UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE FILOSOFIA
PROGRAMA DE PÓS-GRADUAÇÃO

Epiphenomenalism looming large: "mental quausation" and the threat of exclusion

Victor Nicolau Sholl de Freitas Lima

São Paulo
2022

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE FILOSOFIA
PROGRAMA DE PÓS-GRADUAÇÃO

Epiphenomenalism looming large: "mental quausation" and the threat of exclusion

Victor Nicolau Sholl de Freitas Lima

Dissertação apresentada no Programa de Pós Graduação em Filosofia do Departamento de Filosofia, Letras, e Ciências Humanas da Universidade de São Paulo, para obtenção do título de Mestre em Filosofia.

Área de concentração: filosofia da ciência

Orientador: Osvaldo Frota Pessoa Junior

Versão Corrigida

São Paulo

2022

Catalogação na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

L732e Lima, Victor Nicolau Sholl de Freitas
Epifenomenismo à espreita: "quausação mental" e a ameaça da exclusão / Victor Nicolau Sholl de Freitas Lima; orientador Osvaldo Frota Pessoa Junior - São Paulo, 2022.
95 f.

Dissertação (Mestrado)- Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. Departamento de Filosofia. Área de concentração: Filosofia.

1. FILOSOFIA DA MENTE. 2. CAUSALIDADE (FILOSOFIA). 3. METAFÍSICA. 4. FILOSOFIA ANALÍTICA. I. Pessoa Junior, Osvaldo Frota, orient. II. Título.

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS

ENTREGA DO EXEMPLAR CORRIGIDO DA DISSERTAÇÃO/TESE

Termo de Anuência do (a) orientador (a)

Nome do (a) aluno (a): VICTOR NICOLAU SHOLL DE FREITAS LIMA

Data da defesa: 04/11/2022

Nome do Prof. (a) orientador (a): OSVALDO FROTA PESSOA JR.

Nos termos da legislação vigente, declaro **ESTAR CIENTE** do conteúdo deste **EXEMPLAR CORRIGIDO** elaborado em atenção às sugestões dos membros da comissão Julgadora na sessão de defesa do trabalho, manifestando-me **plenamente favorável** ao seu encaminhamento ao Sistema Janus e publicação no **Portal Digital de Teses da USP**.

São Paulo, 2 / 8 / 2023

(Assinatura do (a) orientador (a)

I thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the funding provided through my master's scholarship. I also thank the members of the department of philosophy and of the "Serviço de Pós-graduação" of the Faculty of Philosophy, Languages and Human Sciences of the University of São Paulo (FFLCH-USP) for their service, especially with the difficulties posed by the COVID19 pandemic and by my being abroad in the final semesters of my master's.

My advisor, Osvaldo Pessoa Jr., has been a guiding figure for me since I was still studying physics, before switching to philosophy. I am deeply thankful for his countless helpful contributions and for overseeing and guiding my academic trajectory – as professor, advisor, and friend.

I thank the members of my master's committee for the helpful comments made and for understanding the eventual changes in schedule.

I thank my mother, Juliana, for her limitless love and support, and for being the interlocutor (and test subject) for many of my philosophical ideas. I would not have made it this far and would not have been who I am without her.

Friends are one of life's greatest gifts, and I am lucky to have had many edifying and intellectually stimulating relationships. There are two friends whom I owe special thanks to. I am deeply indebted to Otávio Raposo Jr. If it weren't for that discussion we had one night about the nature of qualia (and the many conversations we had in following years), I am not sure if I would have seen the value of philosophy soon enough for it to have the impact it had on my life. The fact that I went on to write on the metaphysics of mind and that I pursued the path of philosophy in general can be traced to that very night, and my life will forever bear the mark of his influence.
I also thank Gregory Gaboardi, for his countless contributions throughout my formative years. As a role model, he has greatly influenced my way of writing and doing philosophy.

In memory of Jaegwon Kim

Resumo

LIMA, V. N. S. F. **Epiphenomenalism looming large**: "mental quausation" and the threat of exclusion. 2022. Dissertação (mestrado) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2018.

Esta dissertação é sobre causação mental, em especial, sobre o chamado problema da exclusão causal. Tornado famoso por Jaegwon Kim, o problema consiste em uma tensão que parece haver entre alguns pressupostos geralmente vistos como plausíveis, mas que não poderiam ser todos verdadeiros ao mesmo tempo. Em linhas gerais, o problema pode ser fraseado da seguinte forma: uma vez que sempre há uma causa física disponível para explicar a ocorrência de qualquer efeito físico, então, se estados mentais não forem redutíveis a estados físicos, como poderia haver causas mentais de efeitos físicos sem que haja um tipo problemático de sobredeterminação causal (i.e, mais de uma causa suficiente ao mesmo tempo para o mesmo efeito)? Capitalizando em cima deste problema, há um tipo de argumento (o argumento da exclusão causal) que coloca o foco da tensão na premissa não reducionista, visando concluir que teses não reducionistas sobre a mente levariam ao epifenomenismo (a tese de que há estados mentais que não causam nada). Minha dissertação pretende mostrar que a ameaça posta pelo problema e argumento da exclusão é mais abrangente e difícil de lidar do que se costuma supor na literatura. Para tanto, defendo que teorias que tentem fugir do epifenomenismo devem preservar a noção de "quausação mental" apresentada por Terrence Horgan, e ofereço uma nova formulação do argumento da exclusão centrada nessa noção. O projeto da dissertação é dividido em três partes: na primeira, introduzo alguns conceitos e tópicos mais gerais sobre causação mental, apresento o problema e o argumento da exclusão e o que estaria em jogo com eles. Na segunda parte, introduzo a noção de "quausação mental", defendo sua importância e discuto maneiras de capturar de modo mais preciso a ideia geral por trás dela. Na terceira parte, apresento uma nova versão mais abrangente do argumento da exclusão, apresento e respondo uma possível objeção vinda da teoria de causação como "fazer diferença" defendida por Menzies & List, generalizo o problema da exclusão na forma do Dilema Epifenomenista e do Trilema da Eficácia Mental e discuto algumas implicações devastadoras do resultado.

Abstract:

This masters' dissertation is about mental causation, in specific, about what is called the causal exclusion problem. Made famous by Jaegwon Kim, the problem consists in an apparent tension between certain premises that are generally assumed to be plausible, but that could not all be true at the same time. In broad terms, the problem can be formulated like this: since there is always a physical cause available to explain the occurrence of every physical effect, then, if mental causes are not reducible to physical causes, how can there be mental causes of physical effects without implying that some form of causal overdetermination occurs? Building up on this problem, the causal exclusion argument is an argument that puts the tension on the nonreductionist premise, seeking to conclude that nonreductionist theses about the mind would lead to epiphenomenalism (the thesis that there some mental states that do not cause anything). My dissertations seeks to show that the threat posed by the problem and argument of exclusion is more serious than what is usually assumed in the literature. To accomplish this, I defend that accounts of mental causation that seek to avoid epiphenomenalism must preserve the general notion of "mental quausation" defended by Terrence Horgan. This project is divided into three parts: in the first, I introduce some concepts and topics about mental causation more broadly, the problem of exclusion, and the exclusion argument, discussing what would be at stake with them. In the second part, I introduce the notion of "mental quausation", defend its importance and discuss some ways to capture its general idea in more precise terms. In the third part, I introduce a new version of the exclusion argument that is wider in scope. I also present and respond to a possible objection to my argument in the form of Menzies & List theory of causation as making a difference. I then generalise the exclusion problem in the form of the Epiphenomenalist Dilemma and the Trilemma of Mental Efficacy, discussing some devastating implications of the result.

SUMMARY

Part I

The Threat of Exclusion

The relationship between mind and body has been an intriguing problem for millennia. The mind-body problem, as it is often called, is multifaceted and multidisciplinary in nature. A central subproblem is the one about the *causal* relationship between mind and body, on which philosophical work comes to the centre light. It is a central topic not only because causation is the prime mode of *interaction* we think of when reflecting on the mind-body relationship in general, but because investigating this causal relationship has implications (and is a source of insight) about the nature of the mind. This problem has immediate important ramifications: it has implications concerning knowledge, language, free will, our understanding of actions and agency, and the legal and moral questions about moral responsibility. In this work, I will focus on the more direct worries about the causal relationship between the mind and the physical world, comprising bodies and the environment. In specific, I will deal with questions concerning one direction of this kind of causal transaction: that from mind to world.

1. Mental Causation

The following is a list of examples I hope will help illustrate the kind of phenomenon this work is about:

1) *Thirst*: John wakes up in the middle of the night feeling thirsty. Because of this, he walks to the kitchen in search of water.

2) *Bitter drink*: Paula takes the first sip out of a new drink she decided to try. Feeling an extremely bitter taste, she decides to throw it out.

3) *Memory*: Frank suddenly remembers the time when he said something silly to a girl he was trying to impress in high school. He feels embarrassed.

4) *Belief*: Jimmy believes that the ball is hidden inside the cup on the left. Wishing to win the bet, he chooses that cup.

5) *Chocolate*: Walter feels a strong desire to eat chocolate from a brand he used to love in his childhood, but that no longer exists. As a result, he feels sad.

6) *Author*: Mary imagines the appearance and character traits of a man. After some time, she is satisfied with this exercise of imagination and Tobuscus, the main character of her novel, is created.

These are all examples of mental causation, which is what happens when there is a mental cause for an effect. By 'mental cause' I mean a cause that involves mental states, such as thoughts, beliefs, desires, memories and sensations[1]. This kind of phenomenon can occur in many ways, so it is useful to classify them according to their distinctions. For instance, there are differences in the kind of mental state featured *in the cause*: it can be a sensation, such as the cases of Thirst and Bitter Drink; a memory, such as the case of Memory; a desire, in the case of Chocolate or the kind of imaginative state featured in Author. Another important distinction has to do with the kind of *effect* brought about by the cause. Concerning effects, a fairly broad distinction is the one between physical and mental effects. Thirst and Bitter Drink would be cases of mental causes of physical effects (which I will refer to as 'mental-physical causation'; I am counting pieces of behaviour as physical effects), whereas examples 3-4 would be cases of mental causes of mental effects (which I will refer to as 'mental-mental' causation). It is common practice in the literature to divide cases of mental causation between mental and physical effects *only*, but this distinction need not be exhaustive. We can try to come up with examples of effects that are neither easily classified as physical nor as mental. The most natural case I could come up with is the creation of fictional entities – hence the addition of 'Fiction'. Some philosophers argue that fictional entities are something like abstract entities, so, if that is the case, their "coming into being" would not be a mental occurrence[2]. Mental-physical causation is the kind that receives most attention in the literature. The main reason for that is because it is the kind of mental causation that most readily generates a problem when coupled with other commonly held beliefs about the world – the exclusion problem, to which I will turn to soon. Since this problem is the central theme of this work, I will follow the tendency and devote attention mainly to mental-physical causation. This does not mean that mental-mental causation will

[1] Although 'mental state' is a widely used term in the philosophical literature, one would rarely find an attempt to offer an explicit definition of it (an exception is Kim, 1982, where mental states are defined as objects having mental properties at a given time. In this account, they are the same as mental *events*). This lack of clarification is rarely a serious problem, however, since the general idea involves nothing too technical or esoteric. It is meant to be a straightforward notion one could satisfactorily grasp through examples of paradigmatic cases. For the purposes of this work, it will be enough to understand mental states as states a mind can be in, usually related to specific mental operations (such as thinking, believing, imagining, and feeling).

[2] It can surely be argued that fictional entities are ultimately mental (just as it can be argued that mental entities are ultimately physical), but what is important is that, in any case, they are not *obviously* reducible to either physical or mental stuff. For this reason, cases involving fiction should be awarded the status of a different type of mental effect, even if provisionally.

be completely ignored, however. I will briefly discuss some consequences of the kind of exclusionary worry that mental-physical cases motivate that extend to mental-mental cases as well.

An important thing to notice is that there is usually a great degree of simplification when describing cases of mental causation, so that some other factors are usually omitted in favour of only one. For instance, it seems that it is not the feeling of thirst alone that causes John's behaviour, but also his belief that there is water to be found in the kitchen and his desire to quench his thirst. The same applies to the case of Belief, where Jimmy's desire to win the bet is essential to understand his action. What to take from this is a subject of dispute: one could, for example, see this as a reason to interpret claims about mental causation as highly dependent on the specific phrasings we give – that what we mention and what we leave out is crucial to assess whether these claims are true or not. Alternatively, one could take this to show merely that we often simplify things when we speak of mental causation, without any deeper consequence. We do not need to try to extract any kind of important consequence from this for now. I only called attention to this point in order to indicate that the examples on the list I provided, which are common in our day-to-day discourse as well as in the philosophical literature, are often simplified. This indicates that classifying cases in terms of the kind of mental state featured in them is not as simple as it might appear and generally depends on what we want to focus on.

If we want to examine further what is involved in mental causation, we will have to deal with traditional questions raised about causation in general. Some examples: what kind of things feature as the terms (the *relata*) of the causal relation? What entities actually "do the causal work"? What does the causal relation consist in? What are the truth conditions of statements about causation? These are all questions about what is sometimes called "the metaphysics of causation", which comprises questions about the nature of causation and the kinds of entities it involves. This is the kind of philosophical question about causation this work will focus on, with emphasis on the special case of *mental* causation. This means I will not be pursuing answers to, say, questions about how we can come to know causal truths (the *epistemology* of causation), though of course positions on the nature of causation will have consequences about other types of question.

As one could imagine, many different answers have been given to these questions about causation in the history of philosophy. Considering, for example, the question about the kind of entity related by causation, some of the main candidates are: events, facts, objects,

states of affairs, aspects or properties, situations and tropes[3]. As for the nature of the causal relation, a central question is the one about the *number* of terms the relation has (or its "adicity", to use a technical term). Does it involve two terms? Three, four? Again, answers vary. While the "traditional" position is that causation is a relation between two terms (one being the cause and the other the effect), there are many accounts that postulate three or more terms, generally by focusing on a more fine-grained description of the roles played by specific entities in causal processes. One example is the relation of "quausation" formulated by Terrence Horgan (Horgan, 1989), which plays a pivotal role in this dissertation (as the title suggests). It is a four-place relation, relating two events and two properties ("quausal" statements are of the form "c *qua* F causes e *qua* G", where "c" and "e" are token events and "F" and "G" are properties). Many pages will be devoted in this work to discuss "quausation" in detail, especially in Part II.

At this point, it is easy to feel overwhelmed by the branching nature of questions about causation and the plethora of widely different (and often "exotic") positions and qualifications we find. I would like to shake off this feeling, however. We can ignore these difficulties for now in order to advance into the specific case of mental causation; I only wanted to note that these questions are there (even if only on the background) and that, in order to give a full account of any kind of causal phenomenon, it is necessary to answer them. I will come back to some of these in parts II and III, but since we can only cover so much about such a comprehensive topic as causation, we will have to make some assumptions in order to advance. I would like at least to indicate clearly the ones I will make. The usual position about the kind of thing causation relates is that it is *events*. Events are usually evoked because they are things that *occur* (and causation is usually taken to be a relation between occurrences). At least initially, I will follow this trend when speaking about causation in a broad and relatively neutral manner, that is, whenever I do not want or need to address the question about what specific kind of entities it involves. Analogously, I will assume in these neutral contexts that causation is a two-place relation for simplicity. I choose this approach because it avoids repeating wordy attempts at neutrality.

There is one more issue about causation in the broad sense that is important to address. A question that is very divisive from the outset is whether causation is something that happens *in the world*, something in which "the things themselves" are involved, or if it is

[3] Tropes are a very peculiar kind of entity that some philosophers invoke. It takes some time to grasp what they are supposed to be, but for now it suffices to understand them as *particular qualities*, such as "this fragrance of this particular daisy".

just a relation among the expressions or concepts that we use to make sense of phenomena, but that ultimately does not correspond to any real "connection" between things. This is a rough description of the age-old issue of causal realism (corresponding to the first position in the preceding sentence) vs causal antirealism (corresponding to the second). Contrary to the other questions mentioned earlier, it appears that we cannot simply set this one aside and get to the details later. It involves a very fundamental difference in thinking about causation, which in turn leads to very different ways of investigating it. In this dissertation, I will assume causal realism, making only some occasional remarks on implications for antirealist positions. In the end of Part III, however, I will briefly explain what I think is a very good reason to favour a realist view of causation.

I have been speaking of mental causes in general, but I must now announce that this work will focus on a special subtype of these causes. The central problem I will address in this work, the exclusion problem, is centred on those mental states that are deemed to be (or could be) *irreducible*. The exact details about this problem will be covered in the next section, but for now we can see that these are the most problematic cases because they posit a difference between mental and physical items. Once we have this difference in place, there is margin for a conflict between mental and physical causes, which is the heart of the problem we will explore. The mental states that most defy reduction are conscious states and those that have *intentionality* (that is, those that are *about* something, that have a content; examples are thoughts and beliefs). The type of mental state I will focus on are these *conscious* states, states that have a qualitative feeling associated with them. We can also refer to these individual states as sensations, qualitative states, *qualia* (following this terminological tradition in the literature) or phenomenal states (which indicates that the notion of "consciousness" in play here is that of "phenomenal consciousness" or "awareness"). Speaking in a more holistic manner, either when referring to the "mental faculty" associated with these states or to some collection of these states working together, we can speak of "consciousness". In conclusion, then, we can now state that this work will focus on the causal issues related to conscious states – or on the causal influence of consciousness, if we prefer – with emphasis on the special case of mental-physical causation of this type. Referring to my list of examples, this means we will focus on cases like Thirst, Bitter drink and Chocolate[4]. Mental causation

[4] Cases like Memory could also be cited as examples, but whether mental states such as memories, beliefs or "mere thoughts" have a qualitative component is a controversial topic.

in its broader sense will surely remain a topic of interest, but it is with these allegedly irreducible mental causes I will ultimately be concerned.

2. The Exclusion Problem

The exclusion problem has appeared in different forms and under different names in the literature; for instance, it has also been referred to as "the overdetermination problem" (Witmer, 2000) and even as "the mental causation problem"(e.g., Kim, 2005 and Gibb, 2013), a testament of its centrality to the topic of mental causation. The different formulations share a common characteristic, which is the tension between theses each of which seem individually plausible but that could not all be correct. Taking bits and pieces from formulations I liked best, this is the formulation of the problem I favour (Cf. Gibb, 2013 and 2015; Robb, 2013; Ámadóttir & Crane, 2013; Kim, 2005):

Consider these theses:

- Eficacy of the mental (EM): There are mental causes of physical effects.
- Causal Closure of the Physical World (CC): Every physical effect has a sufficient physical cause.
- Nonreductionism (NR): Mental causes are not reducible to physical causes.
- No Overdetermination (NOD): there is no systematic causal overdetermination (that is, there is no kind of effect that is always caused by more than one sufficient cause at the same time).

It is possible to see the problem just by reading the above list of (at least *prima facie*) incompatible theses, or at least to see where this is going. The reader might have some doubts about what exactly some of the statements listed above are supposed to mean, and I will comment on each of them soon. The gist of the problem should be easy to grasp, however, being independent of the specific details: if there is always a sufficient physical cause available to every physical effect and if mental causes are irreducible to physical causes, then it seems that the only way that there could be mental causes of physical effects is if they are causes *in addition* to the physical ones. But this is exactly what NOD blocks. Hence the incompatibility. One way to phrase the exclusion problem, then, is this: given that every physical effect has a physical cause (CC), then, if mental causes are irreducible to physical causes (NR), how can there be mental causation (EM) without systematic overdetermination (NOD)?

An important thing to note is that the tension involved in the exclusion problem can be captured by different phrasings. The key difference between them is which theses are

considered as established premises and which ones are considered hypothetical, the ones that, ultimately, are at stake. This leads to a difference in emphasis: the pressure lies on the theses assumed as hypotheses under consideration, that is, the ones that would figure in the antecedent of a hypothetical conditional. So, for instance, we could phrase the problem as "given that every physical effect has a physical cause and that mental causes are irreducible to physical causes, if there is mental causation, then there is systematic overdetermination". This way of stating the problem puts the pressure on EM and NOD; since the other theses are supposed to be established, either EM or NOD would have to go. We could add to the last phrasing an explicit commitment to NOD, so that the pressure would lie exclusively on EM. But notice that doing so radically alters the "spirit" of the problem: it would no longer be about a "tension" between the theses, turning instead into a simple *reductio* of EM.

I chose that phrasing I offered first, which assumes CC and considers NR as a hypothesis, for two reasons. The first is that it follows the traditional emphasis found in the literature of putting the pressure primarily on NR, which is arguably the most controversial of the theses involved. As we will see in the next section, this particular take on the problem is what motivates Jaegwon Kim's formulation of the exclusion *argument*. Stating the problem that way, then, seems appropriate as a first phrasing because it presents the reader with the usual emphasis found in the literature first. The second reason for preferring that formulation is that it is phrased as a *question*. This seems to best capture the "tension" element I have been alluding to in my characterisation of the problem, since it leaves it as a problem to be solved rather than a mere statement of a contradiction.

Before trying to examine the problem in detail, it is useful to clarify what all of the premises involved mean and what is at stake with them. This is the task I will turn to next.

Efficacy of the mental: This is the most straightforward of all the premises, so there is no need to say much to clarify what it amounts to. It simply states that cases of what I have been calling mental-physical causation, such as the examples Thirst and Bitter Drink from the beginning, actually happen; they are real phenomena. Since the idea that our mental states are able to affect what goes on in the physical world is something we generally take for granted and that plays a fundamental role in the way we conceive of ourselves and our relationship with the world, the only surprising thing revolving around EM is the thought that it might *not* be true. Think of the implications: if it were false, that means our thoughts and desires are not what generates our actions and that our feelings are not responsible for our behaviour. So it would be false that Paula tossed her drink because it tasted bitter, or that

John's late night trip to the kitchen was due to his feeling of thirst (or his desire to quench it), or even that someone cried because they felt sad. All of these very natural ways of understanding our interactions with the world would be illusions if the mental had no efficacy. The striking thesis consisting of the negation of EM is called *epiphenomenalism*. It plays a central role in this dissertation (again, as the title suggests), and I will return to it with a proper introduction later. For now, it is enough to have in mind that the mere threat of excluding mental-physical causation as a real item in the world should be enough to motivate a serious examination of the exclusion problem (hence the name; see Kim, 1993, p. 281).

Even though I said there was not much to clarify about EM, there are a few points concerning formulation that are worth mentioning, however. The first is somewhat broad, affecting NR as well: I used the term 'mental' without any qualification because I wanted to give a general formulation of the problem, as is usual in the literature. However, following the comment I have made before, I am particularly interested in the case of *conscious* mental states. So the argument can be read in this broadest sense, concerning all mental states. Doing so would render it weaker, since there would be less reason to accept the truth of NR. In any case, it is worth keeping in mind that the special case of conscious states is the one we will be ultimately interested in.

The second point addresses a worry some attentive readers might have had. Both CC and NOD are stated in terms of *sufficient* causes, whereas EM only mentions causes *simpliciter*. This could immediately be a way out of the problem: there is no tension, since the kind of cause EM is about need not be sufficient causes. This is not the case, however. I chose to phrase EM in terms of causes *simpliciter* just to arrive at a neat, slogan-like statement that conveys something that has a common-sense appeal to it. Adding "sufficient" would suddenly turn it into "philosophy talk" with some technical aura to it. I wanted to avoid that because, in the end, EM just is this very innocent and commonsensical idea we all are expected to accept. But I do mean, in the end, that the mental causes in question are sufficient causes in the same sense involved in CC and NOD. I will discuss just what this sense is in my comments on CC. Furthermore, I do not believe this is in any way at odds with a more commonsensical reading of EM, for it seems to me that this is precisely what we assume in our everyday musings about mental causation (however rare they may be). Ideally, I would need some kind of experimental evidence to back up my impression, but for now I would like to simply invite the reader to consider what the alternative is. The usual contrast to sufficient cause is *necessary* cause, and, just to keep with the simple spirit, that simply means a cause that adds a necessary condition for the occurrence of an effect, but does not on its

own guarantee that it *will* occur. So if our everyday notion of mental-physical causation, the one behind EM, were not one of *sufficient* causation, than it seems that the alternative would be centred on necessary causes[5]. But this would mean that EM (and our usual, non-philosophical take on the matter) would amount to the idea that our mental states provide a necessary element to the occurrence of physical effects, but do not determine or guarantee the fact that they happen. So they either happen in virtue of something else that guarantees it (either in conjunction to this necessary element provided by the mind or not) or not in virtue of anything sufficient at all (such as what seems to be the case with the radioactive decay of atoms). Not only is this a rather odd, complex and arguably more intelectualised view about mental-physical causation but the simpler alternative in terms of sufficient causes seems more adequate to capture our usual thinking. What sounds more like what we generally believe about mental-physical causation is the idea that our mental states are "the real reason" why the physical effects happened; that they happened *in virtue* of the fact that we had those mental states. It also seems that the "command" of the mind is stronger than providing a mere necessary condition[6]; our willing *makes our arms move*, for example.

