence of physical variables and thus uniquely, in principle, determine the evolution of the physical system. When all the initial conditions are known, the system (3) is exactly solvable. Naturally, in a realistic, complex system we almost never know all the relations comprising the system (3) and all of the initial conditions at a time, so that it is almost never possible to solve the Equations (3) exactly.

This is the reason why complex, many-particle systems are usually described by the methods of statistical mechanics. The current state of the system is represented by a point in the phase space whose coordinates are canonical variables, and its dynamic evolution – the trajectory of the phase point in this space. The states of the system represented by the phase points in a phase cell of small enough volume cannot be differentiated. Each of these elementary cells corresponds to a specific microstate of the system. The probability of a macrostate is determined by the number of elementary cells in the phase volume corresponding to this state, that is, the number of microstates that make up the macrostate.

Now, deterministic laws that would establish functional relations between the mental and physical parameters of the system are not known and the very possibility of their existence is controversial. On the other hand, an argument I made in Section 3 demonstrated the conceivability of mental causation accompanied by local decrease in entropy, i.e. by the violation of the Second Law which is, as we saw in Section 4, immanently statistical in character. Acceptance of the empirically corroborated claim that transference theory of causation is not applicable to the domain of mental causation leaves as the most plausible option for the interactionist the construction of a probabilistic explanation of mental causation. Therefore, I believe that the most promising interactionist strategy implies the commitment to the following thesis:

(MC) *The causal power of mental properties is instantiated* by altering the state probability distribution within the living system, which leads to the redistribution of energy, momentum and other conserved quantities, without altering the total amount of energy or momentum of the system.

At a first pass, the interactionist has no reason to engage in a discussion of whether the probabilistic character of mental causation indicates an immanent property of the assumed psyscophysical laws, analogous to the way in which the quantum mechanics laws seem to be immanently indeterministic, or there is a deeper, deterministic level in their foundations. As was the case with the interpretations of quantum mechanics, that is a problem that need not be solved immediately. Instead, for the interactionist it is of primary interest to demonstrate that it is possible to build a convincing naturalistic account of mental causation based on (MC). If successful, this demonstration would demote CCP to the status of a weakly grounded hypothesis and thus undermine the basis on which physicalism rests. By the end of this section I will attempt to outline some of the basic features that an interactionist account should contain.

The probability of finding a system that exchanges energy with the environment, i.e. is a part of a canonical ensemble, in the macrostate characterized by energy E_i can be obtained from

$$W(E_i) = G(E_i)p_m(E_i) \tag{4}$$

where $G(E_i)$ is the number of microstates with the energy E_i, or the statistical weight of this energy level, and $p_m(E_i)$ – the probability of a microstate corresponding to that energy.[15] It should therefore be shown that the dualistic account can explain the way in which the introduction of irreducible mental variables leads to a change in the number and/or probability of microstates, bearing in mind that (MC) implies that the presence of mental state variables does not change the energy of the system as a whole. It turns out that both possibilities are conceivable, provided the initial interactionist assumption that there are fundamental mental variables, which together with the physical state variables determine the properties of the living system, is true.

Let us first consider how the statistical weight $G(E_i)$ changes with the introduction of mental state variables $\{m_j | j = 1, 2, \dots, l\}$. Clearly the introduction of new degrees of freedom necessarily leads to an increase in the number of ways in which macrostate can be realized, thus changing its probability. The state *Pa* in which *a* has the physical property *P* now differs from the state *Ma* & *Pa*, where *M* is a relevant mental property. The situation resembles the splitting of energy levels into sublevels; however in this case levels do not differentiate in terms of energy, but in terms of the values of mental parameters. The state of the system can now be represented

[15] The discussion can be easily generalized to grand canonical ensembles, which exchange both energy and particles with the environment. Although such an analysis would be more realistic, given the open nature of the biophysical systems, for the sake of simplicity and illustrativeness I will limit myself to the consideration of canonical ensembles.

by a phase point in the generalized phase space, with as many new dimensions as is the number of mental state variables.[16] The phase volume in this generalized phase space with a larger number of dimensions can naturally be divided into larger number of phase cells than in the case of the physical phase space, which is equivalent to the increase of $G(E_i)$. Each of the phase cells corresponds to the specific generalized, psychophysical state of the observed system. The phase trajectories of the system in the generalized phase space are determined not only by physical, but also by psychophysical laws. There is always, obviously, a family of possible phase trajectories whose projections on a physical phase space, obtained by setting all mental variables to zero, are identical. These trajectories correspond to different ways of performing the same physical effect. It follows that mental causation amounts to the choosing between different phase trajectories, i.e. between different system dynamics'. In other words, the mind chooses what operation is to be done and how it is to be performed.