This point about our conception of mental causes being "the real reason" why some physical effects happen is related to my last comment about EM. An important feature of the way we conceive of the role of our minds in affecting our behaviour and our environment is that the occurrence of the mental states in question are *non-redundant*, that is, they are not merely acting on bringing about something that was already bound to happen for other reasons. This is a very important point for the development of the thesis of this work. It is intimately connected to the "spirit" of NOD (and I will return to this topic in my comments on it) and to the notion of 'mental quausation', central to this work, to which the whole of Part II is devoted.

The causal closure principle: this principle together with NOD are arguably the theses in our list that most generate controversy about their precise formulation. The CC principle has appeared in different versions and under different names (such as "the completeness of the physical", e.g. Papineau, 2000; Crane, 1995; for a comprehensive list of many formulations, see Gibb, 2015, p. 628)[7]. A tempting first attempt at offering a formulation of the principle

[5] Matters are not so simple once we consider INUS conditions (Mackie, 1965), which I consider highly interesting, but that will not be discussed in this work.

[6] Not only that, but if the mental contribution were in any way akin to a fixing of necessary conditions for radioactive decay (say), less regularity should be expected between the mental cause and the physical effect.

[7] The reason why I preferred the 'causal closure' label is due to a fortuitous ambiguity of what 'closure' invokes. In a first sense, it has the technical meaning of a closure property of a set, such as in "being closed under an operation". In this case, the set of all physical events (or effects) would be closed under a causal

would be this: "every physical event has a sufficient physical cause". An immediate problem with this phrasing is that it rules out the possibility that there are physical events that do not have a sufficient cause (or a cause at all). This is a possible interpretation of phenomena such as the oft-cited radioactive decay of atoms. The next move would plausibly be to restrict the principle to those events that *do* have a cause. This is basically the move adopted in the formulation I favoured: it is phrased in terms of physical *effects*, which, by definition, require a cause.[8] There is another reason why I decided to formulate the theses involved in the exclusion problem in terms of "causes" and "effects". As I mentioned before, there are many candidates for the role of causal *relata* and I wished to remain as neutral as possible about this issue at this point. Using the generic terms "causes" and "effects" allows me to do just that.

There are further worries, however: Sophie Gibb (following Lowe, 2000) chooses the formulation "At every time at which a physical event has a cause it has a sufficient physical cause" (Gibb, 2013, p. 2). The explicit mention of the times is an attempt to circumvent a possible loophole in the principle that allows physical effects to have non-physical causes due to the transitivity of causation. Consider this case: physical event P1 causes non-physical event N, which in turn causes physical event P2. Since causation is transitive (arguably), P2 can be said to have P1 as a physical cause. But cases like this violate the spirit of the CC principle. Since I am not sure Gibb's proposed version succeeds in excluding these cases (would it not still be true that P2 has a physical cause in P1 at the time it occurs?), we could simply decide to explicitly exclude this sense of "having a cause" by formulating CC in terms of "immediate" (also known as "proximate") cause, that is, a *direct* cause, with no intermediaries. I decided not to do this in my first presentation of the principle again to keep matters simple, but I would see no problem in clarifying that this is what is actually meant, with the lack of mention to immediate cause being a mere case of contextual ellipsis. Personally, I prefer to phrase the CC principle as follows: for every physical effect, there is a true causal explanation of its occurrence in terms of physical causes only. This will find application in the main argument I develop in Part III.

operation of sorts: if an event is in this set, then its cause is also in the set. A second meaning evoked by the term is of the physical world as a "closed system", in which causal transactions would suffer no "external interference".

[8] This is roughly the same strategy employed by Sophie Gibb in Gibb, 2015. In any case, such possible cases of uncaused physical events should not matter to the causal exclusion discussion, since *effects* (in this case, physical) of mental causes are what is at stake.

I mentioned in the last paragraph the "spirit" of the CC principle – but what would that be? Irrespective of the details of formulation, the general idea behind the principle (which is what ultimately matters and what we should keep in mind) is that, for every physical effect, there is always a physical cause available to explain its occurrence. Another way of trying to capture the idea: no appeal to a non-physical cause would ever be required to explain the occurrence of any physical effect[9]. The idea, then, is that there will always be a true causal explanation for the occurrence of any physical effect cashed out only in physical terms. This is, roughly, the core idea behind the CC principle. But in the case of interest to this work, mental causation, what would the truth of CC amount to? What does it imply about physical effects we usually attribute to mental causes? The standard response is to say that it implies that there is a purely physical causal explanation for their occurrence that, at some point, involves the physics behind the workings of our nervous system. So the physical explanation of, say, "John saying something" would be the causal steps, presumably originated in whatever it is that is the neural correlate of John forming his intention to say something, leading to the release of air from his mouth in a particular way. Since the workings of our nervous system are supposed (plausibly) to be ultimately physical, we would end up with a purely physical causal explanation (even though we might not yet be able to fully describe it given our knowledge). There is a crucial thing to notice at this point: the CC principle, on its own, says nothing about there being or not a nonphysical cause for any physical effects – it only states there is a physical cause available. It does not exclude the possibility of an *additional* nonphysical cause. It is only when coupled with the NOD principle that CC excludes this possibility.

Before addressing the reasons we might have for believing the CC principle, there is a last important clarification to be made. It is common in the literature on mental causation to talk about *sufficient* causes, and the CC principle is usually formulated in terms of them. Since this is common practice, I decided to follow the trend for the same reasons I have evoked before for my terminological decisions. I believe no great changes would follow from removing any mention of sufficient causes altogether, though. To avoid misunderstanding, however, it is important to explain what one means by 'sufficient'. If we interpret "sufficient causes" as causes that *necessitate* or *guarantee* the occurrence of their effects, then one could

[9] The two phrasings I offered do not say the same thing. The first says that there is a physical cause available, while the second only says that a non-physical one is never required. It does not explicitly state that there *is* a physical cause available. Even though it is not as precise, I thought the second was worth mentioning since it captures part of the motivation for believing the CC principle, which is the inductive conclusion that, since no appeal to such causes has ever been required in physics, it will never be required.

immediately object that this formulation presupposes a deterministic view of causation, which is an additional and controversial assumption we should avoid if possible. No commitment to determinism follows from mere talk of sufficient causes, however. The key to see this is that a somewhat weaker notion of sufficiency might be the one in play. In this case, a sufficient cause would be compatible with indeterminism if we understand that it is sufficient merely to *fix the probability* of the effect happening (Hitchcock, 2012, pp. 43-44). Alternatively, we could appeal to the even weaker consideration that, once the physical cause happens, additional nonphysical factors make no difference whether the effect occurs or not (*ibid.*, p. 44)[10]. This weaker notion of "sufficient" is what I intended to use in the formulation.

Why should we believe the CC principle? Philosophers frequently assume that this principle is something of a scientific platitude, something we learn from physics. The idea is that it would be a general principle like "all elements have a fusion point", which is assumed and corroborated by scientific practice. But is that really the case? When we examine the principle, it does not seem something we could even expect physics to address directly. Physicists do not usually try to define what the "physical domain" is in their work, let alone directly ponder about possible causes for phenomena lying outside of it. The principle is clearly philosophical. At best, it is some kind of principle scientists are expected to adhere to and to which they could agree if asked. The question should be reframed thus: can physical science provide us with good or even decisive reasons for believing the CC principle? There are at least two main avenues for supposedly arriving at a positive answer for this question. The first (and perhaps the most common or intuitive) is an inductive appeal to the great success physics has had in explaining and predicting phenomena (see, e.g., Gibb, 2013, p. 3). Since nowhere in its history as an established scientific discipline has physics found the need to appeal to any kind of phenomena we could regard as nonphysical to explain anything, and since we can reasonably suppose that our bodies are ultimately physical, we could safely assume that a recourse to nonphysical causes will never be required to explain physical phenomena (including our behaviour). The fact that physics has been so successful in relying on very comprehensive laws to explain phenomena seems to leave little to no space for any kind of influence outside of its domain. This in turn leads us to the second main strategy for defending the scientific support to CC I identified: the appeal to laws of conservation (most notably, of energy and momentum). Now the idea is slightly different. If we suppose that any causal influence exerted by a nonphysical entity would violate in some way the

[10] These weaker interpretations of 'sufficient' are not without their discontents, however; see, for instance, Montero, 2003.

conservation of physical quantities (see, e.g, Dowe, 2000), we could appeal to the fact that these conservation laws have never seen an exception and are central to our physical theories to conclude no such intervention could actually occur. Alternatively, we could conceive of causal transactions in physics as always involving some kind of transference of physical quantities. In this case, the mere possibility of a nonphysical cause would be barred as a matter of principle, since presumably nonphysical entities could not enter into this sort of exchange. This line of argument has been subject to criticism, however, as it is at least not obvious that nonphysical causal influences would require violations of conservation laws (see Gibb, 2015; Montero, 2006, and Pessoa Jr & Melo, 2015). In any case, it is worth noting that these responses to the challenge of finding scientific support for the CC principle (the latter especially) appeal to this picture of the physical world as a closed system.

Nonreductionism: I have used "Nonreductionism" as an abbreviate term. The expanded, and more correct, name to this thesis would be something like "nonreductionism about mental causes". This is arguably the most controversial of all the theses involved in the exclusion problem. Many philosophers invested in the topic of mental causation accept it, however; especially those who adhere to some form of nonreductive physicalism. The NR thesis in question here should not be confused with nonreductive physicalism, though; they are different theses. In general terms, nonreductive physicalism can be characterised as the view that mental items (states, events, properties, predicates, etc.) are not reducible to physical items, but that the former depend on the latter in some way (through notions such as supervenience or realisation). This preserves the "physicalist spirit" of the primacy of the physical through the asymmetrical relation involved: the physical fixes or determines the mental, not the other way around. NR, on the other hand, merely states that mental causes (whatever they are) are not reducible to physical causes. It is clear that these statements say different things, nonreductive physicalism involving further commitments. As an example of their independence, we can note that one can be a nonreductive physicalist while rejecting NR. The most famous example (though likely not the best due to the subtleties involved) would be Donald Davidson's "anomalous monism" (Davidson, 1970; 1993). Though it can be considered a form of nonreductive physicalism in some way, since it maintains that mental and physical predicates are not reducible to one another, this theory posits that mental causes "just are" physical causes (since causation is supposed to relate events and, on this view, mental events are identical to physical events).

As is the case with some of the other theses involved in the exclusion problem, the formulation of NR I chose is extremely general. This is intentional. The idea is to avoid as

much as possible any kind of commitment concerning the details about causation. Speaking in terms of mental and physical "causes" allows us to remain neutral not only about the exact nature of the causal *relata*, but of the causal *relation* as well. An important consequence is that it can be interpreted according to either "*de re*" or "*de dicto*" views on causation. By "*de re*" views on causation I mean the broad family of theories that share the standard causal realist presupposition that causation relates entities *themselves*, that exist "out there in the world". Theories of causation in terms of objects, events, or properties are examples of what I mean by "*de re* views". "*De dicto*" views, on the other hand, are those theories according to which causation is a relation not between "the things out there", but between linguistic items (or representational items, more generally) we use in making discourses about the world. So theories of causation in terms of event descriptions or sentences would be examples of what I mean by "*de dicto*". The "choice" between *de dicto* and *de re* views is as fundamental as the one between causal realism and antirealism, since it deeply affects the way we understand causal matters. It is noteworthy, then, that a formulation of NR can be meaningful to either side of this debate.

What *kind* of reductionism does NR involve, however? The general idea behind all forms of reduction is this: to say that A is reducible to B is to say that A "just is" or "is nothing over and above" B. So NR can be read as the denial that mental causes "just are" or "are nothing over and above" physical causes. That, on its own, is not very informative, however. Adding a little more detail, there are roughly two main types of reductionism in philosophy: *ontological* reductionism, which is about reducing entities, and *representational* reductionism, which is about reducing linguistic or conceptual items, such as theories or concepts. These in turn have further subtypes, such as elimination and identity for ontological reductionism, but we may leave this aside for the moment (for a comprehensive taxonomy, see van Gulick, 2001. I followed his terminology for the "representational" branch). Which of these, if either, is presupposed in NR? Once again, the answer is that, as it was phrased, NR is neutral between these types of reductionism, but there is an interesting point we can now appreciate. My brief description of the ontological and representational varieties of reductionism might have seemed highly reminiscent of my characterisation of the *de re* and *de dicto* views, respectively. I believe this is no coincidence and that there is some kind of "affinity" between these pairs. Given the way they are usually formulated, I believe there is a tendency for those who have *de re* views to consider causal reductionism in ontological terms, whereas those on the *de dicto* side would consider it in representational

terms[11]. If I am correct, this establishes connections between these broad divisions that are not usually considered. The most neutral possible reading of NR would probably be enough for most of our purposes. However, there will be times where I will talk about my specific inclinations on the matter. Since I have already mentioned that I will follow some form of causal realism when this commitment matters, I must now also say that, since I believe causal realism is best understood in *de re* terms, I will ultimately tend towards an ontological reading of NR whenever this distinction matters. This will be important in my take on the notion of "quausation" in Part II, leading to my reformulation of the exclusion argument in Part III.

When it comes to the *reasons* people have for believing that NR *in specific* is true (that is, not reasons for believing in causal nonreductionism in general, but that *mental* causes are irreducible to *physical* ones), we find that they are usually the same reasons for believing in mental irreducibility in general (of which nonreductive physicalism is an especially famous case). The general idea is that there is something about mental states (especially conscious, qualitative states) that makes them peculiar and significantly different from physical entities. This, in turn, would motivate the idea that the kind of causal influence that mental entities exert is different from the causal influence available to physical entities. The usual considerations brought up to motivate this general nonreductive claim are epistemic (the most influential being the knowledge argument, tracing back to Jackson, 1982) or modal (such as the conceivability argument; Chalmers, 1996) and the inverted spectrum hypothesis (dating back to Locke, but more recently brought to the fore by many philosophers, such as Block, 1990, Shoemaker, 1982, and Kim, 2005). Notice that, curiously, these more widely discussed considerations are not directly ontological; that is, a difference in the nature of mental and physical items is argued for *indirectly*. A possible exception is the "multiple realisability" hypothesis (for "classics", see Putnam, 1967 and Fodor, 1975; a "recent classic" is Polger & Shapiro, 2016), which is not as easy to categorise – it surely involves modal considerations, but it involves ontological or conceptual aspects as well.

A last thing to note is that, in general, the more "fine-grained"[12] the theory of causation favoured, the more likely it will be that differences between mental and physical

[11] To say that there is "a tendency" is understating for a "safe bet" on my part. I would actually go as far as to say that these associations (*de re* – ontological and *de dicto* – representational) are the only meaningful parings.
[12] By 'fine/coarse-grained' I mean the extent to which a given theory of causation appeal to subtleties (usually, ontological minutia). The more fine-grained the theory, the more these fine distinctions play a role. So, for example, if one recognizes that events are complex entities (i.e., they have structure, constituents) and maintains that causation occurs at the level of events, then one is favouring a coarse-grained theory (since the constituents are not called in to explain the causal transactions). For an interesting classification of famous theories of causation in the "coarse/fine-grained spectrum", see Schaffer, 2016, sec. 1.

items become causally relevant. It is easy to see why: in "coarse-grained" theories of causation, it is possible to maintain that mental causes reduce to physical causes despite the differences there may be between the mental and the physical. After all, these differences might reside in specific aspects that do not figure in what is causally relevant. However, once the specific details become important, as is the case with more "fine-grained" views, there will be less space for the possible differences to hide, so to speak. So, for example, if one defends a theory of causation in terms of events (a "coarse-grained" view), it is possible to maintain that mental events are the same occurrence as some physical events, even though mental and physical items might differ in some specific aspects. But if one favours a theory in terms of property instances (a very "fine-grained" position), unless they defend a more radical form of reductionism in which mental property instances just are physical property instances, property differences will likely result in causal differences. The case for a notion of "quausation" explored in Part II will favour a "fine-grained" view of causation, so that the peculiarly *mental* aspects become indispensable in an account of mental causation.

No Overdetermination: This is probably the thesis that most requires "unpacking", being the most technical of them. The gist of it is the idea that there can be no more than one sufficient cause for an effect at the same time[13]. If there were, the effect would be overdetermined – that is, it would have more than enough factors determining that it should occur. It is worth mentioning that the kind of picture involved here is not what happens in cases such as, for example, when two people lift a sofa, each pulling up from a side. This would not be a case of causal overdeterminaiton (and thus not a counterexample to NOD), since each person lifting their side is not sufficient to bring the sofa to rise. Instead, what they do individually is what is usually called in philosophy a *partial cause* for the movement – an occurrence that is only a part of the cause of an effect[14]. The NOD principle is not about partial causes, but about sufficient "whole" causes instead (if we call the opposite of partial causes that).

But what exactly does the *systematic* in 'systematic overdetermination' mean? This detail is important and traces back to a distinction made by Jaegwon Kim between genuine and systematic overdermination (for instance, in Kim, 1998, p.65). For Kim, the following case would not be a counterexample to NOD: confused about the set piece choreography

[13] For physical effects at least, but this should not be of worry since this is exactly the kind of effect in question.

[14] In the case of the sofa, the actual cause would presumably be best understood as their combined movements.

they have rehearsed, two footballers take a free kick at the same time. The ball flies into the "top bin"(the top corner of the goal) and a goal the keeper could not have even dreamt of saving is scored. Adding fortuitous coincidences to the story, each player struck the ball individually at such speed and with such good technique that the goal would have been scored in a similar manner regardless of the aid of their teammate.[15] If it is not already clear, in this example both individual actions would be sufficient to bring about the effect and they happened at the same time. Kim calls cases like these examples of "genuine overdetermination". These are cases in which causal overdetermination truly occurs[16]. But, according to Kim, they are rare, coincidental occurrences. By contrast, *systematic* overdetermination would involve cases in which such harmony of sufficient causes occurs frequently, in a systematic manner. This is the kind of overdetermination NOD forbids.

Kim does not go into detail about what characterises systematic overdetermination; he seems to be only interested in describing the *genuine* cases and saying that the systematic ones are nothing like that. In his take on the exclusion problem, Cristopher Hitchcock offers a way of interpreting what systematic overdetermination amounts to (Hitchcock, 2012, p. 45). His proposed reading is centred on frequency: genuine cases are the rare occurrences, whereas systematic ones would be the ones that happen often, such as mental causation. He comments that the vagueness of the "common/rare" distinction is a potential problem, stating that the no-overdetermination principle only requires that overdetermination be not as common as mental causation is (*ibid.*). While I do believe Hitchcock's proposed criterion is satisfactory, I believe that speaking in terms of *types* of phenomena (like I did in the clarification of my formulation) helps to mark the distinction in a somewhat straightforward and concise way. For instance, occasional overdetermination (I do not think 'genuine' is a

[15] Kim's rather gruesome preferred example is about two shooters standing in opposite ends of a line that shoot and kill a person sitting at the midpoint at the same time (the bullets meet halfway in the victim's head) (Kim, 1998).

[16] Are they, though? Recall that in the example I gave I said that the goal would have been scored "in a similar manner" regardless. This is different from saying that the *same* effect (or even a "perfectly similar" one) would ensue. And it surely seems that, in strictly physical terms, the resulting effects in these alternative scenarios (single shooter vs two) could not be "perfectly similar": even if the ball's trajectory ended up being "the same" (something rather miraculous and hard to conceive), the causal transactions would occur in different ways in each scenario (for example, the scenarios would differ regarding the position and movement of the feet when touching the ball). This suggests a way of maintaining that even *occasional* overdetermination never occurs: if we give more fine-grained descriptions of the events, all appearance of overdetermination would be dissolved (since the effects would turn out different, so there could not be multiple causes for *the same* effect). This in its turn hinges on the issue of how dependent on descriptions our causal discourse is. In our example, if we describe the effect as "a goal to be scored" or even "a goal to be scored on the top bin", it seems that there is a margin to speak of overdetermination (since the individual shooter would be sufficient for the same effect). However, if we give a more fine-grained description of the resulting goal, the effects would turn out different in each scenario.

good name for the opposite of 'systematic') would consist, on this reading, of particular (or isolate) cases of causal overdetermination. We might concede that these eventually happen, as is at least *prima facie* plausible that they do. The real problem that the NOD principle blocks is having every instance of a kind of causal phenomenon to happen overdeterminately; that is, every time that mental causation (for example) happens, it happens overdeterminately. This is what seems to "offend nature" (I refer here especially to the quote by Newton Hitchcock reproduces), since it postulates a *redundant* kind of phenomenon[17].

But why should we believe in the NOD? Kim goes so far as to say that denying it is absurd (Kim, 1993, p. 180), and that one of his versions of the principle is an analytic, a priori truth (Kim, 2005)[18]. We do not have to believe that the NOD principle satisfies such a high standard in order to believe it, however. The main reason for believing it is true is what I take to be the "animating spirit" behind the principle: the aforementioned denial of redundancy. To believe in the falsity of NOD amounts to believing that some types of phenomena always occur in and overdeterminate manner – be it mental causation, rain or chemical bonds. But that means that some of the multiple sufficient causes involved (it makes no sense to specify which) are exerting no new causal influence in the process, since the other cause (or causes) is already sufficient on its own to guarantee the effect will happen. This would imply that nature works in such a way that certain types of phenomena always involve redundancy. This seems to violate some kind of principle of parsimony that supposedly rules the way things are; it would imply that nature systematically requires unnecessary additional steps.

Even if systematic overdetermination of some kind turns out to be possible, there would still be reason to resist the idea that this is what mental causation amounts to. As I said earlier, it seems that the idea of mental-physical causation we have and would like to

[17] One could attempt to trivialise my distinction by pointing out that we can generate a kind of event for every specific example of occasional overdetermination (albeit a rather artificial or very specific kind). For instance, we could say that "a concurrent shooting on the head by two shooters" and "a breaking of a window by two rocks acting simultaneously" form two kinds of events (generated from the usual examples of occasional OD). But if we assume that instances of those types occur (which presumably we do) and if these types are by definition (by my "definition", that is) systematically overdeterminate, then we should believe that there are cases of systematic overdetermation (contra NOD). The core idea behind this objection is that it is not obvious how we can individuate kinds of events, so that distinguishing between "natural" (which could not be overdetermined) and "artificial" (which could) kinds of events is not a task we should obviously be able to do. I have yet to think of a decisive response to this objection, but the fact that "mental causation" (and any other kind of phenomenon that I can imagine that is seriously at stake with exclusionary worries) is a kind of phenomenon that is importantly different from any of the artificial and specific kinds the objection is able to point to gives me a reason not to worry.

[18] The principle in question is Kim's "exclusion principle": if a state S1 is causally sufficient for a state S2, then no distinct state obtaining at the same time as S1 can cause S2.

preserve is one in which the mental states are "the real reason why" the physical effect in question happened. This seems incompatible with avoiding the exclusion problem by assuming that mental-physical causation is a case of systematic causal overdetermination, since, in this case, the mental influence would be redundant (which would amount to no influence at all). The idea of mental causation we seem to have is one in which the mental *makes a difference*, and it is not possible to make a difference and be causally redundant at the same time. We could presumably apply this same reasoning to other types of causal phenomena and conclude that, generally, causation involves this kind of difference-making[19] that systematic overdetermination precludes of. This would be another avenue for believing NOD in its general formulation.

When phrased so as to put the pressure on EM, the main peril becomes rejecting this thesis, arriving at the unpalatable view that there are no mental causes of physical effects. In other (more technical) words, it highlights the threat of *epiphenomenalism*. To be an epiphenomenalist about something is to believe that the thing in question is real, in some way dependent on something else, and unable to cause anything[20]. In the most usual case, 'epiphenomenalism' is used to mean 'mental epiphenomenalism', that is, epiphenomenalism about the mind (or, more specifically, about consciousness). So mental epiphenomenalism amounts to the thesis that the mind (or consciousness) is real, dependent in some way on a physical substrate, but unable to cause anything. On this view, the mind would be like a causally inert by-product of the workings of its physical basis, a "mere epiphenomenon". When I speak of "epiphenomenalism", I will have "mental epiphenomenalism" (or, being more precise, something like "consciousness epiphenomenalism") in mind.

I would like to close this section by stressing the seriousness of the exclusion problem. As Jaegwon Kim notes (Kim, 2005, pp. 9-10), it manages to combine worries raised by two other "big" philosophical problems: those of scepticism and of free will. The sceptical worries arise from the common assumption that knowledge requires causal contact. So if our mental states are epiphenomenal, it seems it would be impossible for us to have knowledge about them. This means that innocent knowledge claims, such as that we can know that we are in pain or that we have visual experiences, would be barred since our sensations could not come in causal contact with anything (so, for example, my true belief that I am in pain

[19] 'Difference-making' is a term used to name the core tennet of a strand of causal theories, espoused by philosophers such as Peter Menzies (most notably) and Jonathan List, which I address in Part III. I am not using the term in this specific technical sense in the above passage of the main text. I am only alluding to the general idea of making a difference.

[20] One could, for instance, be an epiphenomenalist about shadows or composite objects.

could not be caused in any way by the pain I feel)[21]. This connects to another highly problematic result of epiphenomenalism, which is linguistic: suppose I feel pain on my left arm and then say "I feel pain on my left arm". If epiphenomenalism is true, my utterance could not have been caused by the pain I felt. This is already very counterintuitive and problematic, but it does not stop there. It seems that our words could not refer to the epiphenomenal mental items they supposedly represent (especially if we adhere to the popular causal theory of reference). For if there is no connection from our mental states and what we say about them, how could the terms "pick out" the corresponding states? These are all symptoms of a more general (and I dare say "darker") consequence: if epiphenomenalism is true, then it seems that no communication about our mental states could be made at all. Without a "causal bridge", it seems that our "inner lives" would have to remain forever incommunicable, since there could be no transmission of information coming from our mental states. There would be no connection from linguistic content to mental content, and nothing we say could possibly "trace back" to our mental states.

The connection with the problem of free will can be seen when we pay attention to the role we attribute to our mental states in our decisions. When it comes to free will, the direct impact of epiphenomenalism lies less on the "free" and more on the "will"; it is a threat to agency. If our thoughts, desires or feelings are causally inert, then every action we might attribute to them is actually caused by something else (usually, the neural correlates of these states). The picture of ourselves sitting in control, interfering in the world through our minds, becomes illusory. Our minds are rendered a mere result rather than a part of the process; the parts that actually do the work, physical in nature, become impervious to any kind of interference. This alienates us from our actions, a result that reverberates across other topics, such as responsibility (how could we become accountable for our actions if epiphenomenalism is the case? We could not truly say, for instance, that someone "acted on bad intentions" or "on a fit of rage" if intentions and feelings are epiphenomenal). This result can be attenuated if we are epiphenomenalists about only *some* mental states; for instance, we

[21] In the cited passage, Kim focuses on the case of perception. It is interesting to note that *perceptual* knowledge can be affected by the kind of scepticism motivated by epiphenomenalism. This might go overlooked if we think that the "direction of fit" of perceptual knowledge is not the one at stake here. That is, since it is a problem about mental causation, the impression is that only those cases with mental causes involved are under consideration, and most of the problematic examples often cited fall under this type (they are about reports or introspection). Since perception is a case of influence "from world to mind", we could easily conclude that it is unaffected by the problems I discussed. But matters are not that simple, since perceptual knowledge requires "processing" and (arguably) a relation to beliefs. The general idea is that perception might not be a mere case of "one-way" influence; it arguably requires *interactions*. But epiphenomenalism bars all possible interactions with the epiphenomenal states. It leaves no place for any kind of mental "feedback".

may see reason to think that conscious states are epiphenomenal, but that thoughts, beliefs and intentions are not (we could see reasons for believing the first states, but not the others, are irreducible, for example)[22]. In this case, we could maintain the efficacy of many mental causation claims: we could still say that John went to the kitchen because he believed there was water to be found there, for instance. But anything conceded to epiphenomenalism might be too much, since we could still find many cases where we attribute causal powers to these excluded states. In the example of phenomenal states as the only ones excluded, this would still imply that every mental cause that would involve a phenomenal mental state is in danger (which becomes especially problematic if we defend, as some philosophers do, that states such as memories and beliefs have a qualitative component to them). Think of the kind of decision featured in cases such as, again, Thirst and Bitter drink. What element of agency would be left to these decisions if the sensations involved were epiphenomenal?[23]

Jerry Fodor neatly captures this dreadful character of epiphenomenalism in this oft-cited passage:

> [I]f it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying […] if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world." (Fodor, 1989, p. 79)

Epiphenomenalism, like the truth of scepticism or the lack of free will, is an "end of the world" result. Theses like these are the end of the world in a relevant sense because they imply that we are deeply mistaken about very fundamental (or basic) aspects of our worldview. In the case of epiphenomenalism, its truth means that we are mistaken when we say that the itch was the reason for the scratch, that the fragrance brought a memory, that we ate because we felt hungry, that we can describe what we feel, and perhaps even that we act on our will. In essence, it means that we err every time we attribute a causal role to any of the epiphenomenal mental states in question (whichever they may be).

Since epiphenomenalism seems so absurd, it is natural to think that, if a thesis entails it, then so much the worse for this thesis. This is precisely the strategy behind a famous kind of argument that builds upon the exclusion problem: the exclusion *argument*, which attempts to show that nonreductionism (generally, in the form of "nonreductive physicalism") leads

[22] This is roughly the picture motivated by Kim, 2005.
[23] For a broad discussion about epiphenomenalism and free will, see Merricks, 2001 (especially chapters 3 and 6).

to epiphenomenalism. In so doing, it is usually interpreted as a *reduction ad absurdum* of nonreductionism.

3. The Exclusion Argument

As we have seen, we can arrange the premises of the exclusion problem in many ways. In so doing, we can switch the focus of the tension behind the problem. The usual formulation of the problem puts the tension on NR. There is a strand of argument that builds upon this emphasis to argue that nonreductionism leads to epiphenomenalism. This is what is usually known as the "exclusion argument" (though we could speak of a *family* of exclusion arguments; see Hitchcock, 2012), made famous by Jaegwon Kim[24].This kind of argument has appeared in many versions and under different names[25]. I will mainly address the specific version I favour and reproduce next. This version I now reproduce is based on Jaegwon Kim's expositions on the matter, though it differs slightly from the specific formulations he has given (Cf. Kim, 1998; 2005 and 2006):

We assume CC and NOD as hypothesis (I will not list them, but will refer to them). EM is in dispute. We then start by assuming NR as an hypothesis:

1. Mental causes are irreducible to physical causes. (Hyp.)

Assume further that a given mental event M of the kind non-reductionists say is irreducible (e.g, a sensation) causes a physical event P.

2. M causes P. (Hyp.)

3. By the causal closure (CC), P has a sufficient physical cause – let us call it P*.

4. If M causes P, then either M is reducible to P* or P is overdetermined.

5. M is not reducible to P* (by NR).

6. So, if M causes P, then P is overdetermined (2,4,5). But this contradicts NOD, so

7. M does not cause P (2–6, *Reductio ad absurdum*). This result generalizes for every M and P. Therefore,

C. If NR is true, then (irreducible) mental events do not cause physical events (that is, EM is false). (1–7, Conditional Proof).

---

[24] Though most of the attention brought to the exclusion problem/argument came from Kim's work, he was not the first to mention this sort of problem. The exclusion argument traces back to at least Malcom, 1968.
[25] Sophie Gibb, for example, calls a version of the exclusion argument (reproduced from Lowe; see Gibb, 2013, note 2) "the argument from causal overdetermination" and arguments of the same family "causal closure arguments" (Gibb, 2013, p. 2).

In summary, the exclusion argument aims to conclude that if NR is true, then EM must be false (which amounts to epiphenomenalism being true). A first thing to notice is that I formulated the argument in terms of NR, while the argument is usually formulated in terms of "nonreductive physicalism". As I have remarked earlier, these theses are not the same, and, depending on the definitions adopted, they might be independent (that is, one can accept nonreductive physicalism without accepting NR and vice-versa). It is clear, though, that the thesis that is immediately relevant to this argument is something like NR, which is an eminently *causal* thesis. We could attempt to bridge the two theses, however, by defending that some sort of nonreductive physicalism (either in an ontological or "representational" variant) implies NR, usually by indicating that the mental entities (or items, in general) that "do the causal work" are rendered irreducible by this strand of physicalism. For example, someone may defend that the "causal work" is done by properties, and then formulate nonreductive physicalism in terms of properties to conclude that it leads to the truth of NR[26].

A second thing to notice is that, as I made clear in my discussion of the tension involved in the exclusion problem, the tension here lies on NR. The argument is, after all, normally put forth as an attempted *reductio* against NR. By taking this formulation as the starting point of my exploration of the exclusion argument, I will preserve this manner of presenting it, with the tension lying primarily on NR. In Part III, however, the points I will raise and defend will inevitably lead to a switch of focus, so that EM (or a refined version of it, as will become clear by the end of Part II) will be at stake. Taking the exclusion argument seriously, I will defend (and this is nothing new), unveils problems about the very possibility of mental causation.

4. Expanding exclusion: the mental-mental case

The exclusion argument I reproduced clearly targets what I have been calling mental-physical causation. It seems essentially tailored to address this variety of mental causation, since it relies on the CC principle to raise the problem for NR. That is, the tension involves the possibility of mental-physical causation in face of CC and NOD. This simply follows the usual way of presenting the exclusion problem. This might lead one to think that cases of mental-mental causation are left untouched by exclusionary considerations, especially under a non-reductionist (or full-fledged non-physicalist) view about the relevant mental states.

[26] This is, for example, the strategy behind Hitchcock's version of the argument in his 2012.

After all, causal interactions between mental items, given the assumption that these mental items are not reducible to physical items of the sort CC is about, will seem to lie outside the purview of the causal closure of the physical domain. This would be a hasty conclusion to derive, however, as there are natural ways to generate a similar argument in the same "spirit" that would also affect mental-mental causation.

Here is one way to extend exclusion to the mental-mental case[27]. Suppose, for starters, that the occurrence of mental item M causes the occurrence another mental item, M*. Now, for the next step, there is a further assumption we will have to make that was not required in the mental-physical case, but it is a rather innocent one, and one that nonreductionists usually make. This assumption is the fairly minimal claim that every time a mental item occurs, it occurs together with some physical item, and that every time that physical item (or that *type* of physical item, it will not matter here) occurs, that mental item also occurs. That is, the occurrence of one is a guarantee, a sufficient condition for the occurrence of the other. This is a claim that most dualists probably accept, and that nonreductive physicalists normally maintain. It follows from this that M* will have a corresponding physical item, call it P*, such that it occurs when M* occurs. Not only that: the occurrence of P* is sufficient for the occurrence of M*[28]. Now, given that P* occurring is sufficient to account for the occurrence of M*, how can M be the cause of M*? A natural assumption to make, and the only sensible one according to Kim (2006), is that M causes M* by causing P*. But, given the CC principle, P* already has a sufficient physical cause – call it P. And now we run into the traditional problem: to suppose that M causes P* given that P is already causally sufficient for causing it would make P* overdetermined, which NOD would not allow. Hence our initial assumption that M causes M* cannot come to be.

I will follow the standard approach in the literature and focus on cases of mental-physical causation in the remainder of this work. I believe it is fruitful, however, to notice that the exclusion problem (and argument) has ripples that travel through these classifications, the mental-mental cases being also in peril.

---

[27] An argument in these lines can be found in Kim, 2006.

[28] Whether the relation between mental items and their underlying physical items in cases like these is that of causation (that is, whether it can truly be said that P* *causes* M*) is subject to debate. Kim (e.g., 1992, 2006) defended that this relation should not be understood in causal terms, citing as a reason the assumption (disputable and potentially problematic) that mental states occur simultaneously with their underlying physical state (that is, there is no delay between the occurrence of one and the other), and that there is no simultaneous causation (that is, causes must occur before their effects). I believe the argument in the main text does not depend on an answer to this question, as I hope my exposition makes clear.

A central task of this work is to develop a new version of the causal exclusion argument. It involves a form of causal exclusion principle I have not encountered yet in the literature, and concludes that only a very restricted form of reductionism could evade the considerations laid out by the argument. In order to formulate this argument, I must first go through what will be a major component behind it: the general notion of mental "quausation". This is what I devote my attention to in Part II.

Part II

Mental Quausation

In Part I, I outlined some key questions that any account of the nature of causation must answer. In Part II, I will examine the notion of "mental quausation", a term (if not the idea) created by Terrence Horgan (1989). This notion will be relevant in addressing the question "In virtue of *what* do causes cause their effects?", which can arguably also be phrased as "what constituents of causes "do the causal work"?", when applied to the case of mental causes. The discussion will also inevitably involve commitments about the answer to the question "what kinds of entities feature as terms in the causal relation?". I will begin by introducing the notion, explaining what it amounts to according to Horgan's original exposition. I will try to defend the relevance of the general idea behind this notion for the mental causation debate. In addition, I will offer my own interpretation of what mental quausation amounts to and what are some key implications of a commitment to this idea. A central result from this discussion will be a revised version of the Efficacy of the Mental thesis (EM), which I call Quausal Efficacy (QE). I will defend that this is the thesis we should preserve if we aim to preserve our intuitive, pre-theoretic stance on the matter.

1. "Mental Quausation" as a term of art: what it means and some background

If asked "what is required in order for EM to be true?", many would probably consider the answer to be obvious: that there be mental causes of physical effects, that is, that we can truly identify as the cause of some physical effects (usually, of our behaviour) some event with a mental component. Put in another way, that there are some mental events that cause some physical effects. This point might seem so straightforward as to seem uninteresting. The problem, however, is that some philosophers have argued that matters are not so simple. One such philosopher is Terrence Horgan, who in his Horgan, 1989 defended that there are some relevant subtleties when we try to capture what mental causes are expected to involve.

Before I explain Horgan's thesis, it is worthwhile to understand the context in which his seminal article on the notion of mental quausation was written. It is situated amongst the larger class of debates that occurred in the 1970s and 1980s concerning the implications of Donald Davidson's views on mental causation, following the influence of Davidson's thesis of "Anomalous monism" developed in his paper "Mental Events" (Davidson, 1970). More specifically, Horgan's paper addresses common worry at the time that Davidson's picture of mental causation would lead to a form of epiphenomenalism. I will not offer a detailed explanation of Davidson's views here[29], but the relevant features to note regarding this specific question is that, on his view, causation is a relation between events and the occurrence of mental states would be the same occurrence as the occurrence of some physical state. They would be one and the same event, only described in two different ways (one with mental predicates, the other with physical predicates). This would show how, in a sense, we could evade the threat posed by the causal exclusion problems: there can be mental causes of physical effects in a way that does not violate the CC principle or posits overdetermination because these mental causes are the same event as what we would call a physical event; they are the same event. Causation being a law-like relation between events, then, guarantees that EM is preserved: mental events can feature as causes of physical effects. However, there is a problem that leads to the alleged collapse to a form of epiphenomenalism. On Davidson's view, mental phenomena are anomalous (hence the name of the view) in that there are no laws governing the occurrence of mental states and that could relate them to the occurrence of physical phenomena (Davidson, 1970/1980, p. 208, 224). Since causation involves, according to Davidson (*ibid.*, p. 223), that the relation between cause and effect be subsumed under a law (the "Principle of the Nomological Character of Causality", *ibid.*) and that only physical phenomena could be governed by laws in this manner (that is, the relevant laws are physical laws, *ibid.*, p. 224), there is place for the suspicion that it is in virtue of the law-like behaviour of *physical* phenomena that the effects of what could describe as mental causation occur[30]. That is, even though mental events just are physical events, the physical component would be doing all the "causal work", leaving no influence to be made by the mental component. Therefore, even though we could say that mental events "in a sense" can be causally efficacious on Davidson's picture, it seems it

[29] A comprehensive summary of the features of Davidson's anomalous monism (as well as auxiliary topics) is given by Yalowitz, 2021.

[30] Ted Honderich (1982) raises this line of objection. It is worth noting that Honderich's considerations anticipate Horgan's in appealing to the relevance of properties in causation. We can say that Honderich was one of the first in the literature, together with LePore & Loewer (1987), to identify this broad notion to which I want to draw attention, which Horgan later labelled "mental quausation".

would not be in virtue of the *mental* features (what makes these *mental* events) that the effects occur. It is this worry that Horgan touches on in his 1989, as will be clearly stated in shortly. This general epiphenomenalism controversy regarding Davidson's view reached its peak in the late 1980s literature in the early 1990s, leading to Davidson's own reply to this worry in his "Thinking Causes" (1993)[31]. For a commented summary of the history of this debate, see Campbell, 2003.

With this contextualisation in place, I will introduce Horgan's general idea in his 1989 paper. As the line of objection to Davidson's picture of mental causation above already indicates, we can start by noticing that events usually involve many properties – colours, shapes, noises, movement occurring in a given angle, momentum, etc. When events constitute causes, clearly not all of the properties are relevant when it comes to making the effects happen. Horgan offers two examples to illustrate this (*ibid*., p. 49): a high pitch sound of meaningful words shattering glass (imagine audio from a speech being blasted from a speaker) and a loud gunshot that kills someone. The examples come from Dretske (1989) and Sosa (1984), respectively. The point with these examples is that neither the meaningfulness of the sound in the first case nor the loudness of the bang in the second are relevant in breaking the glass and killing the person, respectively. Using now my own example (inspired by Honderich, 1982), suppose we put a cantaloupe on an analogical weighing scale and, as a result, the plates of the scale move. In this case, the scale did not move because of the smell or the colour of the cantaloupe. Neither was its roundish shape relevant. The relevant property here is the cantaloupe's mass. It is because of the particular amount of mass that it had (together, of course, with its contact with the scale and the influence of gravity) that the scale moved as it did. What these examples all seem to clearly show is that some properties involved in events that are causes are relevant in producing their effects and some others are not. In many cases, when we want to be precise in attempting to tell a true

[31] Davidson's reply, in essence, is to mark more clearly his disagreement with the way his opponents have framed the question. On his view, it is not in virtue of properties that events cause anything; it is events themselves that cause, without appeal to properties. Furthermore, as I stated in my brief summary in the main text, he does not appeal to mental and physical properties in his view, but to mental and physical *predicates* (a point that Gibb (2006) points to in her defence that Davidson is not an epiphenomenalist). Davidson's views are notably hard to classify (e.g., the debate over whether he is or not a property dualist), but I believe this reply allows us to place his view more clearly on a point further to causal realism in a spectrum of views, making it a position in which causation is a matter of how we *describe* the transactions. We can also say that, in response to the question "what constituents of causes "do the causal work"?", if he considered it even meaningful, would be "no constituents; it is events themselves that cause". In this view, our search for causal factors would stop at the level of events, then.
His reply, of course, has not satisfied some. After all, it still seems reasonable (unless otherwise argued) to ask what factors (properties or not) of a cause were relevant in producing the effect, and to conclude that certain factors not contributing seems a problematic result, leading to a form of epiphenomenalism. For an example of such a reply to Davidson (with which I mostly agree), see Kim, 1993b.

story about a causal transaction, we seek to capture those features of the cause that were relevant in producing the effect, "the real reason" why the effect happened. It was because the sound had *a high pitch* that the glass shattered, and it was because the cantaloupe had a given *mass* that the scale moved. This notion of "qualified causation", where the relevant features of the cause in producing the effect are highlighted is, in essence, the notion of "quausation".

Using Dretske's glass-shattering example, we can say that it was in virtue of the fact that the sound had a high pitch that the glass broke. Said in other words, the event of the emission of sound, *as* an event of a high pitch sound being emitted, caused the event of the glass shattering. Using the Latin term "*qua*" (synonymous with "as") to express the same[32], we would have that the event of the sound *qua* being high-pitched cause the glass shattering, and that it was *not* the event of the gunshot *qua* being loud that caused the killing in Sosa's example. This is the origin, then, of the term "quausation". It is causation with a "*qua*" clause qualifying it, which means that an event causes another *qua* (i.e., in virtue of) having a certain property. "Mental quausation", then, would merely be the application of this broader notion of quausation to the case of *mental* causation. In this case, we would say that mental events cause whatever it is that they do *qua* mental, that is, in virtue of having the peculiarly *mental* properties that they have. In my example of Thirst and Bitter Drink, respectively, it would be *qua* being a feeling of thirst that the cause lead to John walking to the kitchen, and *qua* being a bitter taste that the cause lead to Paula spilling the drink. This seems intuitively correct. In the case of Bitter Drink, it seems that the fact that a *bitter taste was felt* is required to make the effect occur, or is "the real reason why" it occurred. As the above examples illustrate for causal claims in general, there are cases where specifying a particular feature of the cause seems required in order to give a true description of the causal transaction that occurred. The same would hold for cases of mental causation, where specifying the particular mental features of the cause seems required to account for the effect. This explains the motivation behind the charge of epiphenomenalism made against Davidson's view: to give a proper account of mental causation, it will not do to have as a result that there are mental causes "in some sense", with mental items merely being present in the cause. A proper account should maintain that mental events can cause in virtue of the mental features they

[32] A matter of stylistic preference. Horgan defends that constructions in terms of "*qua*" are quite natural and common in specifying relevance in an explanation (Horgan, 1989, p. 50).

involve (what makes them *mental* events), that is, causation of the mental *qua* mental – mental quausation. As Horgan puts it in this key passage:

> [W]e believe not merely that mental events and states *are* causes, but also that they have the effects they do *because* they instantiate the specific mental properties they do. In short, common sense holds not merely that mental events and states are causally efficacious, but also that their mental properties are explanatorily relevant to the causal transactions in which those events figure as causes. Accordingly, any view which denies the latter claim surely qualifies as a kind of epiphenomenalism, a kind I shall call *quausal epiphenomenalism*. (Horgan, 1989, p. 51)

Horgan points out that Davidson's picture of mental causation could not, on its own, preserve this intended result (*ibid*, pp. 48-50, 51-55). I will not reproduce his objections here, but they are similar to the line of objection I presented above, only focused on the phrasing in terms of *reasons*. He offers, in addition, some criteria that any account of the quausal relation should capture. Before outlining these, it is worth mention that Horgan says he is assuming, as background assumptions, that

> (i) that causation is a relation between concrete, spatio-temporally located, events or states; (ii) that every token mental event or state is identical to some token physical event or state; (iii) that every action is caused by a reason of a certain sort, viz., a so-called "primary reason"[…] [33](Horgan, 1989, pp. 47-48)

As will be important for my own interpretation of the notion of mental quausation, Horgan remarks that he doubts any of these assumptions are required to formulate the problem outlined or the notion of mental quausation (*ibid.*). With this important qualification in place, I will now reproduce his four clauses that characterise quausation in broad terms:

"For any two events c and e and any two properties F and G, c *qua* F causes e *qua* G iff:

(i) c causes e;

(ii) c instantiates F;

(iii) e instantiates G; and

[33] I omitted his fourth assumption since it is not relevant to the general topic of mental causation I want to cover. It concerns what he thinks constitutes a primary reason in the explanation of actions (Horgan, 1989, p. 48).