The way in which the introduction of mental variables affects the calculation of the probability of microstates $p_m(E_i)$ depends crucially on the nature of the mental variables and laws that determine their relation to physical variables – psychophysical laws. It is common knowledge that physical laws cause different systems to submit to different statistics. In such a way, classical statistics assume that particles can always be distinguished from one another and that each energy state can be filled by an arbitrary number of particles, while quantum statistics are based on the postulates of the existence of discrete states of the system, the identity of elementary particles, and the impossibility of simultaneously exactly determining the coordinates and momentum, derived from the Heisenberg uncertainty principle. The Pauli Exclusion Principle additionally applies to half-integer spin particles, or fermions, which leads to a significant difference between Fermi-Dirac and Bose-Einstein statistics. There is no reason to believe that it is impossible to determine the dependence of probability of microstates on all, including mental variables. Positing that the unobservable mental parameters have irreducible causal powers means that it is conceivable that they figure in the probability distribution function as independent state parameters, thus causing the redistribution of physical properties of the system, in accordance with the values ascribed to them by the mind. To the observer, this would appear as a consequence of anomalous correlations. On the other hand, the example of Maxwell's thought experiment shows that it is also conceivable that the mind could perform its causal action by changing boundary conditions, by means of introducing temporary selection rules that establish

[16] In a similar manner, in order to define the state of a system under certain contextual conditions Bishop introduces 'contextual topology', "constructed by picking out particular reference states and defining the appropriate observables for these states" (2012, 71).

constraints on the transition from one state to another. The selection rule by which the demon retains the fast molecules in one, and slow ones in the other chamber, using its power to control boundary conditions at the door dividing the chambers, is formally equivalent to redistribution of the probability of microstates corresponding to different kinetic energies of the molecules. Whatever the mechanism, the final result of the process is the redistribution of energy between the subsystems that the demon has performed without changing the total energy or momentum in the system. Knowing both the probability $p_m(E_i, m_j)$ and the statistical weight $G(E_i, m_j)$ would in principle allow the calculation of the probability of the state of the system in which the mental causation is manifested. In this way, the probabilistic theory of mental causation would intelligibly explain the local decrease of entropy that follows the instantiation of the causal power of the mind and pave the way to a better understanding of processes at the illusive boundary of the mental and physical.

The main task of interactive dualism, therefore, consists in finding nomic relations that will causally connect mental with physical properties and facts. Note that only after formulation of the comprehensive theory of mental causation will it be possible to individuate the unobservable mental properties that make up the class of state variables $\{m_j | j = 1, 2, \dots, l\}$. The aggravating circumstance is that all our measuring techniques and instruments are designed for measuring only physical parameters and rely on the application of known physical laws. In addition, the assumption that the mind affects physical systems without causing accelerations, imposed by (F 1) and (F 2), must also apply to the effect of the mind on all physical measuring instruments, which cannot therefore be expected to directly register the effects of mental causation.[17] Hence, the existence and causal powers of mental variables can be deduced only indirectly. The logical starting point is the empirical study of anomalous correlations between neural processes, as well as between different mental states and their physical manifestations and consequences. A dualist should hope that over time this analysis will become more refined and will ultimately enable the discovery of at least provisional, probabilistic laws of psychophysical causation, despite the aggravating fact, well-known from physics, that a system can have the same observable properties for a very wide range of values of unobservable parameters, or the theoretical models in which they appear. The starting point in this analysis is to establish classes of correlated physical, especially neural events, leading to particular behavioral manifestations. If seemingly random correlations factually lead to directed, mentally preconceived physical effects, it is hard to see how these facts can be physically explained. An interactionist may appeal to the

[17] *Cf.* Lowe (2008, 74).

inference from the best explanation in order to account for this anomalous correlation by the presence of the latent common cause in the form of a mental variable, event or state M. For example, if physical – particularly physiological, or neural – events A and B are regularly correlated in a way that $P(A\,\&\,B|M) \gg P(A)\cdot P(B)$, viz., if their correlator's value is inexplicably high, i.e. $C(A,B) = P(A\,\&\,B|M) - P(A)\cdot P(B) \gg 0$, an interractionist may rightfully expect that a thorough enquiry of this correlation should result in a better understanding of M and its causal role. Since mental properties are unobservable, this indirect, physics-based enquiry is the only way to determine their existence, individuate and study them. On the other hand, if the conclusion of this research is that all anomalous correlations can be physically explained, we will be able to convincingly verify CCP – bearing in mind our earlier inference that there are probably no anomalous accelerations either. This would certainly represent a major triumph of the ontology of physicalism. I believe that it is very important that both possibilities can be tested conclusively within the framework of the theoretical system of physics.