(iv) the fact that c instantiates F is explanatorily relevant to the fact that e occurs and instantiates G." (*ibid.*, p. 51).

There are a few important things to note about this formulation. First, he chooses the standard picture of causation in terms of events, introducing properties as terms of the relation as well. Second, his formulation of the relation of quausation treats it as a four-place relation: e *qua* F causes c *qua* G. Third, he includes a "*qua*" clause for the effects; this means that, under this formulation, the effects are subjected to the same sort of qualification that applies to causes.

There is not much to be said of clauses (i)-(iii) as requirements. Clearly any theory of quausation formulated in these terms must preserve them. As Horgan himself notes (Horgan, 1989, p. 50), it is clause (iv), which he dubs the "quausal relevance requirement" (*ibid.*), that is the most interesting and peculiar. As meaningful as it is, it is also clearly under-specified: for example, what is for something to be "explanatory relevant" in a causal explanation? This way of phrasing the requirement might even suggest a less realist reading of causation, one in which what matters is making explanations that make sense in an appropriate way. This was not, however, Horgan's intended reading, as he makes clear in stating that quausation is a "metaphysical" notion, not an epistemic one:

> *Bona fide* quausal relevance is not merely epistemic, but metaphysical: relevant properties must somehow pertain directly to the causal transaction *itself*, and not merely to our knowledge that it is a causal transaction. (Horgan, 1989, p. 54)

Attempting to capture clause (iv) in a theory of causation that can salvage mental quausation is the task that Horgan devotes section 4 onwards of the paper (pp. 56-64). We can even go as far as saying that offering an account of quausation *just is* offering a proper account of the Quausal Requirement Horgan phrased as clause (iv). Horgan's own proposed way of doing so (which he calls a broad sketch in need of completion, *ibid.*, p. 56) is a rather complicated account of counterfactual dependence which appeals to "pertinently similar worlds" (PSW) (*ibid.*, p. 58). I will not reproduce or explain this account in this work since my goal is merely to work with the *general idea* of quausation. In fact, I believe there are important problems with Horgan's proposal of capturing the quausal requirement, a type of problem that I believe besets any merely modal theory of causation (of which theories centred on the notion of counterfactual dependence are one type). If I am correct, no amendment to a proposal like Horgan's would work, so no "completion" of his project would be forthcoming. I will make these objections in Part III, when I address the theory of causation as "making a difference"

put forth by Menzies & List, which, as we will see, bears a remarkable structural similarity to Horgan's proposal.

This concludes my introduction of the general idea of quausation (and, consequently, of mental quausation). I believe Horgan was right in bringing these considerations to fore, so much so that a central motivation for writing this work is my belief that the notion of mental quausation should receive more attention than it now does in the literature, which I hope to indicate with its application to my own take on the causal exclusion problem. To list some of the key theses that I believe Horgan was right about, I can mention: that true statements about causal transactions often involve qualifying clauses that specify the features of the cause that made the effect occur; that this is salient in mental causation, where for ordinary cases (especially, those involving conscious states) it seems that appealing to the peculiarly *mental* factors of the cause is required to explain its effects; that it is worthwhile to formulate this notion in terms of events and properties, and that this quausal relation is a relation between items in the world, those involved in the real causal transactions happening, not items in the conceptual apparatus we use to describe these phenomena. These points of agreement are, to me, the most important, so I believe the results I argue for in this work are in alignment with the animating idea behind Horgan's work. There are, however, significant points in which I will deviate from Horgan's original proposal in my interpretation of what mental quausation involves. These are what I will turn to in following sections, starting at section 3. Before that, in section 2, I will give a brief defence of the importance of incorporating the general idea of mental quausation in any account of mental causation that seeks to preserve the common-sense, pre-theoretical view of mental efficacy.

2. Mental Quausation rediscovered: the importance of the general idea

I said in the previous section that one of the central aims of this work is to try "put mental quausation in the spotlight" again, as it were, showing that this notion should earn a resurgence in discussions about mental causation. I would like to qualify this claim. To say that what I am aiming for is a rediscovery of the notion of mental quausation, as the title of this section puts it, is a bit of an overstatement. The notion surely is not an obscure one, neither one that has been completely forgotten. Horgan's 1989 paper saw some influence at the time, after all, and is still cited in comprehensive introductions to the topic of mental causation (e.g., Robb & Heil, 2021). It is also very important to note that the general idea, even if not under the label of "quausation", has been present in works in the literature since then. One example is the type of objection raised by, for example, Noordhof (1998),

Shoemaker (2003), and Macdonald & Macdonald (2006), against David Robb's (1997) trope theory approach to mental causation, saying that it still faced the problem of whether tropes cause in virtue of being mental (the intended, "quausal" result) or of being physical[34]. There seems to have been waning interest, however, in addressing this notion centrally and explicitly[35]. While Horgan never retracted from his main points raised in his 1989 paper and considers his subsequent work to be compatible with it[36], he went on to focus on other aspects when addressing mental causation, such as his commitment to a contextual element in causal claims (e.g., Horgan, 1998, 2001) and what can be described as the phenomenology of agency (e.g., Horgan, 2011, 2015, forthcoming). What I have as an intended aim, then, is that this general idea behind the notion of mental quausation be addressed more explicitly in the literature, especially if, as I argue, it is required to capture the common-sense view of mental causation.

I would like to defend Horgan's points and the central claim that, indeed, the notion of mental causation that we pre-theoretically have is a notion of mental *quausation*. Therefore, any account of mental causation that is minimally oriented by this pre-theoretical notion (either by assuming it in framing the problem or by attempting to actively vindicate it) must accommodate the "quausal" requirement – that is, it must be an account of mental quausation. As it is usually the case with claims about what would be our common-sense or pre-theoretical notions[37], it is hard to defend one's reading without appealing to what seems "intuitively" correct. Though I believe most of the appeal of my defence will indeed come from such a recourse to what is intuitively correct (and it would be hard if at all possible to convince those utterly unengaged by this appeal), I do think there are some independent reasons to consider the notion of mental quausation relevant. I will start, then, with an attempt to mobilise the relevant intuitions to make attractive the claim that what we pre-theoretically expect of mental causation is mental quausation, and will then proceed with a brief exposition of independent reasons to consider the general idea relevant.

2.1 The intuitive appeal

If one does not see any problem with maintaining that (i) there can be causes in which mental states participate, (ii) the effects of these causes are not in any way "indebted" to the mental aspect, and that (iii) these cases are genuine cases of mental causation

[34] For replies, see Robb, 2001 and 2013, where it is labelled "Objection 4".
[35] A notable recent exception is Bernstein & Wilson, 2016.
[36] Terrence Horgan told me this in private correspondence.
[37] As should be clear by now, I am using these terms interchangeably.

nonetheless, then I do not see what could be said to convince this person of the contrary; apparently, they would simply have a different idea of what mental causation is about. What I can do is highlight the factors that make the notion of mental quausation seem very appealing. Let us go back to the cases of *Thirst* and *Bitter Drink*. It seems that feeling of being thirsty and the bitter taste, respectively, are crucial parts we should mention if we wanted to describe what happened in these cases. After all, it seems clear that it was *because of these feelings* that the effects occurred. John went to the kitchen in search of water because he felt that dry throat feeling associated with being thirsty. Paula threw away the drink because she felt a bitter taste when drinking it. Or consider a similar case where a person cries because a song sounded beautiful. It seems that it was because the song sounded the specific way that it did to that person (that is, that they had that specific auditory experience) that they cried. In cases like these, not only does it seem correct to say that the agents did what they did because they experienced those feelings, but it also seems correct to say that they would not have so acted were these feelings not present. They are indispensable factors in any true and complete description we can give of the causal transactions that occurred (a point that will be crucial in my version of the exclusion argument in Part III). Alternatively, we can say that the causal chain involved in these examples would have to eventually trace back to those feelings. Any causal story that tried to accurately describe what happened and that did not trace the effects back to those feelings would seem to be missing the *key* causal factor. As Fodor put it in that famous passage I cited in Part I (Fodor, 1989, p. 79), we intuitively take the itching to be an indispensable factor in causing the scratching. We scratch because if itches (because we *feel* the itching feeling). Any view that has a result that this is not the case, that is, that it was not because of the itching feeling that we scratched, or, in other words, that the itching feeling played no role in causing the scratching, would go against our basic assumptions about what mental causation is supposed to be. It would imply we are deeply wrong about it. In essence, if all the alleged cases of mental (conscious) causes were like this, with the mental factors being causally idle, we would be inclined to say that EM is simply false. Or, at least, that it is "as good as" being false, since the unpalatable results that follow from EM being false would presumably also follow if the mental factors play no role in making effects happen, even if we could still maintain that there are mental causes "in some sense" (such as Davidson's account according to its critics or a notion of supervenient causation[38]). In other words, if

[38] Supervenient causation is a type of position that has a form like this: if P supervenes on Q and Q causes E, then P can be said to cause E. I believe this view is better understood as a form of "dependent causation", where the relevant factor is the fact that P is in some sense ontologically dependent on Q (that is, P is related to Q via dependence relations such as reduction, realisation, or constitution). That is, P would be said to cause E simply because it is in some way ontologically dependent on Q. In this case, it would be assumed that

there is no mental *quausation*, then the unwanted results follow even if there could be mental causation in some sense. After all, we would not be able to say truthfully that we got startled because we heard (experienced) a loud noise, that we cried because the music was beautiful, that we threw away the drink because it tasted awful, or that we scratched an arm because it felt itchy. Provided that, in such a view, these mental states are still real, we would end up with a form of epiphenomenalism that seems to be epiphenomenalism enough given its results.

A different way to motivate the same idea involves a thought experiment. It is formulated in terms of the idea of a philosophical zombie, made famous by David Chalmers (Chalmers, 1996). Philosophical zombies ("p-zombies" or just "zombies" for short) are, in short, beings who are just like us in their physical constitution and behaviour, with the exception that they are not conscious. The relevant notion of consciousness here is the one I said I would be using, that of phenomenal consciousness/awareness. It as topic of contention whether p-zombies are metaphysically or even logically possible, but let us consider, at least *in arguendo*, zombie versions of the people in the examples of mental causation I used before. By this I mean beings who are just like those people in their physical make-up, but that lack consciousness. Now, notice that I left out the assumption that these zombies would *behave* just like the original characters – I only asked us to consider them as perfect physical duplicates. Though this makes the zombies I ask us to conceive in this case different from how they are usually defined, this is on purpose, as this is precisely the claim my thought experiment seeks to assess. The conclusion I think we would intuitively derive from these zombie scenarios is that we would not expect the zombie counterparts to produce the same effects that the conscious people produce in the same circumstances. For example, take the zombie version of Paula from *Bitter Drink*. If we assume that Paula only spilled the drink after drinking it because it tasted terribly bitter, it seems we are committed to the claim that, if she had not felt the bitter taste (or another unpleasant taste that would elicit a similar reaction), she would not have spilled the drink in that circumstance. Zombie Paula would, by hypothesis, feel nothing when drinking a similar drink in an analogous situation. She, just like any other zombie, is "dead inside", has no phenomenal consciousness. Since Zombie Paula would not feel a bitter taste when taking the drink, it seems that, if we agreed that Paula only spilled the drink because of the taste she felt, we would be committed to the result that

supervenience is a form of ontological dependence, which, strictly speaking, it is not (see Kim, 1990). Notice this could hold in cases where P plays no role in causing E, with all the "causal work" being done by Q. In a case like this, we could have a position in which there is mental causation (with mental states causing an effect by supervening on physical states that cause that effect) but no mental *quausation*.

Zombie Paula would not react like original Paula in that case, that is, not spilling the drink (and, presumably, not having a disgusted look on her face right after drinking it, etc.).[39] Or consider again the perhaps more compelling example of scratching an itch. If we think that someone scratched a given part of their body because (and only because) they felt that part itch, then we would expect a zombie version of that person not to scratch that part, since they cannot feel anything.

If you felt compelled by the previous thought experiment and considered the indicated result that zombies would not produce the same effects as their conscious counterparts in those circumstances, then this exercise in thought was successful in making mental quausation seem at least more plausible as a prerequisite for any theory of mental causation. After all, it lends support to the idea that the phenomenal aspect of conscious states is indispensable in producing the effects we attribute to them. But perhaps you did not feel moved by this thought experiment. Worse yet, it might have made you reconsider an initial commitment to the claim that the phenomenal aspect is crucial in leading to the expected effects, as you might have now concluded that the merely physiological phenomena occurring in the bodies of the zombie replicas would have been sufficient to produce the same effects. You would then conclude that p-zombies are like usually defined: just like us in physical make-up *and* in behaviour. If that is the case, that does not dismantle any intuitive appeal the notion of mental quausation could have, as there is another thought experiment to consider that I believe makes the case for it even more compelling.

Consider now beings who are just like us in their physical make-up and who, unlike p-zombies, are conscious. This time, the peculiar feature they have is that they have experiences that differ in kind from ours given the same stimuli and underlying physical processes.[40] For example, while we feel a burning pain when touching a hot stove, they would feel a tickle, or when eating vanilla ice cream, they feel the taste we normally feel when eating pickles. That is, the phenomenal features of their conscious states for at least some given states differ radically from ours – they experience different kinds of feelings under the same conditions. Let us call them "altered" versions of their normal counterparts. Suppose, then, that Altered Paula feels not a bitter taste, but in fact a pleasant, sweet taste when drinking the

[39] The conclusion here is not that zombies could never produce the same instances of behaviour (that is, the specific pattern of bodily movements) as their conscious counterparts. Rather, the conclusion is that they would not produce the same effects *given the assumption that* the specific mental states the original characters experienced is indispensable in causing the original effects.

[40] These cases involve an idea similar to the famous inverted spectrum scenarios. They are more general, however, as they do not imply an idea of inversion (i.e., of having an "opposite" phenomenal quality) and are, of course, not restricted to chromatic qualities.

same drink of the original example. Or that, instead of an itchy feeling, the altered version feels pins and needles on their leg. What would we expect the reactions of these altered counterparts to be? It seems natural, I assume, that we would say Altered Paula would *not* throw away her drink (in fact, she might gladly continue drinking it), and that the pins and needles person would probably not scratch their leg. If we conclude this, then it seems it is because we attribute causal relevance to the *specific* felt phenomenal qualities in producing the effects in question. No other feeling (surely no other feeling of a radically different *kind*) could take their place in producing the same effects. If this is so, then it appears we are intuitively committed to the idea that it is because of the specific phenomenal features of their experiences that conscious beings produce the effects that they do. And this is nothing other than a commitment to a view of mental quausation.

I take it, then, that a position that maintains that there is mental causation but denies mental quausation would be one that maintains that the mental features of the alleged cause were not causally relevant for producing the effect or did no causal work in making the effect happen. What the considerations I raised above would indicate, if I am correct, is that our intuitive position is to assume that mental causation requires mental quausation.

2.2 Independent motivating factors

When it comes to independent reasons to consider the notion of mental quausation relevant, there are two that I can outline at present. The first is related to the phenomenology of agency, that is, the phenomenal (experiential, conscious) component associated with being an agent, a being capable of acting. As I mentioned above, this is a topic that Terrence Horgan devoted more attention to since his 1989 paper on mental quausation. As Horgan (forthcoming) discusses, some, such as Nida-Rümelin (2018), view our phenomenal awareness of ourselves as agents as providing evidence in favour of views about the underlying ontology of what is involved when conscious agents act in the world. Specifically, that we cause our actions in a way that is not determined by a previous chain of events (especially, of events that are outside this phenomenal, subjective perspective, such as physical events involving our brains and our environment).[41] This is a very controversial view, but it shows how one can find support in the realm of philosophy of action to the view that it must be because of the eminently mental (in this case, phenomenal) aspects of our experience that we can cause certain effects in the world (our actions), a view in line with the

[41] Whether or not this commits one to a theory of agent causation is open to debate, as Horgan remarks in the first section of his forthcoming paper.

quausal requirement. This general point can be found in earlier works in the literature, such as Kim's defence (2010a) that we need to incorporate the agent's first-person point of view (which we can understand as being subjective, phenomenal) in explaining actions.

The second independent consideration in favour of the relevance of mental quausation is similar to the first, but this time focuses on the topic of free will. Related to the topic of the phenomenology of agency, the idea now is that, in order to vindicate an adequate notion of free will, simply defending an account in which beings like us can willfully cause their actions in some sense will not do. Some further qualifications in the cause are required. As Bernstein & Wilson (2016) put it, we would like to vindicate a picture of free will in which the causes of our actions cause them *qua* being free, that is, *qua* being instances of mental states that are "qualitative, intentional, and freely deliberative" (*ibid*, p. 315). The problem of maintaining this result in face of considerations that seem to undermine it (the case in point are considerations in favour of causal determinism being true) is, they argue, analogous to the exclusion problem in terms of mental *quausation* as Horgan identifies it. In fact, they argue that both are instances of more general problem involving mental quausation, which they call the "general problem of mental quausation":

> *The General Problem of Mental Quausation.* How can a mental event M of a given type be efficacious vis-à-vis an event E in virtue of being the type of mental event it is, given that there is reason to think that events of M's type are causally irrelevant to the production of events of E's type? (Bernstein & Wilson, 2016, p. 314)

If Bernstein & Wilson are correct, the problem of quausal exclusion, as we might call it (how can mental causes cause in virtue of their peculiarly mental features given the premises of the exclusion problem?), would be an instance of a broader type of problem of which a problem for free will would also be an instance of. This allows them to draw parallels between positions in the mental causation debate and positions in the free will debate; specifically, between hard determinism[42] and eliminative physicalism on one side and soft determinism and nonreductive physicalism, the ones usually pressured by the general problem, on the

[42] "Hard determinism", as the term is usually employed, names the thesis that causal determinism is true (that is, every event (except, perhaps, the first) is causally necessitated by previous events together with natural laws) and that we (and beings like us; I'll omit this qualification henceforth) do not have free will. "Soft determinism", on the other hand, usually names a form of compatibilist view about free will – that is, that even though causal determinism is true, we still have free will in some relevant sense.

other (*ibid.*, pp. 315-319). If this approach is correct, then we could see reasons to see the relevance of the general idea of mental quausation coming from discussions about free will.[43]

I believe (and hope) that the considerations I raised in section 2.1 are enough to make a decent share of the philosophical audience feel persuaded to believe that the intuitive notion of mental causation we would like to vindicate is a notion of mental quausation. At the very least, I believe my points serve to show that (and why) the notion of mental quausation might seem attractive to relevant parties of the debate on mental causation, so that it becomes at least interesting to assess the implications a commitment to this general idea would have for the problem of exclusion. Though I believe the prospects of convincing someone who was not moved by the intuitive considerations of the relevance of the general idea are bleak, I believe it was worthy to include in section 2.2 the two types of possible independent reasons to adopt the general idea, as it exemplifies that not all the motivating factors for the commitment to it need to come from direct appeals to intuition.

Having made my case for the general idea of mental quausation, I can now explain the relevant tweaks I apply to Horgan's original proposal to arrive at the version I will favour and employ. This is the task I turn to in section 3.

3. Mental Quausation revised: How my approach differs from Horgan's

As I said above, I believe Horgan is correct in many of the points he identified about mental causation, but I also said there would be important differences in the way I will interpret and use the general idea of mental quausation and Horgan's original proposal. In this section, I will outline what I believe are the most important ones.

3.1 The role of reasons

Horgan's exposition in his original article involves frequent mention of "reasons", most centrally in his assumption (number (iii) in the passage from his 1989, pp. 47-48 reproduced above) that every action is caused by a "primary reason", which he understands as a reason in a sense in line with Davidson's view of explanatory reasons of actions in terms of beliefs and desires[44]. This makes sense when attempting to construct a theory that explains actions,

[43] For other works that address the relationship between mental causation and free will (and especially the implications of epiphenomenalism to free will), see Kim, 2005, and Merricks, 2001, chap. 6.

[44] Horgan defines a primary reason thus: "a token primary reason for performing an action of kind K is a complex state consisting of a token belief b and token desire d which jointly "rationalize" performing an action of kind K, in the sense that b is a belief that performing a K-action would (or probably would) bring about the object of the desire d." (Horgan, 1989, p. 48).

especially in the original context of directly addressing the points raised by Donald Davidson. The relationship between the mental cause of a certain action and the motivation or rationale behind is an intimate one. In cases like my examples *Thirst* and *Belief*, we would feel inclined to point to the beliefs, desires, and perhaps even feelings of the agents as constituting both causes and motivations for their actions. We appeal to the fact that John was feeling thirsty and wanted to quench his thirst or to Jimmy's belief and his desire to win the bet to make their actions *intelligible*, as *actions* coming from rational agents. We invoke reasons to make actions intelligible, but also in this purely causal sense: the action occurred (at least in part) because the agent entertained certain reasons to act in that way. The "because" clause has this ambivalence where we can interpret it in these cases either in the causal sense or in this broader, explanatory sense. While I recognise the naturality of this connection between causes and reasons for action and believe any complete account of actions will require addressing reasons, it is not a connection I will make. I will be interested in the purely causal factors of the phenomenon we call mental causation, considering mental items as elements of causal transactions irrespective of their constituting reasons or not. This will simplify matters and not derail the topic of discussion to the domain of reasons and the philosophy of action, where other considerations become pertinent.[45] [46] I believe this choice will have no serious consequences for what I will defend about the notion of mental quausation,

[45] I am also interested in the broadest category possible of mental causation in the way I defined it (causal transactions with a mental cause). This will include cases that do not count as reasonable actions made by agents, or even actions at all.

[46] Though intimately connected, causes and reasons behind actions are still relevantly different. A puzzling example that helps illustrate this difference is the following: suppose someone shoots at you, but you escape the shot by crouching. You avoided being shot because you crouched, dodging the bullet. But you only crouched because there was a bullet coming towards you. The bullet would explain in some way (e.g., would be the truthmaker of "you crouched" or of "you dodged a bullet") your dodging, but it presumably would not be a causal explanation of why you were not shot. It would be odd, after all, to explain that what caused you not being shot (or your survival) was the shot that was fired towards you. I think the contrast between explanation (even causal explanation) and causation as illustrated by this example is very puzzling and I would have to think more before taking any more definite stance. What I think, for now, that can be extracted from it is that it might suggest that the relation of explanation is not transitive. The fact that you were not shot (or killed) is explained by the fact that you crouched. This fact in turn is explained by a bullet coming towards you. But the fact that there was a bullet coming towards you presumably does not explain the fact that you were not shot (or killed). I think the most interesting conclusions to be taken from the example concern explanation, not causation. "Not getting shot" or "Not being killed", being cases of negative events, should require something like paraphrase suggested by Dowe to be dealt with (see Dowe, 2001). This would involve identifying some positive event that, had it not happened, would have led to you getting shot. So if we go back in the causal chain, we would find that a cause for this positive event was the event of a bullet being directed at you. But this no longer seems to be a problem: the fact that there was a bullet coming is causally responsible for you detecting it, which in turn was a cause for your reaction – crouching. No problem here. So, as has always been the point, the problem seems to depend on the kind of description we give of the phenomena involved. And this seems to affect explanation (understood as the kind of explanation we evoke to make things understandable), not causation. Put in another way, it is not a puzzle about the phenomena that occurred in the world, but about how we make sense of them. I thank Pedro Galvão (who was tempted to draw the opposite conclusion, that it affected causation rather than explanation) for bringing up this example.

especially as I have no intention of preserving the original notion exactly as it was elaborated by Horgan.

3.2 The number of places in the Quausal Relation

The quausal relation as formulated by Horgan, as we have seen, is a four-place relation: C *qua* F causes E *qua* G, with E and C being events and F and G being properties. The "*qua*" clause shows up twice, qualifying both cause *and* effect. I will not consider the *qua* clause for the effects, so my general idea of a quausal relation will be three-placed: C *qua* F causes E. Put in another way, without the "*qua*" construction, C causes E in virtue/because of F. This time, this is much more a choice for brevity than a substantive point of disagreement with Horgan. I believe the qualification for effects is a natural part of our discourse about causation, a feature that deserves serious philosophical attention. I am just supposing that it is something we can isolate from the discussion in terms of what qualifies the causes, which is the central topic of discussion. Given a mere decision of what to focus on in this work, then, I will be devoting attention to this three-place version of the relation. There is a brief remark I would like to make about the qualification of the effects. What I believe is behind most of these common constructions, and that I do indeed lose the ability to directly accommodate in my favoured phrasing of the transactions, is a contrastive claim involving counterfactuals. That is, claims of the form "this rather than that" involving comparison with scenarios that did not actually occur. Examples would be "the bitter taste that I felt caused me to throw away the drink rather than keep drinking it", "Johns feeling of thirst together with his desire to quenched it caused him to move to the kitchen instead of the garage", or "Taking the medicine caused my survival rather than my death". These contrastive clauses could then be phrased with the *qua* construction like in the following example: "the taste *qua* being bitter caused my action *qua* spilling". I do not think, however, that we lose the ability of making these claims by taking my approach. They just happen to be different, comparative claims. Though I will not address these contrastive considerations in this work, I believe Dowe (2001) has shown an elegant way of plausibly handling them that is "agnostic" when it comes to the preferred theory of causation (see note 18). One would be in more serious disagreement with my view, however, if one believed that these contrastive clauses are behind *all* causal claims. This is a deeper question I will not address here due to topic and space constraints.

3.3 The Role of Properties in Mental Quausation

A feature of Horgan's formulation of mental quausation is that properties show up as terms of the relation. That alone is a significant point of departure from the standard position of treating causation as a relation exclusively between events. A peculiar feature of the notion of quausation, then, is that properties have centre stage. Horgan's formulation in terms of "properties" is ambiguous between property types and specific instances of properties. To be more precise (and I do not know if this would constitute a point of departure from Horgan's formulation), I will treat quausation as involving property instances, that is, specific occurrences/exemplifications of properties at a given place at a given time. For example, suppose I want to specify that a red ball caused some effect in virtue of being round or of being red. Instead of "roundness" or "redness" entering the relation, it is the specific round shape of that particular ball or the specific shade of red it exhibits that would feature in the quausal statement instead. We could still say that the ball caused something *qua* red or *qua* round, but that should only be understood as an abbreviation. The more precise description would be that the ball caused what it did in virtue of the specific shape or hue it had at the relevant moment.

Talk of "property instances" might suggest that I am presupposing that properties are universals which can be instantiated. This is not an assumption I am making – in fact, I will make no substantive assumptions about the correct view on the nature of properties. I am using the term "property instance" since it has become widely used in the literature, but a term like "token property" or anything that indicates that the specific properties occurring at a given moment are relevant would work just as well. Though I believe it will not matter for my exposition here, I am considering some restrictions, however, on the kind of property that I believe can feature in causation (and in quausation, for that matter). I will be considering only concrete properties[47], and I think especially of non-relation properties (properties related to location being some of the few plausible exceptions). Examples of the type of property I have in mind include those related to the shape of objects and those related

[47] The term "concrete" is famously hard to define precisely. Usually, concrete entities are assumed to be those that exist (or can exist) in space and time, but that is still not very precise and contentious. I do not have a favoured characterisation (let alone an outright definition) of what "concrete" amounts to. What is usually safe to do, though, is contrast concrete entities with abstract entities. And that is something minimally informative about the kind of property I am considering: they are not abstract properties. Again, I do not have a precise characterisation of what *those* are, but examples would be properties related to presumably abstract entities (such as numbers; see my examples about the number of spikes in the main text) or any view in which properties instances are assumed to be abstract entities in general (such as what is usually assumed in trope theory).

to physical quantities like mass, charge, spin, etc. I am aware some of these physical quantities are arguably inherently relational. If that is indeed the case, I would have no problem including them on my list of exceptions. The type of "relational" property that I want to exclude from the list of possible types of property that can exert causal influence are those properties that are not in some sense features of the objects themselves, but merely something that can truly be said of them[48]. Properties such as "happening 1750 years after the death of Genghis Khan", or "having a number of spikes that is divisible by three" could not enter the relation in my view. I am assuming that phenomenal properties, the type I am centrally interested in, fit this restriction (that is, they are (arguably) non-relational and concrete, at least in a broad sense of "concrete" in which non-physical things (which they might be) can count as concrete). Again, I doubt this restriction is at all required or will make any relevant difference for my main points, but it helps to understand what kind of property I will have in mind when discussing them.

Taking a step further than what Horgan originally assumes, I will be assuming a view of causation in which properties have an even more explicit role. I believe we should take the idea of an effect happening because of some property that the cause had quite seriously and at face value. To capture this, I believe we would have to be committed to a view of causation in which property instances do "the causal work", that is, that they are what is responsible for making the effects in question happen. I will be committed, then, to some form of *property causation*, that is, causation in terms of properties.

To explain what I mean by "property causation", it will be helpful to start with an illustrative example and only then move to a characterisation. Consider the example of a ball hitting and breaking a window. Under what we can call "object causation", that is, causation in terms of objects, we could say that the ball caused the window to break. Notice how we related two objects (the ball and the window), with an object (the ball) being the cause of some alteration happening to the window (an object as well). This is a very natural way of describing causal transactions, so much so that we might consider that the naïve theory of causation of most people is a form of object causation. The same example would be described under event causation (where events are the terms) as 'the ball hitting the window' causing 'the breaking of the window'. That is, the event "the ball hitting the window at *t*"

[48] What I surely want to exclude are properties involved in what is usually called "Cambridge changes", which are mere changes about what becomes true of a thing without involving any actual change in that thing. An example is "becoming an uncle".

causes the event "the window breaking at *t**", with *t** being later than *t*.[49] A picture like this might seem more appropriate once we note that it is not just the ball, on its own, that breaks the window. The ball sitting still would not accomplish this, for example. An *occurrence* involving the ball would then seem the more appropriate candidate for the cause, and the same applies to what happens to the window. We might notice, however, that not all properties of the ball are relevant in producing this effect. Surely the colour of the ball and its precise shape make no difference in breaking the window. Its mass, however, and, better yet, its *momentum* do make a difference. We can then entertain a view of causation that makes these distinctions and attributes to certain properties of the object (or collection of objects) the cause of the effect. This would be property causation, and a description of this example under such view would be something along the lines of "the momentum x and the contact area of measure y of the ball hitting the window caused the window to break". The way I will be using the term, "property causation" is characterised as any theory of causation in which the effects are deemed to occur because of (i.e., caused by) properties of the entities involved. Notice that we do not need to stop mentioning events in describing causal transactions according to property causation, as in my suggested description of the case of the ball hitting the window. In order to count as property causation, the central requirement is that properties are being assigned a relevant role as causes; it is because of *them* that the effects occurred.

It is worth noting, then, that a view in which properties might appear as constituents of the terms of the relation, but where it would not be true to say that it is because of *them* that the effect occurs, would not be property causation view. Perhaps Davidson's view would be an example of this, as causation on that view happens at the level of events, even though these events involve the occurrence of properties.

I do not know of any detailed account of property causation. L. A. Paul's aspect causation (Paul, 2000) is the famous notion that comes the closest, but it is relevantly different as it is in terms of aspects of facts rather than properties of objects[50]. It is worth mentioning that I by no means have a strict commitment to properties being the type of

[49] This is a simplification, as each shard of glass moving at a given instant could be considered its own event caused by the ball hitting the window.

[50] Shoemaker's views (e.g., in Shoemaker, 2007) could perhaps count as a such a theory, though he pursues a slightly different direction from the one I will take. His notion of realisation of causal powers can be approximated by the principle that I will assume as P1 of my Quausal Exclusion Argument in Part III. The approximation I see is that, under the principle I defend, a property instance P being ontologically reducible (in a broad sense that might include "realisation") to property instance P* would allow P and P* to have the same causal powers. That is similar to Shoemaker's thesis about a realised property having the causal powers of its realiser. For a critique of Shoemaker's view, see Kim, 2010b.

entity required. I only believe it is the natural assumption to make, one that is also prominent in the literature. I find it hard, for example, to understand the difference between "aspects" and "properties" in some usages (those, of course, in which they are not synonyms) – consider, as an example, the task of explaining the difference between property dualism and aspect dualism. My view can easily be reformulated in terms of "aspects" of objects if such change, for some reason, would be suggested as better, as long as the view of aspects in question captures the requirements I outlined above. Any view where "features" of mental items, in the broadest sense of "features", can play this central causal role would work to capture the general idea I intend. Although a detailed theory of what property would be beneficial for my overall argumentative strategy, I do not have such an account to offer here. I will assume that such a theory can be provided and that it would imply no result that contradicts other features of the picture I will defend. If pressed, however, I would currently bet that a very promising route to offer a precise theory of property causation would be a powers-based account, in which the relevant properties have intrinsic causal powers[51].

Horgan's original proposal is rather ambiguous about what the relationship between the properties and the events is, and, specifically, about what are the bearers of such properties. There are passages where it seems that the properties are properties of the *events*. For example, when discussing Sosa's gunshot example, he says the cause of the death was not the event *qua* loud (Horgan, 1989, p. 50), and, more clearly, when discussing an example connected to Davidson's theory of action explanation, mentions the event of exercising, "*qua* the property *being an exercising* […]" (*ibid.*). His definition of the quausal relation, however, are more ambiguous; the first two conditions only require that the respective properties "be instantiated" by the events that constitute cause and effect (*ibid.*). This could be interpreted in at least two ways. According to one reading, the token events instantiate the properties by bearing those properties. They are properties *of* the events. According to the other, the events involve the instantiation of the properties; that is, it is part of the occurrences that constitute them that those properties are exemplified. In this case, the properties are properties of the objects constituting the events and occur "in" the events. A good example to illustrate properties occurring "in" an event (as constituents) versus properties of an event is the event "a red light flashes". The property of being red (or, more precisely, of being of a certain shade of red) is a property that features in the event, as a constituent; it is part of what the

[51] A view along these lines is defended in Gibb, 2015.

event is. The property of occurring quickly, arguably, is a property of the event; it is an event that has a quick duration.

I will assume the second interpretation, focusing on properties of *objects* occurring *in* events, not properties of the events. Since Horgan's own definition of the relation seems ambiguous with respect to these two readings, I believe it will not prove to be a very substantive point of departure from his original idea if he happened to favour properties of events instead.[52]

All of this might seem very broad, but I believe these commitments already have significant results. At the very least, a commitment to the general notion of mental quausation so construed commits us to a rather fine-grained view of causation, at least clearly more fine-grained than views in which analysis of the terms of the relation stops at the level of events. In fact, it seems plausible that a view even more fine-grained than what initial appearances indicate might be required. Some ordinary examples might allude to this. I will offer one example that I believe is particularly good for illustrating what I mean. The example involves wine tasting. Consider what happens when a sommelier takes a sip of a fine wine, attentively introspects on its taste, and makes a descriptive report about it. They might report something like "this wine has some flowery notes, with some vanilla residue typical of oak barrels", or whatever it is that sommeliers say. The important thing to notice is that *notes of a taste* are a relevant part of the wine taster's report. This gives us occasion to think about the important and interesting topic of how precisely we should individuate property instances. While I will not offer an answer here, I believe the wine taster example suggests that a very fine-grained way of doing so might be required. Notice that the taste of the wine, described in this broad sense, could reasonably count as a property instance on its own. We could say, as in the case of *Bitter Drink*, that "the taste of the drink[53]caused Paula to spill the drink". That seems

[52] I certainly see that there is some appeal to the idea that the properties of the *events* are relevant, as we naturally point to these features of events as the causes in many cases. For example, the properties of occurring quickly or slowly are frequently pointed to as causally relevant, such as in "You hit your head on the low ceiling because you were too slow in ducking" or "He lost consciousness because he was raised too fast to the surface". I believe, however (with a not so high level of credence in this assumption), that these features can ultimately all be accounted for in terms of properties of the *constituent* (the objects involved) rather than of the event(so, keeping with the examples, we could offer paraphrases like "you hit your head against that particular place in that particular time because (a relevant part of) your head moved towards it at that time"; the pressure case is more complicated, but would involve appeal to the physiological processes that the body underwent and their cause in the environment). I believe these examples usually involve the modal contrastive claims I addressed earlier in section 3.2, which I believe can be accommodated in paraphrases of a Dowean kind. See also note 18.

[53]The taste of the drink *at that moment*. I will omit these time-stamp qualifications henceforth in my examples, just remember they are tacitly there.

correct, and to already capture the quausation spirit. But notice that we could also qualify it further, as "a *bitter* taste", which sounds even more appropriate. What the wine tasting example shows is that we can take this even further, speaking of notes of a taste (which, remember, could already be regarded as a property). This almost suggests talk of "properties of properties", though of course that is more a manner of speech than a suggestion in ontology. The upshot is that any plausible account of property causation most accommodate cases like the wine tasting one, where features as specific as the notes of a taste might be relevant.

There are some further commitments we can extract from the general view of quausation I outlined that I believe are relevant to highlight. Going back to the questions about causation that I listed in Part I, we can already identify at least some broad answers to the following questions: (I)"In virtue of *what* do causes cause their effects?", (II) "What constituents of causes "do the causal work"?" (which I said above can be considered a rephrasing of (I)), and (III) "What kinds of entities feature as terms in the causal relation?". The answers would be, respectively: (AI) Of the specific properties they instantiate at that moment (and the relevant specification might be very specific, as the wine tasting example indicates), (AII) The property instances, and (AIII) events (occurrences) and properties of the constituents[54] of these events.

A final comment I believe is important to make here is that the "*qua*" construction used in Horgan's formulation can be paraphrased in other terms. Instead of saying "event c, *qua* F, causes e", we would say, according to my reading and in a little more detail, "event c, in virtue of property instance F occurring in it[55], causes e". Despite the issue I discussed above about property instances being properties of events vs occurring "in" events, which may or may not be a point of departure from my view and Horgan's, I believe this paraphrasis that replaces the "*qua*" terminology with "in virtue of" was already available to Horgan's original account. The important takeaway here, then, is that the "*qua*" construction is not essential to the notion of quausation. The key feature, instead, is an appeal to specific *features* of the cause as being relevant in bringing about the effects.

[54] Objects being the natural choice for these property-bearing constituents.

[55] Alternatively, and in a way that would perhaps be more natural to many, we could phrase it such that "F" stands for a property type, so that we would talk about "instances of F". This seems closer to Horgan's original phrasing. This can be faced either as a stylistic choice or as one that at least alludes to a more substantial difference. I chose to have "F" stand for the specific property instances involved because it is property instances that have centre stage in my view. The fact that they are instances of a broader type will not be what is appealed to.

3.4 Causal relevance as causal indispensability

The last feature of my view of mental quausation that I believe is relevant enough to deserve its own subsection is the way I interpret the notion of causal relevance, and, more generally, Horgan's clause (iv), the "quausal relevance requirement". The way Horgan phrased this requirement was, recall, that "the fact that c instantiates F is explanatorily relevant to the fact that e occurs and instantiates G." (Horgan, 1989, p. 51). And we can understand "explanatory relevance" here to be "causal relevance", as the type of explanation in question are causal explanations. As I mentioned before (especially in section 3.1), talk of "causal explanation" blends together elements of what is supposed to occur between the relevant entities, the phenomena "in the world", and epistemic elements related to how we describe causal transactions and how we understand them. Since I also said that I would be assuming a causal realist view and that I would not focus on these epistemic matters involved in a broad notion of "explanation", I think it will be more precise to frame my discussion in terms that are not directly connected to explanatory considerations. As a first modification, then, I would rephrase the quausal relevance requirement in terms of the (still broad) notion of "causal relevance":

(iv*) "The fact that c instantiates F is causally relevant to the fact that e occurs."

A crucial term that I will be using instead of explicitly explanatory terms like "causal explanation" is "causal description". What I mean by a causal description is a statement that describes (or attempts to describe) a causal transaction. To illustrate, suppose you see one of those drinking bird toys (those that rock back and forth due to the heat engine due to the liquid they carry) pushing a ball. The ball then rolls down a slide. You could then say "the toy bird pushed the ball, causing it to roll down the slide" or "the ball rolled down the slide due to the push given to it by the toy bird". These statements would be examples of causal descriptions. I am assuming that causal descriptions can be true or false, and I am also assuming a broad correspondence theory of truth about their truth conditions. That is, roughly, a causal statement will be true if the transaction you described matches (in some way) a relevant[56] transaction in the world – otherwise, it is false. That is, more specifically,

[56] I will only briefly explain this qualification, as we can get lost in a rabbit hole very quickly. Not allowing just any transaction in the world that matches the description serves to block certain cases where someone seeks to describe one phenomenon, gets it wrong, but accidently says something true about some other phenomenon. For example, someone sees a magic trick and utters "the three of hearts was shuffled into the deck". That happened to be an illusion; the card the person thought was shuffled into a deck was instead quickly hidden in a sleeve, say. But suppose now that it just so happens that another magician who was practicing behind that person did indeed shuffle a three of hearts into a deck. That would make the statement uttered true, in a sense, but false in another (what the speaker intended to say; in this case, what was meant by

"the toy bird pushed the ball, causing it to roll down the slide" would be true if the toy bird indeed pushed the ball, causing it to roll down the slide (which is true in the example). This is the terminology I will use in phrasing my version of the causal exclusion argument in Part III.

Horgan's proposed way of offering an account of the quausal requirement involves a rather complex theory of counterfactual dependence. While I will not offer a proper *theory* to underpin what is behind my version of the quausal requirement (iv*), suggesting only that it must be a theory of property causation, there is a point about the way I will interpret it that, though subtle, will prove important. The point is this: I will be interpreting "being causally relevant for" as "being causally *indispensable* for". The intuitive motivation behind this reason is the following: if F is causally relevant for e, then it seems this means that any causal description that tell the whole story about the causal transaction in question must mention F in one way or another. F is, in this sense, *indispensable*[57] in a causal description of the transaction culminating in effect e. When I talk about "telling the whole story", there is an important element of *completeness* in play. I am surely not suggesting that the relevant causal factor must show up in *any* true description of the causal transaction in question. We can surely omit relevant factors and still say something true when we give a *partial* description of a transaction. In fact, on most occasions, we omit details that are not informative to mention. To illustrate, consider possible descriptions of what happens when we turn on a light by flipping a switch. If I say that "my flipping of the switch caused the light bulb to glow", I could be saying something true (provided that I did turn on the light by flipping the switch). However, there are some relevant steps in the causal chain that we are omitting, namely, those about what happens after the switch is flipped and before the light goes on, about the electricity being transmitted. If I wanted to give you a complete description of the causal transaction, I would have to mention what happened to the circuitry leading from the switch to the light bulb. Crucially, I would have to mention electricity. As facts about how electricity

"the three of hearts" might be a relevant factor). Hence the "relevant" qualification. In any case, I will not get into discussions about these matters (such as Gettier cases), as they take us too far off topic. I will just be assuming, for simplicity, that a true causal statement will be true about the relevant (perhaps, the intended) phenomena in each case, which are usually straightforwardly identified, ignoring these tricky odd cases.

[57] My use of "indispensable" can be understood as synonymous with "necessary", as in a "necessary condition". I prefer the term "indispensable" for two reasons. The first is that the term "necessary" might, in some contexts, naturally lead to an alethic modal reading, which can be misleading and suggest commitments I do not intend to make. The second reason is that I believe the terms "dispensable"/"indispensable" work well in easily conveying my intended sense: that the factors in question must show up in any true and complete description of the causal transactions in question.

travels through parts of the circuit are causally relevant for the lighting of the bulb, they are indispensable (that is, must feature) in a complete accurate description of the events.

To summarise, if a factor is causally relevant for a causal transaction (or, to be more precise, for the description of a causal transaction), then any description of that causal transaction that is true and complete must include that thing. That is, it is indispensable. In other words, any description that omits that factor would be missing an important causal factor, so the resulting description could not be true and complete (that is, it is either false or a partial truth about the transaction, leaving important parts missing).

There are many ways to make sense of what a "complete" causal description is. As I will defend in Part III, I will interpret a complete causal description as one that leaves no gaps in the causal steps.

As a final summary, I believe it will be fruitful to lay out my version of quausation through requirements like Horgan's. It goes like this:

For any two events c and e and any property instance F of the appropriate kind[58], c *qua* F causes e iff:

(1) c causes e;

(2) c occurs and instantiates F;

(3) the fact that c instantiates F is causally relevant to the fact that e. That is, F is indispensable in any true and complete description of the causal transaction leading from c to e.

4. Theses redefined: Quausal Efficacy and Quausal Epiphenomenalism

With the general notion of mental quausation laid out, we can now formulate versions of key theses in the discussion about the exclusion problem: the efficacy of the mental thesis (EM) and the relevant type of epiphenomenalism at stake.

If, as I argued, the notion of causal efficacy of the mental we want to preserve is truly one which involves quausation, then we can give a more accurate formulation of the thesis as

[58] Recall that the assumption I am making, but that is not required, is that the properties are concrete and non-relational.

Quausal Efficacy (QE): there are mental causes of physical effects in which the effect occurs in virtue of the instantiation of mental properties in the cause.

This is the version of the efficacy of the mental I will use in assessing the problem of exclusion. If my defence of the claim that our common-sense/pre-theoretical notion of mental causation is a notion of mental quausation is correct, then it is QE that is at stake with the problem of exclusion. Any theory that manages to vindicate this thesis while at the same time resolving the tension between the other premises could be said to have successfully solved the problem of exclusion in a way that "saves" the pre-theoretical claim that the mental is causally efficacious. Conversely, if QE is false, then we are left with a result that has all the unpalatable features of epiphenomenalism, even if we could say that there are mental causes of physical effect "in some sense".

We can now also give a proper definition of this qualified version of epiphenomenalism, which we can call (following Horgan) quausal epiphenomenalism:

Quausal Epiphenomenalism: there are instances of mental properties, but they do not cause anything[59].

Notice that Quausal Epiphenomenalism is not the same as the negation of QE. The negation of QE would amount to something along the lines of "there are mental causes of physical effects in which the effect occurs in virtue of the instantiation of mental properties in the cause", i.e., there is no mental quausation. This could be the case in two ways: 1. If Quausal Epiphenomenalism is true, that is, there are some mental states that do not cause anything in virtue of their specific instances of mental properties[60]. 2. There are no mental states/property instances whatsoever – that is, some antirealist/eliminativist theory about the mind is true. The former scenario is the one I will be focusing on the most, but the more general threat to QE that involves the latter case will prove relevant in my Mental Efficacy Trilemma in Part III.

[59] Some alternative phrasings could be: "there are mental states, but they do not cause anything in virtue of the instantiation of their mental properties" or "there are mental states, but they do not "quause" anything". The latter is illusively concise since it requires "unpacking" of what "quausing" amounts to.

[60] And this case, on its turn, admits of two possible scenarios: one in which the mental states in question do not cause anything ("classic" epiphenomenalism) and the other where they do cause some effects, but not in virtue of the properties they instantiate (that is, mental causation that is not quausation).

To conclude, I would like to draw attention to an important passage from Horgan's 1989 paper:

> Successfully fending off quausal epiphenomenalism will require doing two things: first, giving a plausible account of quausal relevance; and second, using this account to argue that the mental *qua* mental is causally efficacious. (Horgan, 1989, p. 51)

The main goal of this work is to address the tasks outlined in this passage. Regarding the first, I offered in Part II a very general interpretation of the requirements of mental quausation, a task that should be supplemented by a specific theory of causation (as I argued, the plausible candidate would be a theory of property causation). The central relevance of this work, however, should be found in addressing the second task, so much so that I believe its quality should be judged over how relevant my conclusions about it are. I will argue in Part III that the task of vindicating the quausal efficacy of the mental is harder than usually thought, the threat of quausal epiphenomenalism haunting us at every path we might take.