In short, individuating mental variables by their causal roles and determining the method of their quantification are the first important tasks of the interactionist theory of mental causation. Given their unobservable nature, the only way to accomplish this task is to establish their nomic, probabilistic relations with the physical parameters whose measurable change they cause, by means of the state probability distribution function $f(q_i, m_j)$. Possible discovery of fundamental, in-depth laws that would allow predicting the course and outcomes of psychophysical processes is a matter of the future.[18] In other words, an interactionist may choose to settle for a coarse-grained description at a first pass and leave finer-grained questions for the future.

The significance of the described correlations was emphasized by Lowe (2000, 2006, 2008), one of the most prominent proponents of interactive dualism. According to him, mental causation is manifested as a "unifying factor explaining why those apparently independent causal chains of neural events should converge upon bodily movements in question" (Lowe 2000, 581). Although Lowe did not provide details of his mechanism of mental causation, his main ideas are compatible with the general features I have outlined here. He basically claims that convergence of neural events can be understood as a formal property of certain causal trees. The mind, in a way,

[18] Chalmers (1996) suggested *the principle of structural coherence* as the bridge from features of physical processes to features of experience, in order to explain the structure of the phenomenal domain. His idea is based on the assumption that consciousness supervenes on the functional organization of the brain. Although an interactive dualist will hardly agree with the latter assumption and consequently reject the explanatory value of Chalmers' principle, she may find the suggestion of structural coherence rather useful as a means of quantifying mental parameters, because it establishes a functional link between physical and mental properties.

charts a number of pathways through the maze of independent possible causal events, leading them to the point of convergence – the intended physical outcome. Lowe in this way attempts to replace event-causation by fact-causation: "What is brought about is not an event, but a fact or state of affairs" (2000, 582). I believe that in order to be convincing a model of mental causation should go a step further and provide the account of this fact-causation by advancing an explanation of an underlying event-causation. The way that the mind 'sees' the path through the maze of possibilities, as laid out here, is to 'mark' that path by redistributing the values of mental parameters and consequently altering state probability distribution, in accordance with the rules arising from psychophysical laws.

It is important to note that subsystems in the discussed sense need not necessarily be individual particles, but also more complex systems, such as ion channels, synapses or entire neurons. The persuasiveness of the interactionist explanation of mental causation depends crucially on whether it can solve the problem of the identification of the mind-body nexus, i.e. the site where the causal interaction between mental and physical properties takes place. It has been sought for in the presynaptic membranes of the axons (e.g. Beck & Eccles 1992), in microtubules – cylindrical protein lattices in the neuronal cytoskeleton (Penrose 1994) and other places. It is not my intention to give an assessment of existing interactionist models here, or to propose new ones. The aim of the arguments presented is more modest: to show that the construction of a convincing and intelligible interactionist model is conceivable and realistic, and that CCP does not constitute a serious obstacle to such a possibility as long as the anomalous correlations remain physically unexplained.

To the observer who, by using physical measuring instruments and his own senses, tries to directly perceive mental parameters and mental causation, only physical properties are accessible, so that her description of events within the living system corresponds to the viewpoint of current physics: all events are physically causal and their outcome seems purely coincidental, driven by anomalous correlations of neural events. According to her, the proceedings in the system are purely physical in nature and thus confirm CCP. It is clear that this description may easily be incomplete and that it is impossible to give an account of mental causation by the means of current physics. Moreover, physics is clearly indifferent to the possibility of existence of such causation. Although it is 'blind' to mental properties, it does not *a priori* oppose them. If empirical research shows that the significant properties of living systems – primarily brains – cannot be described without their introduction, physics will simply incorporate them into its theoretical system. In this case, CCP and Papineau's arguments will fall victim to Hempel's

dilemma: being based on the theoretical definition of the term ‘physical’ that originates in current physics, they fail if the theoretical system of physics is expanded by the introduction of irreducible mental properties and psychophysical laws.

6 Conclusion

If we accept the inductive claim (F 2) with the effect that there are no non-physical forces, i.e. no non-physical forms of energy as claimed in (E 2), the only conceivable way in which mental causation could be manifested is the non-physical mind that works analogously to the Maxwell’s demon: by redistributing energy and momentum within the physical system, without changing their total amount, in accordance with (MC). It can only do that by altering state probability distribution in the system, which results in local violation of the Second Law of Thermodynamics. Psychophysical laws that govern this process are prima facie probabilistic. The conceivability of such causal mechanism, at the very least, makes the causal closure principle (C 1), and thus the conclusion of the argument from CCP that all mental events are identical to physical events, inconclusive. Physics does not favor this thesis over the claims of interactionism.