Part III
Epiphenomenalism Looming Large

We like to think that our conscious experiences are the cause of many things we do. We search for food because we feel hungry, look for medicine because our head aches, kiss because we feel an urge and make decisions because a certain option seems like the best one. To think otherwise would alienate us from our behaviour and radically alter our understanding of ourselves and our place in the world. After all, to say that our pains are not the cause of our wincing, that the itching is not the cause of the scratching, or that our wanting is not the cause of our seeking seems to take us "away from the steering wheel" and render illusory our understanding of basically everything we do. Epiphenomenalism is the thesis that this unpalatable possibility is actually the case: our experiences do not cause anything. As shown in Part I, the exclusion problem together with the exclusion argument were established in the specialised literature as the greatest hurdle in the way of a theory of mental causation, putting special pressure on non-reductionist views about the mind and mental causes. The major threat to these views is a collapse into a form of epiphenomenalism. In the third and final part of this work, I will argue that the challenge posed by the exclusion problem and the exclusion argument is harder to overcome than usually assumed. Central to this task is my elaboration of a new version of the exclusion argument, aided by the general idea of mental quausation I defended in Part II, which I call the Quausal Exclusion Argument. The upshot is that the only type of reductionism that would be immune to a collapse into epiphenomenalism is one that is not very popular. However, I argue that those who do subscribe to such a view should not be celebrating just yet, as I believe a more general problem applies to their view as well. I will introduce this problem in dilemma form, as the

Epiphenomenalist Dilemma, and in the even broader Mental Efficacy Trilemma. This problem would indicate that no account could vindicate Quausal Efficacy (QE), a devastating result given the defence in Part II that QE is the thesis we would generally want to preserve. I then present the opposing view I consider the most promising to my Quausal Exclusion Argument, the theory of causation as "making a difference" defended by Peter Menzies and Cristopher List. I offer my reasons to believe that this theory and any view in the same family would be unable to provide a satisfactory response to the exclusion argument, and discuss an important possible implication of this result. Finally, I highlight some problematic results of epiphenomenalism that I find the hardest to accommodate, emphasizing a particular problem about personal ontology that I believe has gone overlooked.

1. The Quausal Exclusion Argument

With the general idea of mental quausation I presented in Part II, I can now present my formulation of a new version of the causal exclusion argument. The motivation behind this reformulation, as should be clear by now, is that the version of the exclusion argument outlined in Part I does not capture the quausal considerations laid out in Part II. It is compatible, as we have seen, with cases in which mental events cause even though the *mental* aspects do no "causal work"; in other words, it allows mental causation without mental quausation. In so doing, it follows that it countenances cases of quausal epiphenomenalism among those that evade its threat of exclusion. It is "too coarse a filter", so to speak. In order to fully capture the threat of (quausal) epiphenomenalism, we need a different kind of argument, one that, as we will see, is more restrictive than the version we have considered.

I will be assuming the Causal Closure principle (CC) as a tacit premise, one I will not explicitly list in my presentation of the argument but that is there in the background. I will formulate the argument such that the tension falls on a form of Nonreductionism (NR), with the mental efficacy premise being assessed now being the Quausal Efficacy (QE) version I defined at the end of Part II. One important difference will be that my version will not rely on an explicit principle of No Overdetermination (NOD). Instead, I will rely on a broader principle of causal relevance/dispensability related to true and complete causal descriptions. This principle is arguably the most original contribution brought by my version of the argument, and one that marks the most noticeable difference to how the argument is usually formulated in the literature.

I believe it will be fruitful to show some of the development stages of the argument, as it engages with important qualifications in response to complications that would naturally

be found when thinking about it. I will start, then, with what seemed like a natural starting point in formulating the argument and show the considerations that lead to its final version. Keeping with the nomenclature, I decided to name this version "The Quausal Exclusion Argument", and the first version I was able to devise goes as follows:

Quausal Exclusion Argument (version one):

P1. If there is a true gapless causal description[61] for the occurrence of a given effect that makes no recourse to property P, then, if instances of P are not identical to instances of any of the properties mentioned as causes in this description, then P is causally dispensable* for this description. (*A property is causally dispensable for a description iff not mentioning it in a true causal description does not make it false or create a "gap".)

P2. For every physical effect the occurrence of which we might attribute to a mental cause, there is a true gapless description of the causal transaction leading to that effect that is not cashed out in terms of mental (phenomenal) properties.

P3. If P is dispensable in a true and gapless description of the causal transaction leading to that effect, then it is not in virtue of P that the effect in question occurs, i.e, P is not a cause of that effect.

P4. Instances of mental (phenomenal) properties are not reducible to instances of physical properties (Hypothesis for Conditional Proof)

Therefore,

C. If instances of mental properties are not reducible to instances of physical properties, then it is not in virtue of them that the physical effects occur.

Now, there is a lot going on in this argument. The first thing to notice is that it makes no mention of the key premises featured in the original version. This might seem more economical, but that does not mean their influence has disappeared, however. For instance,

[61] The notion of a "causal description" is the one I presented in Part II, which, in brief, amounts to descriptive recantations of the causal transactions in question, which can be true or false. Recall that the most important point here is to avoid the ambiguity between epistemic and "purely" metaphysical considerations that might come with the use of "explanation" in "causal explanations". I believe that formulating the argument in terms of a notion like the Davidsonian notion of "causal statement" would work just as well. I just prefer to use causal descriptions as to clearly refer to the specific type of statements that I had in mind. I thank Wilson Bezerra for the comment about the importance of being precise with these terms.

P2 is now our version of the Causal Closure of the physical domain. It is not exactly a *statement* of the principle, but it surely relies on it. The idea behind it is that we can always give a true physical (or, perhaps, neuronal) account of what goes on in alleged cases of mental causation. For example, I could say the bitterness of the taste was the cause of my spitting of the beverage I just drank. This is a "mentalistic" description of what happened. However, it seems true that we could also give a description that makes no mention of "bitterness", mentioning instead only the neural correlate of "feeling a bitter taste" (whatever that is). It would go something along these lines: "I receive such and such stimuli through my taste buds; that sends an electrochemical signal to my brain; I go through a given brain state (the correlate of bitterness, in this case); my brain sends and electrochemical signal to my nerves; I move the muscles in my mouth, spewing the liquid"[62]. Of course, this is all very rough, and some steps have been omitted, but it is sufficient to get the point across. The point where the CC principle enters the stage is when we rely on it as a guarantee that 1. there will always be a physical description (i.e., a description appealing only to physical entities) available for what is going on in our behaviour (and, it is presumed, the causal "story" just provided ultimately boils down to physical interactions) and that 2. they are complete in the sense that there will be no gap in the description, that is, that are no "leaps" (causal steps left unconnected), no glaring omission or occasion for something mysterious and "miraculous" to be evoked.

Premise 3 captures the very broad idea that the factor that causes something to occur is not dispensable in a true and complete causal description for the occurrence. That would be to leave out the very thing that matters! If it is dispensable, then it was surely not the "*real* causal factor" that made that effect occur. This was the point of Horgan's remark about the causal relevance we attribute to mental properties when we think it is *in virtue of them* that mental causation occurs. In this version, it shows up according to my interpretation of the causal relevance requirement of mental quausation as causal indispensability.

Premise 4 is the new version of the non-reductionist hypothesis. The key thing to notice here is that, together with premise 1, it narrows the scope of theories capable of

[62] This is an excellent example of the kind of naïve story philosophers usually come up with to illustrate the "completeness" of neuroscientific explanations of our behaviour. If asked to fill in with details, most of us would probably have some trouble and feel disconcerted. In reality, much less is known about the specifics of the neurobiology (not to say the physics) of mental causation. We should bear this in mind, but I believe it is not a reason to defeat the belief that there is always a description of this kind available. The point is that, even if we do not know the details yet, there is no reason to suppose that what is going on in these cases is something indescribable by our best science. And this is all the general idea behind the CC principle amounts to and requires to establish the main point.

evading the threat of exclusion. The idea that phenomenal properties are not reducible to physical properties (or to any kind of non-phenomenal properties, for that matter) is not without its discontents, but it is less controversial than theses such as, for instance, that mental *events* cannot be reduced to physical ones. For example, anomalous monism without a commitment to *property reduction* could no longer have a margin to escape the argument. Even if mental occurrences are identical to physical occurrences, if instances of phenomenal properties are not reducible to instances of physical ones, then the argument still works[63]. So it seems that the only kind of theory capable of evading the threat is some kind of reductionism about phenomenal property instances. Defending such a thesis is widely regarded as the greatest challenge faced by reductionists, since the nature of qualitative states seems to defy any kind of such reduction. My point that this kind of thesis is harder to defend than token event reduction is motivated by the fact that reducing property instances, arguably the most "fine-grained" kind of psychophysical reductionism possible, faces the problem of reducing the problematic entities (qualities) themselves, which seems a harder task than reducing more complex (or "coarse-grained") entities in which these qualities *participate* in some way.

Completing my comments about the individual premises, I point out that premise 1 seems to "carry the most weight on its shoulders". It was perhaps the premise which required the most original insight to devise. It is motivated by a new kind of explanatory exclusion: instead of making recourse to a problem with overdetermination of sufficient causes, it is motivated by the idea that there can be no two different true and *complete* causal stories explaining the same effect. As I will soon mention, I make use of a slightly weakened version of this premise in my final version of the argument: there can be true stories that are not identical, but, in this case, they must be *reducible* to a single one. As I said in Part II, I am assuming a kind of explanatory realism in my discussion (that is, there is something "in the world" that makes explanations true; in this case, the causal descriptions work like

[63] One may be inclined to assume that for a property instance to be reducible to another it just takes something analogous to what can make an occurrence/event be reducible to another in standard Davidsonian fashion (e.g., the token event "being in pain at t" could be the same occurrence as "being in N at t ", where N is the neural-correlate (and thus realiser) of pain). I believe this misses the point, however, and perhaps speaking in terms of "property exemplifications" might help to see why. In the kind of case we are considering, what happens is that a particular phenomenal property (a particular sensation of redness, say) occurs at a given time. The kind of reduction I am considering is whether that particular phenomenal property is reducible to a different, particular physical (or "non-phenomenal", at least) property. So the kind of reduction required by the formulation I intend is between the particular properties themselves, not their occurrences (that is, the event that occurs when they are exemplified). I chose the term "property instance" to mark the difference from property types, to stress I am talking about particular manifestations of properties, and not about the occurrences or events in which they feature. To summarise: the reduction at stake here concerns specific/token properties, not events.

explanations of what ensued, being true or false in terms of whether they accurately describe what happened or not), but I am unsure over whether this assumption is required to motivate P1 (for a discussion of explanatory realism and issues involving explanatory exclusion, see Kim, 1989).

The upshot is that, if a causal description is true and contains no gaps, then, if a given property is not featured in this description nor is in some way reduced to any of the properties that do feature in it (more on this in the next paragraph), then it seems there is no place left for it in that description – it has no gap to explore and exert its influence. It is then dispensable in the sense outlined. Furthermore, if we suppose that causal claims are made true by that in virtue of which the causal link holds (which is a kind of realist reading of the property "quausation" idea), then the result is that a property that is "left out" of a true gapless description is not the true cause of the effect in question – which is precisely the point of premise 3.

I called the above formulation the argument the "version one". This is because it seems a natural first attempt to formulate the kind of argument required, and it was indeed the first version I devised. There is, however, a subtle but important modification that I believe should be made to this initial version, and it is worth discussing it. It was motivated by a possible counterexample to premise 1. Discussions about causal exclusion are usually held in terms of causal sufficiency. There are reasons, however, to suppose that is not a decisive feature of causal interactions. Stephen Yablo, in his (1992), argues that causal sufficiency is not the same thing as causal relevance; ironically, the presence of the first is not sufficient for the latter. To illustrate this point, Yablo asks us to consider an ingenious thought experiment that serves as a counterexample to this usual assumption. Suppose that a pigeon has been trained to peck at objects that are put in front of its face if and only if they are red. Now suppose a particular object of a particular shade of red is placed in front of the pigeon – a crimson lollipop, say. As expected by its Pavlovian conditioning, the bird pecks at it. It is surely true that "being crimson" is sufficient to account for why the pigeon pecked. But it surely is not *relevant*, since an object with any other shade of red would produce the same result. So "being red" seems to be what is truly causally relevant in this case, not "being crimson" or being of any other shade whatsoever. But "being red" is also sufficient for the pecking, so, according to the reasoning behind the original exclusion argument, we could be led to think that saying that "being red" was a cause in addition to "being crimson" would lead to overdetermination, and thus wrongly conclude that "being red", what is *really* relevant, should be causally excluded.

How to respond to the points raised by Yablo? Cases like the pigeon's would surely count as counterexamples to premise 1, since "being red" is not identical to "being crimson" and the fact that we can devise an explanation (or description) solely in terms of shades should not count as a reason to say that "being red" is dispensable. I think he is at least partially right, and I do not have any fully developed response to his position yet,[64] though I will say more about this example in the following sections. Luckily, I can reformulate P1 in a way that no longer clashes with cases like these without having to commit myself to a specific theory about causal relevance. It seems plausible to assume that things are red in virtue of exemplifying a specific shade of red – after all, nothing can exemplify "pure redness", *just* redness. By the same token, it is reasonable to assume that red things cause things in virtue of their redness *through* the properties of being of a given shade. Following this line of thought, an immediate attempt would be to rephrase the conditional clause at the end of P1 to say that "if P is not identical to any property mentioned in the description and P does not cause *through* the instantiation of any of those properties, then P is causally dispensable for this description". But the last conjunct in the antecedent might beg the question against those that, for instance, defend some kind of "supervenient causation" which, recall, can be formulated in these terms: "if P supervenes on Q and Q causes E, then P can be said to cause E". After all, it would be assuming that instances of mental properties are not the kind of thing that causes *via* the instantiation of physical properties, which, though I believe is correct, is not obvious and a bit trickier to argue for. For now, we should look for something less controversial and/or specific.

It seems that cases like "being red" vs "being crimson" are special in the sense that they involve a special relationship between a general type of property and its instances. The appropriate account of these cases must capture this relationship, whatever it is. I believe there is a relatively simple way to do so: though "being red" is neither identical nor (presumably) reducible to, say, "being crimson", it seems that *instances* of "being red" are reducible to instances of shades of red. That is, to be a property instance of red apparently *just is* to be a property instance of specific shade of red, since redness is instantiated by its shades. So it turns out the modification we needed for premise 1 was rather simple: replacing

[64] I do have some inclinations, though. As I mention in the main text, the puzzling feature of cases like these seems to come from the special relationship between general properties (or universals) and their instances, especially in the case relating colours and specific shades. What I take as the lesson from these cases is that "in virtue of" or "because" admits of further discriminations. We could say, then, that "the pigeon pecked at the lollipop because it was red; more specifically, it pecked at the lollipop because it was crimson". This does not mean that pointing to "being crimson" as the cause is wrong; it is just a more *specific* description of the cause (or, perhaps, a more specific "reason why" the effect happened).

"identical" with "reducible to". I will return to this topic below, in my reply to Menzies & List's proposal to overcome the exclusion problem.

At last, we come to my final formulation of the Quausal Exclusion Argument:

Quausal Exclusion Argument (Final Version):

P1. If there is a true gapless causal description for the occurrence of a given effect that makes no recourse to property P, then, if instances of P are not reducible to instances of any of the properties mentioned as causes in this description, then P is causally dispensable for this description.

P2. For every physical effect the occurrence of which we might attribute to a mental cause, there is a true gapless description of the causal transaction leading to that effect that is not cashed out in terms of mental (phenomenal) properties.

P3. If P is dispensable in a true and gapless description of the causal transaction leading to that effect, then it is not in virtue of P that the effect in question occurs, i.e., P is not a cause of that effect.

P4. Instances of mental (phenomenal) properties are not reducible to instances of physical properties (Hypothesis for Conditional Proof)

Therefore,

C. If instances of mental properties are not reducible to instances of physical properties, then it is not in virtue of them that the physical effects occur.

The last comment I would like to make about the Quausal Exclusion Argument in this first section is to point out that this argument has a conditional thesis as its conclusion: roughly put, if non-reductionism about phenomenal property instances is true, then quausal epiphenomenalism is true. This is the kind of conclusion that causal exclusion arguments traditionally are built to produce (since, recall, they are usually proposed as a *reductio* of non-reductionism). It will follow, then, that if the argument is sound, then the only way to escape quausal epiphenomenalism that the considerations in the argument allow is for a specific kind of reductionist thesis to be true. Namely, a thesis where instances of mental (phenomenal) properties are reducible to instances of physical properties. In other, more precise, words: that the specific phenomenal properties featured in any relevant step of the causal chain be ontologically reduced in the appropriate way to specific physical properties. Establishing this result is already a substantial contribution this argument can bring to the

discussion, as it would force those that seek to avoid epiphenomenalism (most philosophers) to commit to a very narrow reductionist view, one that is not the most popular and that most non-reductionists would not be inclined to adopt. This result would already be enough to show that the challenged posed by the causal exclusion problem is more serious than is usually assumed, as it would require a specific type of theory that is unpopular amongst parties more on the non-reductionist side to evade falling into quausal epiphenomenalism. I believe, however, that the threat posed by the exclusion problem and the exclusion argument is even more serious and the results even more devastating. To show why, I will offer my generalisation of the problem in the form of a dilemma (and a trilemma). Before that, though, there is one salient line of objection to my argument that I would like to discuss.

2. A possible objection: causation as "making a difference"

Naturally, as is usually the case with arguments as controversial as the exclusion argument, there are many ways one could attempt to object to the argument, especially considering how one could object to details in its formulation. Exploring various ways the argument could be resisted is, then, the obvious task to pursue in assessing the quality of the argument. Given time and space constraints, I will only address here one major line of objection, one that is motivated by a counterfactual account of causation as "making a difference" developed by Peter Menzies and Christopher List. There are good reasons why I decided to address this particular objection in detriment of others. The first is that I consider it the most ingenious, challenging, and (perhaps because of the former qualities) promising of the lines of objection I have come across. The second reason is that it shares features in common with other types of theories of causation where higher-order entities can have causal efficacy (theories of what is usually called macro or downward causation). And the third is that, being a counterfactual type of theory, it will give me occasion to raise more general objections to this broader family of views.

I will address here the proposal put forth by Peter Menzies in his (2013), though, as he notes in the article, the general idea was developed together with Cristopher List in previous works (List and Menzies 2009; Menzies and List 2010). For this reason, I will often refer to this account (as I have been doing before) as the Menzies-List theory or account of causation as making a difference. Menzies also offers in the same paper an interesting discussion about the best way to formulate the exclusion argument, including some problems with Kim's past formulations. I will not reproduce this discussion here, focusing instead on

the positive account therein advanced by Menzies and how it is supposed to solve the exclusion problem.

Menzies introduces his account as an account of "causal relevance" – that is, of what makes a given factor relevant in causing or not given effects. It is not a stretch, though, to understand his proposal as an account of *causation* more broadly. This becomes especially clear when we notice that, to say that a given factor is causally relevant in producing some effect, this seems to imply that that factor plays some role in causing it. In other words, if factor F is causally relevant for the occurrence of E, then F is at least a partial cause of E[65]. And is natural to assume that the converse holds: if a factor causes some effect, then it is causally relevant for the occurrence of that effect. So we can understand Menzies and List's account of "being causally relevant" as an account of "being a cause".

Menzies appeals to an intuitive way of making sense of causal relevance in terms of "making a difference". The idea here is that a given factor would be causally relevant for a given effect if it occurring or not makes a difference to the effect occurring or not. That is, if its presence leads to the effect occurring and its absence (that is, it not occurring) leads to the effect not occurring. If it plays this role, we can say that factor *makes a difference* to whether the given event will happen as an effect or not. What Menzies offers as a more formal way of capturing this result is the rather natural interpretation of what was just said in terms of counterfactuals. His proposal takes this form:

Truth conditions for causal relevance (or making a difference): The state S1 makes a difference to the state S2 in the actual world just in case (i) if in any relevantly similar possible situation S1 holds, S2 also holds; and (ii) if in any relevantly similar situation world S1 does not hold, S2 does not hold. (Menzies, 2013, p. 73)

An initial thing that is interesting to note about this formulation is that it appeals to situations/possible worlds that are *not actual* to assess whether the states in the actual world are indeed causally connected. This is a feature shared by all theories in the family of theories that is usually called "counterfactual theories of causation", that trace back at least to David

[65] More precisely, F would be at least a factor of a partial cause of E, that is, an element of the cause, something that features in that cause. In the case of Menzies' discussion, just as has been the case so far in this work, the factors under discussion are properties (or property instances), and causation is also assumed to relate events/occurrences. Going with my property-causation reading, I will treat a property featuring in a cause that is relevant in producing the effect in question as a cause of that effect. If one would like to keep attribution of causes to events, and prefers to talk of constituents of events (such as token properties) as "causally relevant" factors instead of literal causes, then one could treat my exposition merely as short for this expanded formulation, where the properties are causal factors, but not, strictly speaking, causes.

Lewis' *Counterfactuals* (1973)[66]. That is, they turn to what would happen in a selected class of possible worlds to determine whether some actual factor causes or not a given effect. The motivation for this makes sense. To assess whether a factor is causally relevant or not, it seems natural to consider what would happen if the factor were not present. Better yet (so that we are assessing the relevance of that factor *only*): it seems we should consider what would happen if everything else remained unchanged and only that factor was altered. Dealing with these considerations through the apparatus of possible worlds and counterfactual conditions is just a natural move to make philosophically. This is what Menzies does when offering this slightly modified restatement of the truth conditions for causal relevance:

Truth conditions for causal relevance (making a difference): The S1 makes a difference to S2 in the actual world if and only if it is true in the actual world that (i) S1 holds □→ S2 holds; and (ii) S1 doesn't hold □→ S2 doesn't hold. (Menzies, 2013, p. 74)

Where "□→" is interpreted via a notion of closest possible worlds to a given world *w*, which is a subset of all the possible worlds accessible to *w* (given an accessibility relation as is usually defined in systems of modal logic). In summary, the idea is that the less "steps" it takes to get from world z from world *w* via the accessibility relation, the closest z is from *w*. Given this framework, Menzies' account interprets the "□→" operator in this way: "P □→ Q is true in world w if and only if Q is true in all the closest P-worlds to w." (*ibid.*).However, there is no precise definition of what counts as "relevantly close" and how close a given world has to be from the actual world in order for it to be a member of the set of "the closest" possible worlds. I believe this is intentional, as Menzies' idea is that what these requisites would be would probably vary from case to case.

Notice that the point about assessing whether the same (or similar) effect would happen were *everything else* to remain the same and only the factor under evaluation change helps explain another detail in this formulation. The fact that only "relevantly similar" possible worlds are considered has to do with an attempt to construe this *ceteris paribus* condition. We do not have to consider what would happen in *every* logically, metaphysically, or even physically possible world. We do not even have to consider what would happen in every possible world where these events (or states, objects, etc.) or their relevant counterparts in these worlds occur. We want to restrict our domain of consideration only to worlds that

[66] Lewis's counterfactual theory is importantly different from Menzies & List's, as it focuses on something akin to clause (ii) of the formulation.

are, as Menzies rightly put it, relevantly similar to the actual world (or to the scenario under consideration, if it is not actual). That is, to assess whether a specific spark in a fuse box caused a fire in a warehouse, we do not have to consider whether a similar spark would have led to a similar effect in a world where gravity was 100 times stronger, or where our moon was made of cheese, or one where there was no air in the atmosphere. Rather, we consider scenarios that are "close" to our world and ask whether the fire would have occurred given the occurrence (or absence) of that spark. Clause (i) certifies that the relationship of between alleged cause and effect is robust enough. Similar scenarios where the proposed cause is present would lead to similar effects. Clause (ii), on its turn, guarantees that the factor is not redundant, that it indeed matters in making the effect happen. Obviously, what scenarios count as "relevantly similar" or "close" to the target scenario is vague, which is a point opponents of proposals along these lines could press. In any case, I believe the appeal behind restricting the class of possible scenarios that are relevant is relatively clear.

I made clear in Part II that I would not explore the more detailed, counterfactualist account that Horgan put forward in his original paper to capture the *desiderata* of mental quausation. I would just like to present his final version of this account to briefly comment on the striking similarities it bears to the Menzies & List account. Here it is:

> If (i) event c causes event e, (ii) c and e respectively instantiate properties F and G, (iii) F and G are logically and metaphysically independent, and (iv) the causal transaction between c and e does not involve preemption, overdetermination, or the like, then the fact that c and e instantiate F and G, respectively, is explanatorily relevant to the fact that c causes e iff the following *Relevance Condition* is satisfied: (R) For any world w in P[c,e], if c* is the event in w that is pertinently similar to c of W, then (i) if c* instantiates F in w, then c* causes (in w) an event e* which both instantiates G (in w) and is pertinently similar to the W-event e; and (ii) if c* does not instantiate F in w, then c* does not cause (in w) an event which is pertinently similar to the W-event e. (Horgan, 1989, pp.58-59)

Where P[c,e] is defined as a set of "Pertinently Similar Worlds" (PSW) relative to events c and e, such that "[e]ach PSW contains a situation pertinently similar to – although perhaps somewhat different from – the situation in which c caused e in the actual world W." (Horgan, 1989, p. 58). The many similarities are clear, so much so that I believe that the main objections I will raised against Menzies & List's account will apply, *mutatis mutandis*, to Horgan's original proposal (and to any counterfactualist proposal, for that matter).

The conditions for "making a difference" defended by Menzies & List surely deliver results that seem correct for many paradigmatic cases of causation. For example, let us consider again the case of the cantaloupe being put on the scale. If we apply Menzies & List's

proposal to this case, we get the right verdict about the specific mass the cantaloupe has been causally relevant to produce the specific movement on the scale. In any relevantly similar scenario where the cantaloupe has that amount of mass and is put on the scale, the scale moves to indicate the specific amount in kg (say). If the cantaloupe put on the scale had a relevantly *different* mass (say, being a few kilograms heavier), the scale would have moved in a different way, indicating another weight tag (or perhaps not moved at all). So the specific mass the cantaloupe has passes the test for being causally relevant for the movement observed in the scale: in every relevant scenario where it has that mass, the scale moves in the observed way (so (i) is satisfied), and in every scenario where no melon with that exact mass is put on the scale the scale does not move in that way (clause (ii) is satisfied). We could run similar "tests" for most (if not all) the examples discussed previously in this work and arrive at the same kind of correct verdict.

Something remarkable happens, however, when we apply this test to a case like Yablo's Pavlovian pigeon, which is precisely the example Menzies chooses to illustrate his central point. Once again, we will notice a causal difference between "being red" and "being crimson". If we apply it to "being red", the conditions are clearly satisfied. Anything that is put in front of that pigeon that happens to be red would cause it to peck, and nothing that is not red would cause the same reaction. So "being red" satisfies clauses (i) and (ii) and, therefore, is declared causally relevant for the pecking according to Menzies & List's account. What about "being crimson" (or any other specific shade of red)? Well, it certainly satisfies condition (i), as anything that is crimson (or any other shade of red) that is put in front of the pigeon will cause it to peck. As we have seen, this follows from the fact that "being crimson" is sufficient for causing the pecking. This time, however, condition (ii) will not be satisfied; the thing presented to the pigeon could not be crimson and still lead to the expected pecking, provided it was of any shade of red. That is, "being crimson" is unnecessary for the pecking. Since the pecking could occur anyway without the lollipop being crimson, "being crimson" is deemed not causally relevant for the pecking. So far, this result seems similar to Yablo's original remarks that I already discussed above. Notice, however, that the conclusion seems more striking this time. Recall that, in my first discussion of the example, I mentioned that, in a sense, we could say that "being crimson" caused the pecking. The upshot of the example would just be that we would have to reconcile this fact (that "being crimson", the causally sufficient factor, is in a sense a cause) with the causal relevance of "being red". This is not a result Menzies would concede now. His proposal is that only "being red" is causally relevant and, therefore, can truthfully be claimed as the cause in this case. "Being crimson"

is not causally relevant and, therefore, is completely excluded of any causal influence. It is simply not the factor responsible for making the effect happen. It is not a cause even "in a sense". This, in effect, puts the original conclusion we could arrive if we followed causal sufficiency as a guide on its head: the causally sufficient factor is not a cause. According to Menzies, cases like these would show that causal factors must be "specific enough" for a given effect (Menzies, 2013, p. 73). Contrary to my first assessment of the example, his take is that being *too* specific a factor can be a problem, one that might detract from the postulated causal influence of a given factor[67].

A crucial factor Menzies wanted to draw attention to, however, was not exactly the last surprising result, but an ontological "moral of the story". What the pigeon case shows is that more ontologically fundamental entities might not have causal influence in scenarios where less fundamental entities *do* exert causal influence. In other terms, something A of a higher ontological order than B and that is ontologically dependent on/determined by B could cause something that B does not cause. This could mean that "macro" properties, like emergent properties, could exert causal influence when their underlying ontological "basis" is causally inefficacious.

It is not hard to see where this is going. Menzies treats some of the traditionally problematic cases of mental causation as analogous to the pigeon case, being examples of this type of case where something ontologically dependent does all the causal work. The central move to establish this bridge between the cases is his assumption that mental states (including, therefore, the problematic phenomenal states) are multiply realisable. Menzies considers what he takes to be the plausible possibility that the mental states usually involved in the exclusion problem (let us focus, as we have been doing, on phenomenal states) are multiply realisable. That is, they are realised[68] by different types of physical states. Let us

[67] Though it is an interesting point with some undeniable appeal, I believe Menzies is not correct about this. I think there is no way of reconciling his remarks with a causal realist view. They seem to assume that the way we *describe* the events matter, and I believe that if we focus instead on what actually transpired in the world when the alleged causal transaction occurred, we will in most of these cases be able to trace back the causes to the more specific factors he would exclude. So a more promising route for a view like his seems to me to be one that attributes causal relevance to higher order properties, but that does not necessarily exclude more specific factors from having a causal role.

[68] The notion of realisation is tricky to characterise. The intuitive idea is one that is usually explained by an appeal to other notions like that of "implementation", the usual example in recent use being the relationship between computer software and hardware. The same software can be implemented/realised in different pieces of hardware. This relation that software bear to their hardware is generally taken to be a good example of the idea of realisation. "Realisation" is sometimes defined in causal terms, so that A is realised by B iff the "causal role" of A (that is, the effects attributed to it) is caused by B. I believe this more restricted, causal notion of realisation would presuppose some things that are under dispute here, so a more neutral notion of realisation should be preferred. For our purposes, it will suffice to understand the realisation relation as an ontological dependence relation holding between a more higher-level entity, the one being realised, and some

consider as an example a case where we attribute the anger felt by someone (Tim) as a cause of their hitting someone. Let us frame it so that the mental state M is "feeling angry" [69]. Now, suppose M is realised in Tim, our character, in the actual world by physical state P (presumably, some physical state corresponding to a specific pattern of neuronal activation in Tim's brain). Now, Menzies' assumption about the multiple realisability of M is that M could have been realised by other physical states in worlds that are relevantly similar to the actual world. In those worlds, presumably the same effect (the hitting) would ensue given the same (or relevantly similar) conditions. This would make this a phenomenon of mental causation that is realisation-insensitive, as Menzies calls it (Menzies, 2013, p. 80). This means that M can cause whatever it is that it manages to cause irrespective of the specific physical realisers it has in each world. That is, as long as "feeling angry" is being realised, the hitting will occur (given the conditions). It does not matter if M is being realised by P or a different physical state, P*. As it turns out, then, when we apply the conditions for causal relevance to these states, we find that M satisfies them – whenever M is realised, the hitting occurs, and (presumably) whenever M is not realised, the hitting does not occur. The crucial point now is that P does *not* pass the test. In specific, it fails condition (ii). There are worlds relevantly close to the actual world where P does not occur and yet the hitting follows – namely, those worlds where M is realised by something else, such as P*. And this result applies to each and every physical state that might realise M. None of them, specifically, is relevant in producing the effect. So it follows that, in this case, M makes a difference for the effect, and is therefore causally relevant/a cause of it, but the specific physical state underlying it does not. Notice that, like with his verdict about the pigeon case, Menzies' conclusion is not that the claim that multiply-realisable mental states can be causally relevant is compatible with the claim that the underlying physical states are, a result called compatibilism in the literature and defended by, among others, Syndey Shoemaker (2007) (Menzies, 2013, pp. 81-82). Rather, the surprising result of Menzies & List's account is that only the mental states/properties are causally relevant, excluding any causal relevance of the physical realisers. The solution that Menzies & List offer to the exclusion problem is then that mental items of an appropriate kind (those that are multiply realisable and realisation insensitive) can be causally efficacious even though they are not ontologically reducible to any physical states, and even though their

other more fundamental entity, the realiser, in such a way that entities of the type of the realised thing exist/occur only by being "implemented" in some way by realisers of that lower type. They are ontologically "parasitic" on their realisers in some sense.

[69] Keeping closer to Menzies' original terms, I will discuss this point in terms of mental states. We can easily transpose the discussion to properties and property instances to approximate them to my preferred way of framing the causal transactions.

underlying physical states are causally sufficient for the occurrence of any effects attributed to them! This striking result holds because, on their view, it is the *mental* items, not the physical ones, that make a difference in leading to the effects. We could say that it is in virtue of the *mental* items that the effects occur, so we could, following their view, claim a form of mental quausation. If their proposal is correct, then my quausal exclusion argument is not sound. It would show that P2 of my argument is false (since the causal story told solely in terms of the physical properties would not, strictly speaking, be true on their view, since they do not capture the factors that *do* make a difference), and, perhaps, that P1 is also false.

Now, naturally, the above exposition is a simplification of Menzies & List's views on the matter. Many details and qualifications they carefully made were left out here. I believe, however, that I managed to capture in a faithful manner the core ideas behind their conclusion. As I mentioned before, I consider this proposal very ingenious and worthy of discussion, and I do not have a fully fleshed out response to all the points it raises. I believe, however, that there are good reasons to believe this approach is not successful in solving the exclusion problem and in objecting to my quausal exclusion argument. I will offer these reasons in the following sections, starting with general remarks about the adequacy of the truth conditions for causal relevance and then addressing the specific application of this account to mental causation.

3. Why "making a difference" does not solve the exclusion problem

My first line of objection to the account of making a difference defended by Menzies & List consists of offering counterexamples to their proposed truth conditions for causal relevance. First, I will show that there are examples of factors that are clearly not causally related but that would come out as such according to their analysis. That is, their conditions are too inclusive to be a proper account of causal relevance, letting in unwanted results. This is what I focus on in section 3.1. Secondly, in section 3.2, I will show that there are examples that seem to clearly involve causation that are nonetheless excluded from their account. This means it is also too restrictive, as it does not capture all cases of causation.

3.1 Non-causal connections that satisfy Menzies & List's criteria

I must start by admitting that I am generally not very good with coming up with good illustrative examples. Whenever I try fishing for examples, the first ones that come to mind are usually quite odd ones that provide none of the appeal that examples involving more familiar elements and circumstances provide. I will try my best to illustrate the types of

counterexamples I outline here, and I believe some of them will turn out to seem like good examples. As a safety net, I will offer the overall *form* these counterexamples would take, and hopefully that on its own might be enough to get the point across. Perhaps the reader will be able to come up with their own, better examples that fit these forms.

The class of counterexamples I intend to cover in this section are those containing elements that would satisfy Menzies & List's conditions for causal relevance, but that seem to be clearly *not* causally connected in the slightest. There are two types of such counterexamples that I have identified:

I. b and c are constituents (the best example would be (proper) parts) of a. If a occurs (or exists), b and c could both satisfy the conditions of causal relevance laid out by Menzies & List, yet it can be that only of them is the cause of an effect e[70].

The idea here is that two factors might be counterfactually connect such that one occurs/exists iff the other does. In this case, they are both also counterfactually connected to a third factor (a, in my scheme above), on which they might depend ontologically. Yet they might differ in their causal influence: one factor, but not the other, causes a given effect.

Here is a possible example, involving diseases, symptoms, and their causes. Suppose there is a rare disease that involves two peculiar symptoms (that is, symptoms not caused by any other occurring disease). Let us suppose it is a rare type of migraine, and the symptoms in question are a special type of visual disturbance (what is usually called an "aura") and some unusual pattern of blood flow in certain regions of the brain. Suppose they both occur due to a given factor F that causes the condition. Now, suppose the blood pressure symptom, but not the aura, is responsible for generating another symptom, like a peculiar type of tinnitus. Given the peculiarity of the conditions and the symptoms, it seems that both the aura and the pattern of blood flow would be related in a way that would satisfy the conditions for causal relevance in producing the tinnitus. If the aura did not occur, the tinnitus would not occur. Whenever that kind of aura occurs, the tinnitus occurs. So it seems the aura would count as causally relevant for the occurrence of the tinnitus, and thus be a cause of it according to the view of "making a difference". Yet, as the example assumes, it is the pattern of blood flow increase, not the aura, that is causally responsible for the tinnitus. The aura is simply a strongly correlated symptom.

[70] Or, perhaps, that neither of them are. Think of factors that are strongly correlated with a cause, but that are not those that are causally responsible for a given effect that cause produces.

Here is another, more outlandish example. Suppose that, on a faraway planet, there is a type of nut with a shell so hard that only one thing in that planet can crack it. The thing in question is a drill-like structure in the mouth of a type of creature endemic to that planet, which we can call a Drilly. Suppose that, when performing the strenuous operation of drilling through the hard shell of the nut, some structures in the Drilly's body make a peculiar accompanying movement (say, their peculiar Drilly horns vibrate or shake in a given way). The way I am envisioning it, this bodily motion is something only Drillies do, and that, given the biomechanics of the Drillies' bodies, is something that always happens when they drill the nut. It seems that performing that specific bodily motion and the actual drilling would both satisfy conditions (i) and (ii) of the Menzies & List account of causal relevance for producing the effect, yet only the drilling would actually be a cause of the effect of cracking the nut.[71]

Careful readers might have identified a glaring problem with the conditions for causal relevance provided by Menzies & List: necessarily cooccurring factors would both satisfy conditions (i) and (ii) related to each other and to themselves. As a result, they would come out as being causally relevant for/causes of each other and of themselves. That is clearly preposterous, as the causal relation is supposed to be both asymmetric and irreflexive. There is an obvious and charitable way to remove this issue, however, which is to suppose a temporal asymmetry: factors that are to enter as causes must occur before the effects. This solves these immediate absurdities, but I believe there are examples of factors that have this temporal order and are related in the appropriate way and yet are not causally related[72]. These constitute the second type of counterexample I identified:

II. b and c are both effects (to make it even more dramatic, necessary effects) of a, and, for some reason, b always occurs before c occurs. Cases like these would satisfy my amended version of Menzies & List's conditions and have as a result that b and c are causally related (with b being the cause of c), while, in fact, they are not causally connected at all (apart from the fact that they are both effects of a same cause).

I believe this class of counterexamples is more straightforward. Nonetheless, I have what I believe is the best example of this kind of counterexample to help illustrate. Consider the relationship between the electric discharge we identify lightning with and the sound

[71] I believe we can also modify an example like Kim's example (Kim, 1990, p. 15) involving the relationship between I.Q. and dexterity to provide a counterexample along the lines of the ones I presented.
[72] The examples of the migraine and Drilly above could be amended so that the factors have this temporal asymmetry built in.

(thunder) and the flash it emits[73] . The electric discharge would be "a" in this case, with the flash being "b" and the thunder being "c". The point is that, given the difference in speed between light and sound, the flash always occurs before the thunder occurs, and it is such that one occurs iff the other occurs[74]. This would make the flash of a lightning the cause of its thunder under Menzies's account. That is, in every possible world close to ours, the conditions are met for the flash and the lightning. It results, then, that in *our* world (the world we are conducting the analysis in), instances of flashes of lightings would be the cause of their following instances of thunder, which is clearly false.

It is worth noting that, in Horgan's own counterfactual account of mental quausation, he was careful enough to require that the properties entering the relation must be "logically and metaphysically independent" (Horgan, 1989, p. 56). I believe his motivation was to exclude potential cases similar to the ones I presented. I do not have a clear idea of what "metaphysically independent" properties might be and whether we should indeed exclude those from an account of quausation[75]. In any case, I believe the properties I outlined are surely logically independent, and perhaps also metaphysically so.

I hope the two types of counterexamples and the specific examples I gave were persuasive in conveying what I believe is the correct result that the analysis of causal relevance provided by Menzies & List gives the wrong verdict about non-causal connections. This is already a serious flaw to their view, but I believe there are also good reasons to think that it also leaves out factors that are clearly causally related.

---

[73] I am not referring here to *image* of a flash that is seen or the *sound heard* of a thunder; that is, to those *mental* items that would be non-occurring to beings who are blind and deaf. I am referring to the flash and thunder as the physical processes of an emission of light and of soundwaves, respectively, regardless of whether they are seen or heard.

[74] Someone could object to this claim, claiming that there might be physically possible (that is, not a mere logical or metaphysical possibility) scenarios in which the flash from the discharge occurs but is not followed by thunder, perhaps because of some property of the medium in which the discharge occurs that is conductive to light, but not to sound waves. While I do not know whether such scenarios are indeed possible (notice that cases where a vacuum is a medium would probably not work, as the phenomenon we are associating with the electric discharge has to do with the presence of gases), we could, in any case, restrict the characterization of the phenomenon so as to only apply to lightning in our world given the observation about the indexical nature of Menzies' conditions. Deriving the odd result that, in our planet (and others like it), the flash of a lightning would count as the cause of its thunder under Menzies' conditions would be sufficiently problematic for his account.

[75] One natural way I have of interpreting this idea is that two properties are metaphysically dependent if they always occur together in every possible world in which they exist, and they are independent if they are not dependent. If that is so, I do not see a reason to suppose that metaphysically dependent properties should not be causally connected. Think of the possibility that event C and event E always occur together. In that sense, they would be metaphysically dependent. Could they not be causally related? I see no reason why not. This is probably not what Horgan had in mind. But then I don't know what was.

3.2 Causally related factors that the analysis does not capture

When it comes to examples of factors that would be clearly causally related but that would not count as such according to the "making a difference" proposal, there is only one class of examples that occurred to me. This time, they are nothing new. They involve causal "back-ups" or what is usually called in the literature "pre-emptors.". The challenge that causal back-ups pose to counterfactual theories of causation has long been acknowledged and much has been written in response. I mention this to clarify I of course do not assume I am providing anything surprising in raising these counterexamples, and I would also not be able to do justice to the debate over this issue in this section. I believe it is a relevant issue to bring up, though, as nothing in Menzies & List's account offers a direct way of dealing with these examples, and they surely cannot be simply shrugged-off as being fringe cases.

The causal back-ups I have in mind follow this structure: if a occurs, it causes c; if a were not to occur, however, b would occur and cause c. As an illustration, suppose I make a contraption to hit a cue ball to send a given billiard ball to the pocket. Suppose, for simplicity, that there is only one (approximate) way the ball could be hit and pocketed (it is in an extremely difficult angle, say). Suppose, further, that I am very good and consistent with my shots. If I hit the ball within 30 seconds, the ball will, naturally, be pocketed. If I do not, a back-up machine will be activated (via a timer) and hit the ball in (again, approximately, to any degree we may find interesting) the same way I would have done. Suppose, then, that I do indeed take my shot and pocket the ball. Clearly, my shot caused the ball to be pocketed. Since the ball would be pocketed in either case (in every close possible world, at least), given the back-up device, my hitting the ball would not satisfy Menzies & List's condition that, were the event not to occur, the alleged effect would not have occurred as well. The same would hold for the alternative scenario in which the machine is the one that pockets the ball. My shot would not, therefore, count as a cause of the ball being pocketed in this case according to Menzies & List's conditions. But it surely is. Hence another inadequacy of the proposal.

Again, it is worth noting that Horgan also explicitly mentioned the absence of causal pre-emptors as a further assumption made in his detailed account of quausation (Horgan, 1989, p. 56). This is a common move, with "absent pre-emptors" becoming a frequently used qualification in the literature of causation. As I mentioned above, I do not see how we can so easily ignore pre-emptors. They are frequent enough in the real world to not count as some odd exception, and, what I believe is the most important part, they are clearly examples

of causation, whether one of the option causes the effect or another. They must be taken seriously and handled satisfactorily by any theory of causation. Menzies & List's proposal, as it stands, has no clear indication of how to accomplish this and avoid the obvious counterexamples.

If the examples I raised are well-formed and the conclusions I indicated we should extract from them are correct, then the account of causation as "making a difference" cashed out in terms of the causal relevance requirement put forth by Menzies & List fails. It counts as causally related factors that are not and does not allow factors that clearly are causally connected. These general remarks are, on their own, enough reason not to accept their proposal. It might seem weird, however, to leave the case against their view at that and not address the specific way they apply it to the case of mental causation, deriving the interesting result that higher-order entities might turn out to "steal" the causal powers of their determiners. I do have some special replies to make about that application, though I believe matters are much more difficult to express in this case. There are also important points where I am still unsure of what to think exactly, so a less straightforward reply than the above can be expected to this point. Still, I believe I have relevant things to point out. This is what I do in the next section.

3.3 The issue with "making a difference" only as mental

As we have seen, the application of Menzies & List's theory of causation as "making a difference" to the mental causation has as a result that, at least as long as mental items are multiply realisable and the causal influence associated with them is realisation independent, we could maintain that these mental items (and not their causally sufficient physical realisers) are the causes of their attributed effects. And we might have good reason to think these mental items *are* multiply realisable in just this sense, as Menzies seems to suggest in his (2013). I have shown in the previous two subsections that there are good reasons to think that the analysis of casual sufficiency their theory employs has serious problems and should not be adopted. Nevertheless, we could ask whether, leaving these general objections to the criteria for "making a difference" aside, there are any other problems with the multiple realisation-based solution they offer to the exclusion problem. I believe there are.

A first thing that I believe is important to consider concerns the issue of the unclear conditions for a possible world/scenario/situation to count as relevantly similar or close to the one under consideration. In their defence, as I indicated my exposition the proposal, it seems there will always be some vagueness or indeterminacy involved in addressing whether

some possible worlds are "close enough" to another one, and that the required measure for this might vary depending on context. This on its own would not mean the notion is meaningless or useless, and that we cannot extract informative conclusions about causal phenomena without a precise definition of how similar is similar enough (or how close is close enough). Nevertheless, a weakness of this recognized imprecise character of these notions is that this imprecision might become relevant in reaching a verdict about some less clear-cut cases. I believe this issue arises when considering whether it is true that the relevant mental items realisation insensitive in the sense claimed, that is, whether it is true that in all of the "closest" possible worlds, there are different physical properties that could realise the same mental item and produce the same effects. Notice that there are two important claims involved here: 1. That the mental items in question are multiply realisable in the closest possible worlds and 2. That in these worlds the same effects happen. I think both of these claims, perhaps especially 1, are not at all obvious. It is simply not obvious (if even fully meaningful) to say that a given phenomenal property P could be realized by a different physical state in worlds that are "sufficiently close". Assuming that these worlds *are* close enough seems to beg the question in favour of the view. It is something that requires some further argumentation.