According to my exposition of general features that an intelligible interactionist account should possess, the causal power of the mind stems from its ability to directly perceive mental states and to redistribute physical properties in the body, especially in the neural system, by changing the number and probability of microstates of the system. The mind may be able to do this by altering state probability distribution function, or by establishing selection rules that would dictate the way of populating unobservable mental states, analogous to the way in which Hund’s rules or Pauli Exclusion Principle dictate the way of filling the atomic orbitals with electrons. The mind, however, establishes these rules temporarily and intentionally, in accordance with still unknown psychophysical laws. Hence, it leads to a decrease in entropy and local violation of the Second Law, without changing the total energy and momentum in the system. Redistribution of physical quantities in a particular system can lead to physical consequences in the form of bodily movements that do not occur coincidentally, as one could infer from a physicalist account, but as the result of the intentional action of the mind. The system soon returns to equilibrium under the effect of purely physical laws and during this process its entropy increases again. Ultimately, the described mechanism corresponds to the known activity of the mind: its actions are intentional and directed, and come down to controlling the probabilities of physical actions. The effective

causes of bodily movements are physical forces, whose actions result in the dissipation of energy that contributes to the increase in the overall entropy in the Universe.

The main motivation behind this work has been the perception that we are not much closer to the idea how the mind interacts with the body than a century ago. It is of little avail to discuss the merits of various metaphysical arguments concerning mental causation without even suspecting how that causation may manifest. The purpose of my argument is not only to outline the features of a conceivable interactionist model of mental causation, but also to indicate the way of distinguishing the manifestations of the effects of physical and mental causes. I believe that no interactionist account will be considered intelligible until this demarcation is made. The final answers to questions concerning the existence and nature of irreducible mental properties and mental causation can only come from induction and depend on the discovery of hypothetical psychophysical laws. Establishing mental properties and mental causation as the basic features of the world would be one of the most significant extensions of our insights into the nature of reality. The failure to confirm their existence and irreducibility would convincingly verify CCP and thus the ontological doctrine of physicalism. Either way, I believe that the question can only be resolved by physics.

References

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

Beck, F., & Eccles, J. (1992). Quantum aspects of brain activity and the role od consciousness, *Proceedings of the National Academy of Sciences of USA*, 89, 11357-11361.

Bennett, C. H. (1987). Demons, Engines and the Second Law, *Scientific American*, 257, 108-116.

Bennett, C. H. (2011). Notes on Landauer's principle, reversible computation, and Maxwell's demon, *Studies in History and Philosophy of Modern Physics*, 34 (3), 501-510.

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.

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

Broad, C. D. (1925). *The Mind and its Place in Nature*, Routledge & Kegan Paul, London.

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

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

Francesco, O. & Swoyer, C. (2017). "Properties", *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/properties/>.

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.

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

Hempel, C. (1980). Comments on Goodman's Ways of Worldmaking, *Synthese*, 45, 193-200.

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. (2001). From Physics to Physicalism, in Gillett, C. and Loewer, B. (eds.) *Physicalism and its Discontents*, 37-56, Cambridge: Cambridge University Press.

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

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.

Montero, B. & Papineau, D. (2005). A defence of the *via negativa* argument for physicalism, *Analysis* 65 (3), 233-237.

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

Papineau, D. (2001). Rise of Phisicalism, 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.

Penrose, R. (1994). *Shadows of the Mind*, New York: Oxford.

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

Robinson, H. (2017). Dualism, *The Standard Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.standard.edu/archives/fall2017/entries/dualism/>.

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

Stapp, H. P. (1993). *Mind, Matter and Quantum Mechanics*, New York: Springer-Verlag.

Schwartz, J. M. et al. (2004). Quantum physics in neuroscience and psychology: a neurophysical model of mind-brain interaction, *Philosophical Transactions of the Royal Society B*, published online, 1-19.

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

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

Stoljar, D. (2017). Physicalism, *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/physicalism/>.

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

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.

Teller, P. (2004). The law-idealization, *Philosophy of Science*, 71, 730-741.

Wilson, J. (2006). On characterizing the physical, *Philosophical Studies*, 131, 61-99.

Wilson, J. (2007). Newtonian Forces, *British Journal for the Philosophy of Science*, 58, 173-205.