The above is a minor, albeit important issue the supporter of the "making a difference" approach must address. The deeper problem I have with this proposal, however, has to do with the type of ontological relation that is supposed to hold between the relevant entities. I think there is something special behind Yablo's pigeon example that explains why it seems to be so compelling. Specifically, I believe there is something special about the ontological relation that holds between colours and their specific shades. Bernestein & Wilson briefly comment on this seemingly unique relation of "increasing specificity" (Berstein & Wilson, 2016, p. 323)., referring to it as a determinable/determinate relation. I think this special relation (I am not exactly sure how to call it) is not easily compared to other relations holding between alleged property types and their instances; specifically, I do find it all clear that it is in any way analogous to the relationship that is supposed to hold between phenomenal properties (or their types) and physical properties they might depend on. That is, there seems to be a relevant difference between the way redness is realized in specific instances of red shades vs the way phenomenal properties would be realised by physical properties. The relationship seems to be something akin to constitution in the first case, but not in the other. I find this point I am trying to make especially hard to put into words, but I feel tempted to say that being crimson is *a way of being red* in a sense in which physical

properties are not ways of being a certain phenomenal property; they seem to be of radically different types. One alternative way to put it that might be helpful is in terms that seem epistemic. There seems to be something peculiar about the relationship between redness and specific shades in that something *reveals itself* (for a lack of better words) as red *through* its specific shades. Conversely, something being crimson *reveals* redness through it. One can immediately see that something is red by seeing that it is crimson, and one can only see that something is red through seeing that something is of a specific shade of red (that is, no one witnesses pure redness). This relation would surely never be present in the case of phenomenal properties and their underlying physical properties, no matter how radical the reductionist theory one believes is correct. Now, this peculiarity may or may not turn out to be relevant, but I believe it at least helps to show that there is something peculiar about the ontological relation in the case of colours and shades, one that is not replicated by other entities. This might help show why we might be so easily persuaded of the causal efficacy of redness in the pigeon case, as, given this special relationship, it might be that something only causes something in virtue of being red *through* (or by) being of a specific shade of red. And what would naturally warrant this transferral of causal efficacy from instances of redness to instances of shades would be this *je ne sais quoi* relation that shades bear to their colour types. The upshot is this: we might be willing and happy to try to accommodate the result that redness plays a causal role for the pecking in the pigeon case, as I am and tried to do in my revised formulation of the argument. I think an instance of redness has the same "causal powers" as an instance of a shade of red, since an instance of redness can only exist as an instance of a shade of red. If this is explained by some peculiar feature of the relationship between colours and shades, however, we should not easily draw the parallel from a case like the pigeon's to cases involving other relations, such as the relation that is supposed to hold between phenomenal properties and their realisers. Hence, there is no forcing the conclusion that, if we concede redness must be awarded causal relevance in the pigeon case, we must do the same in the phenomenal property/physical realiser case.

My final remark about the specific application of "making a difference" to mental causation is that it seems odd to say that mental type M could be the cause of something if, in every instance of it, the physical property instances seem to do all the causal work. Recall that Menzies never denied (and, in fact, agreed) that the physical factors remained causally sufficient in producing the effects in every case under the multiple realisation hypothesis. What I now want to draw attention to is that he never offered any reason to make us doubt that P2 of my Quausal Exclusion Argument is false. That is, that there is always a true and

complete causal description we could give of the relevant causal transactions that is given solely in terms of physical property instances. He would surely not frame the issue in this way, but he would be committed at least to denying that this sort of description could be true. However, his major considerations were not made in an effort so show some kind of glaring hole or falsity in this type of causal story, but rather to show that causal sufficiency was not the relevant notion we should focus on. So he has offered us no direct reason to believe the following way of putting is false: the causal chain could be clearly traced back to the physical properties instances in a complete fashion. Moreover, the same could, arguably, not be done in tracing it to the mental property. If this result remains, as I believe it does, then we get the impression that, in every case where the alleged mental cause leads to the effects, the underlying physical properties seem enough to account for what happened. If this is so, then it is odd to maintain that the mental property exerts causal influence, and even more so to maintain that only the mental property, and not any physical property, is causally relevant.

Notice further that it is not part of Menzies & List's proposal that the mental properties in some way cause *via* the physical properties that realise them. Instead, their point amounts to an analysis of what factors are counterfactually connected to what. This suggest to me that their approach is closer to an account of causal explanation than one of causation[76]. That is, it is about what we can interpret as being relevant in explaining a causal transaction. While this is by no means a problem of the view, I believe it might be conflating the issues I tried to keep separate (that of explanation and causation), and might indicate that their view deviates from the causal realist assumption I am following. This way of interpreting what is in play with their proposal neatly connects to what I believe is the main takeaway from the position they defended. Instead of their conclusion that the mental factors are the only causes, what I interpret is that, when it comes to causation strictly speaking, being crimson *is* a cause of the pecking in the pigeon case, and the specific physical properties associated with the neural realisers *are* causes of the effects we attribute to the mental state. When it comes to *explaining* these causal transactions (that is, of making them informative, intelligible), appeal to more general considerations might make more sense. We might learn something informative when we say that "P_1 causes e, and P_2, P_3,..Pn would also cause e under similar circumstances. What they all have in common is that they are somehow related

[76] I thank Minseok Kim for this observation.

(realise, determine, etc.) to property M". When it comes to causes, however, they are strictly speaking not causal factors.

What I believe the actual "moral of the story" of Menzies & List's remarks is, then, is something like this: P_1 indeed causes (is causally relevant for) the given effect E. We could learn that any other physical property that happened to be ontologically connected to mental property M in the appropriate way (say, by realizing it) would also produce the same effect, and that no other type of physical property would. This is something informative that is learnt about the phenomenon, but would not on its own show that M is causally responsible for anything. We just gain additional explanatory knowledge.

This concludes my examination of Menzies & List's response to the problem of exclusion. As I stated before, I believe it is very ingenious and that there are important lessons to be taken from it, even though they might not be what the authors intend. There are multiple points on which I simply do not have a clear stance yet, and others where I wish my inclinations could be expressed in a clearer manner. Nonetheless, I think at least some of the reasons I offered should show we should not accept their proposed view, at least not as originally formulated.

3.4 Causation as a hyperintensional notion

I want to conclude section 3 with a brief discussion that can be faced as a standalone subsection. While my main points in this work do not depend on the position I will defend here, I believe it is an important idea to explore, one that I believe has not received much attention in the literature. In brief, I will suggest here that causation is a hyperintensional notion – or, if one prefers to classify the relations themselves *vis-à-vis* their intensionality, and not our notions about them, that causation is a hyperintensional relation.

Before I offer my reasons in favour of this claim, I must explain what hyperintensionality is. It is probably best to start by explaining what intensionality is (and what intensional equivalence is) and then define hyperintensionality in terms of how it differs from it. I will follow Nolan (2014, p. 151) in treating intensional items as "merely intensional". To be more precise, it is best to define these notions in terms of *places in a sentence* as the items being predicated (as Nolan, 2014, does). A place in a sentence is intensional if and only if necessarily equivalent terms can be substituted for one another in that sentence without changing its truth value (i.e., substitution *salva veritate*). Interpreting necessity in terms of possible worlds, A and B are said to be necessarily equivalent if and

only if they exist (or occur) together in the same possible worlds. Analogously, terms picking out the same thing in all possible worlds are said to have the same intension. To illustrate, take the famous example of Hesperus and Phosphorus (both names of the planet Venus). The current standard view in metaphysics is that Hesperus and Phosphorus are necessarily equivalent, since they exist in the same possible worlds (since they are one and the same thing, Venus). The place occupied by "Hesperus" in the sentence "Hesperus is visible in the sky" is intensional according to this definition, as we could substitute the occurrence of "Hesperus" with "Phosphorus", terms with the same intension, and end up with a sentence that is still true. Hyperintensional places in a sentence are those that do not allow for substitution *salva veritate* of terms with a same intension. Classic examples involve intentional notions, such as believing, knowing that, wanting, desiring, thinking that, etc. For example, in "Cassius believes that Hesperus is visible in the sky", the place occupied by "Hesperus" would be hyperintensional. If we substituted "Hesperus" by "Phosphorus", a term with the same intention, in this sentence, we would not be guaranteed to end up with a sentence with the same truth conditions. After all, Cassius might not know that Hesperus and Phosphorus are the same thing, that is, that "Hesperus" and "Phosphorus" are different names for the same celestial body. He might believe that they name different things. Since a key factor governing the truth conditions of the sentence in this example is Cassius believes in, "Cassius believes that Phosphorus is visible in the sky" might turn out to be false even though the original sentence is true, and event though "Hesperus" and "Phosphorus" have the same intension.

Intentional notions, then, are a good illustration of hyperintensionality, and they might count as paradigmatic examples of it. There has been defence, however, that some non-intentional notions should be regarded as hyperintensional, some of them being important metaphysical notions, such as the notion of essence (famously defended in Fine, 1994). This is what is behind what Nolan (2021) calls "the hyperintensional revolution" in early 21st century metaphysics. My proposal here (and that I intend to argue for in detail in future works) is to add to this revolutionary turmoil by suggesting that causation, or, more precisely, that "causes" or "is a cause of", should be understood as a hyperintensional notion – that is, a notion that can show up in hyperintensional places in some sentences.

If "causes" or "is a cause of" is hyperintensional, then this means there are at least two necessarily equivalent items (in this case, let us assume they are events) such that causing a certain effect can be truthfully attributed to one but not to the other. That is, we would have items A and B that are necessarily equivalent, but "A causes E" and "B causes E" do

not have the same truth-value. To give a more precise formulation, we can say that, if "causes"/"is a cause of" is hyperintensional, then the following conditional must be false[77]:

Intensional Causal Conditional (ICC) $\Box$ (Occurs(A) $\leftrightarrow$ Occurs (B)) $\rightarrow \Box \forall X$ (A*C*X$\leftrightarrow$ B*C*X)

Where A and B are events, X is a variable ranging over events, *C* is modelling the causal relation (so A*C*X reads "A causes X" or "A is a cause of X"), and "Occurs(x)" is an operator that applies to events which, naturally, is supposed to mean that a given event occurs.[78] An English gloss of the ICC would then be: if it is necessary[79] that events A and B occur together[80], then, necessarily, A causes every event that B causes and vice versa. We could also formulate the ICC in terms of quantification over possible worlds:

ICC, possible worlds version: $\forall w \in R$ (Occurs(A)$_w$ $\leftrightarrow$ Occurs (B)$_w$) $\rightarrow \forall y \in R \, \forall X$ (A*C*X$_y$$\leftrightarrow$ B*C*X$_y$)

with w and y being variables over possible worlds, R a set of relevant possible worlds, and the added subindex for the "Occurs" operator and the causal relation, specifying in which world the events occur and causation between events occurs, respectively. As is, this formulation makes a weaker claim than ICC does, as it quantifies over a set of relevant possible worlds, whereas the original ICC is naturally understood as covering *all* possible worlds. We can easily make them equivalent by removing the restriction in scope for the quantifiers in the possible worlds version of the ICC. It is convenient, however, to offer a formulation in terms of a set of relevant possible worlds because, as we have seen, this is a move many counterfactualist proposals make.

Now, if causation is indeed a hyperintensional notion, then it means that the ICC must be false. This means there must be at least two events that occur together in all possible worlds they occur, and yet that they differ in what they cause. I believe examples of type I that I discussed in section 3.1 above, such as my fictional migraine symptoms and the

[77] At least for some sentences involving the relevant causal term. I am leaving open the possibility that some, but not all, places in sentences occupied by causal terms are hyperintensional. I believe this will hold for all of them, but this is not an assumption I feel like I need to make at this point. If causation happens to behave hyperintensionally in some places, that would be enough for it being regarded as a hyperintensional notion in at least some contexts, which is enough for my purposes.

[78] I thank Michael Rieppel for this specific suggestion of formulating the ICC, and Roderick Batchelor and Daniel Nolan for helpful comments about ways of making my claims about causation being hyperintensional in general more precise.

[79] As usual, it is debatable what is the appropriate notion of necessity that should figure here. I assume this will be metaphysical necessity (whatever that turns out to be), but I do not think this will matter for what I will say in this work.

[80] If we want to be more precise, we could also build into the formulation a temporal index for the occurrence of the events and for the causal relation. I just omitted these to avoid cluttering of the formulation.

motions in the bodies of Drilies, satisfy these requirements[81]. The pattern of increase in blood flow and the aura occur together in all relevant situations, but it is the patter of blood flow increase, not the aura, that causes the tinnitus. Analogously, the motion in the Drillies' bodies always occur together with their drilling, but it is the drilling, not the bodily motion, that causes the cracking of the nut.

What these examples together with those of type II seem to have strongly indicated already is that no "merely modal" analysis (that is, an analysis in terms of factors existing or occurring together in a class of possible worlds) could capture all that is involved in causation. If causation is hyperintensional, we can now better understand why: it is not a notion that can be captured in purely intensional terms. This would mean that no counterfactual theory of causation could be a complete account of causation, and so that no refined version of Horgan's original counterfactual account of quausation would complete his project.

There is a related possible result that I would like to call attention to. It has been traditionally assumed that causation involves some strong systematic, regular component. If something causes an effect, then a similar factor under similar conditions should cause a similar effect. It would seem odd to suppose that something could be the cause of a given phenomenon in one situation and fail to produce similar effects in a situation with the same (or relevantly similar) background conditions. One way to attempt to capture this is through a formulation like this

Causal Homogeneity Principle (CHP): If A causes B under circumstances C in situation S, then, in all situations relevantly similar to S, A* causes B* under circumstances C* (where A*, B*, and C* are factors occurring in a situation relevantly similar to S that are relevantly similar to (think of counterparts of) A, B, and C, respectively).

Given the intuitive appeal of the sort of systematic regularity (which I tried to capture under the term "homogeneity") traditionally associated with causation, the CHP seems to be true *prima facie*. While I do believe the conditional in the CHP holds true for most of the causal transactions we usually encounter or even think about, I believe it is not true in general. That is, it is not a requirement for causation to hold. I will illustrate my point considering the possibility that causation involves some kind of generation, or production of the effect (see the Appendix for an exploration of this idea). Given this assumption, suppose that A generates (in the appropriate sense) E under circumstances C in situation S. Now, suppose

[81] Especially if we slightly modify them to have a temporal asymmetry.

further that, for whatever odd reason, an item relevantly similar to A, A*, would fail to generate a similar effect, E*, under similar circumstances in a relevantly similar situation S*. According to the CHP, this would entail that A does not cause E in S. I think this is false, and would be a very odd thing to maintain. The right verdict here seems to me that, as long as A generates E in the appropriate sense of generation that is supposed to be involved in causation, this should be sufficient to claim that A causes E *in S*. It seems that A (or something very much like it) would not cause a similar effect (like E*) in situations like S*. But that seems no reason to detract the result that, in S, A *does* cause E. If this result is correct, this also adds to the impression that causation is hyperintensional and that a "merely modal" account of causation is not possible, as it indicates that what happens in other possible worlds (or situations), however similar they are to the one under consideration, should not determine whether a causal transaction occurs *in that world* or not.

If what I said is correct, then causation should be included on the list of other metaphysical non-intentional notions that have recently been caught up in what Nolan calls the "hyperintensional revolution". This is a topic that deserves further exploration, and is a project I intend to take up in future works.

4. The Epiphenomenalist Dilemma and the Mental Efficacy Trilemma

The challenge posed by the Quausal Exclusion Argument should be seen as serious if my defence of it has been successful. As we have seen, it already makes the challenge to be met harder by restricting the scope of possible reductionist theses that can evade it. Still, it might be seen as the same type of argument as before, targeting mainly forms of nonreductionism and serving as a *reductio* of them. It would only serve to show that more types of nonreductionist views are affected by exclusionary considerations. Though that, as I stated before, can be seen as an interesting result on its own, it might also be interpreted as strengthening the case for a more radical form of reductionism. It seems to make clear what we have to do to avoid epiphenomenalism: accept that phenomenal property instances are reducible in some appropriate way to physical property instances. As I indicated before, however, matters are not so simple. I believe the issues involved in the Quausal Exclusion Problem are part of a more general threat to mental quausation. This time, it seems that not even reductionist views about phenomenal property instances could escape, and that QE seems to be impossible to vindicate. As a result, quausal epiphenomenalism, which I argued is just as bad as epiphenomenalism *simpliciter*, seems the unavoidable conclusion. No matter where we look, epiphenomenalism seems to "stick out its ugly head", menacingly stalking in

the shadowy corners, just waiting to inevitably meet us somewhere or other, regardless of the path we take. I called my argument for this sombre conclusion the Epiphenomenalist Dilemma.

The Epiphenomenalist Dilemma

P1. Either instances of mental properties are reducible to instances of physical properties or not. (the kind of reduction I am assuming here, as before, is a form of ontological reduction).

P2. If instances of mental properties are not reducible to instances of physical properties, then instances of mental properties are not causally relevant, i.e., QE is false. (because of the Quausal Exclusion Argument)

P3. If instances of mental properties are reducible to instances of physical properties, then there is no relevant difference between mental and physical properties that would warrant the claim that it is in virtue of the *mental* properties that alleged mental causes cause whatever it is that they do. Hence QE is false in this case.

Therefore,

C. QE is false.

This time, the only premise that needs clarification is P3. I believe this is a good way to explain the motivating idea behind it: whatever the relation of reduction that is supposed to hold between token phenomenal properties and token physical properties, it will either be the case that there is some ontological distinction between them or not. If there is, it seems the same exclusionary worry we have seen in the exclusion problem could be raised once again at this level: are the mental aspects doing any "causal work" in this case? Can we in any relevant sense trace back the causal influence in producing the effects in question to *them*? It seems, given the motivation for P2 of my Quausal Exclusion Argument, that there are always plausible candidates of a true and complete "causal story" we can tell that accounts for the occurrence of the effects that only appeals to the physical properties. Given this (and it is not clear how any reductionist account of the kind in question here could challenge this), the mental aspects are, once again, threatened with causal exclusion given the considerations of my Quausal Exclusion Argument. On the flipside, if, on this reductionist account, there simply is no ontological difference between token phenomenal properties and some token physical properties, then it seems we lose the ability of making any "*qua*" qualification that discriminates between mental and physical factors. The only way I envision where we could keep this possibility is if we move away from a causal realist view, defending instead that a

factor causing "*qua* mental" or "*qua* physical" would be a matter of how we describe or conceptualise the causal transactions. But this is not a move available given my assumption of a causal realist view.

We can construe a version of the same problem that is a trilemma instead. In this case, we would consider the question of whether there are instances of mental properties in the first place. The dilemma version I presented can reasonably be called epiphenomenalist because it would entail that a form of epiphenomenalism is true in either case: in both cases, the mental property instances are real and causally inefficacious. This trilemma version, however, would have one of its horns leading to a case where QE is false but where epiphenomenalism is not true (the case where there are no mental property instances). Therefore, I would suggest a different name for the trilemma version, the more general "The Mental Efficacy Trilemma". A version of this trilemma can be formulated as follows:

The Mental Efficacy Trilemma

P1. Either there are instances of mental properties or not.

P2. If there are no instances of mental properties, then QE is false (trivially).

P3. If there are instances of mental properties, then either they are reducible to instances of physical properties or not.

P4. If instances of mental properties are not reducible to instances of physical properties, then instances of mental properties are not causally relevant, hence QE is false.

P5. If instances of mental properties are reducible to instances of physical properties, then there is no relevant difference between mental and physical properties that would warrant the claim that it is in virtue of the *mental* properties that alleged mental causes cause whatever it is that they do. Hence QE is false in this case.

Therefore,

C. QE is false.

This is just an expansion of the considerations of the dilemma to include the possibility of some kind of eliminativism being true. Eliminativism about the mind, about consciousness, or about mental properties in specific are positions that would entail that there are no mental properties.

This trilemma is the broadest and most serious incarnation of the causal exclusion problem I ever encountered. Let us pause for a moment to appreciate its conclusion. It would show that, no matter what stance we take on the exclusion problem, we will end up with a position in which the Quausal Efficacy thesis is false. If what I defended in Part II was any good, QE is supposed to be the version of the thesis of mental efficacy we would like to salvage if we seek to retain the minimally common-sensical, pre-theoretical view of mental causation we possess. QE being false means that the bitter taste did not cause Paula's spilling of the drink, that John did not go to the kitchen because he felt thirsty, that the itching did not cause the scratching. It seems to be, very much to Fodor's demise, the end of the world. It means that either a form of eliminativism is true or a form of quausal epiphenomenalism is true. In the latter case, that means our phenomenal states are very much real, but that all causal influence we might attribute to them is purely illusory. We are wrong in assuming our behaviour (or *anything*, really) occurs in virtue of the conscious experiences we have. What would that imply? The deeply counterintuitive results abound. While I do believe some of the usual worries are less devastating than usually supposed (such as the idea that consciousness could not have evolved if it were causally inefficacious), there are some serious problems I currently see no way of responding to. Aside from the alienating aspect I mentioned in the introduction, I highlight (and repeat) the problems about knowledge, reference, and personal ontology. It seems we can know (and know with paradigmatic certainty) what we feel. But how could we come to know something we could not have causal contact with? And, for the same reason, how could we refer to phenomenal states when we talk about them, something we obviously assume we can do[82]? How to shake-off the feeling suggested by this fragmented picture of the world that, when we say something true about our sensations, we are simply *lucky* to match what we feel with what we say or believe?

Finally, there is another problem that I have not seen explored in the literature before. It has to do with personal ontology, that is, our theorising about what kind of entities we (or persons in general) are. This problem affects a particular variant of epiphenomenalism. Suppose we accept that some, but surely not all, mental states are epiphenomenal. We could, say, believe that only *phenomenal* states are epiphenomenal (because they are irreducible), but that beliefs, desires, memories, etc. are not (likely because we believe these *are* reducible)[83]. But we surely suppose that we have access to both our beliefs and our conscious experiences, or that our phenomenal states can communicate in some way with these other types of state

[82] Such as when someone says, correctly, that "this pain I feel is stingy".
[83] As I mentioned in Part I, this would be a view similar to the one defended by Kim in his (2005).

(such as when I correctly remember something I felt). Furthermore, if we believe in the very broad thesis that our conscious states, as well as our beliefs, desires, memories, etc. at least partially constitute what we are (that is, if we believe in a very broad, or minimal, version of a psychological theory of personal ontology), then *what kind of thing are we*? Is it even possible for an object to satisfy these criteria? If we keep the idea that reducible mental states are reducible to physical states and that irreducible states are non-physical, then it seems we would have to be some kind of "hybrid" entity, partially constituted by physical states and partially by nonphysical ones. This would already be enough to pose problems of interaction, but if we also state that the irreducible states are causally inert (as epiphenomenalism requires), then any form of interactionist dualism is discarded and we are left with an utterly mysterious, fragmented and seemingly incomprehensible entity. Surely there must be some kind of communication between the phenomenal aspect of our minds and the non-phenomenal (such as beliefs, for example), and we are supposed to be the kind of thing that can have some kind of access or contact with both. But how could that be possible? What kind of thing would be capable of satisfying all of these requirements? At the end, we are left in a challenging and awkward situation: these results seem so absurd that epiphenomenalism *must* be false. But if it is, then something must be wrong with my Quausal Exclusion Argument or, more generally, with the Menal Efficacy Trilemma. What could that be?

Conclusion

In this work, I set out to establish my contribution to the literature on the problem of causal exclusion. I believe the threat it poses to our dear claim of causal efficacy of mental states is more serious than philosophers usually assume. A good part of the motivation for this belief of mine was the appeal I found in Terry Horgan's defence of the importance of vindicating a notion of mental quausation. Since it requires a more fine-grained theory of causation to avoid epiphenomenalism, I suspected the challenge of avoiding epiphenomenalism to be harder than those who do not acknowledge this need for quausation would normally assume. With my interpretation of Horgan's general idea of mental quausation, I formulated a new version of the causal exclusion argument, concluding that any view that does not reduce phenomenal property instances to physical property instances would fall prey to a relevant form of epiphenomenalism, which I called quausal epiphenomenalism, following Horgan. I presented and replied to what is the most challenging account to my argument I found to date, the theory of causation as making a difference defended by Peter Menzies and Cristopher List. Finally, I generalised the threat of exclusion to include even the forms of

reductionism that the Quausal Exclusion Argument seemed to leave untouched. We are left, then, with the result that one of the most abhorred theses in the history of philosophy must be true. My overall argument screams for refutation, and I eagerly wait for insightful responses to it that might come.

I was clear in Part I (and throughout) that I was assuming a form of causal realism in my discussion of the causal exclusion argument. This leaves open an obvious avenue for resisting my line of argument: rejecting my assumption that a form of causal realism is true. I do believe there is fruitful ground to explore in the antirealist side of the debate on the nature of causation, and I am by no means confident that a form of antirealism might not turn out to be the best view or that there might not be a promising way out of the conundrum available to antirealists. I do think, however, there are good reasons to favour a realist rather than an antirealist view of causation; more specifically, to favour a view of causation being a real phenomenon in the world, instead of being a linguistic or conceptual artifact we deploy to make sense of phenomena. I would like to briefly present what I currently see as the best reason to favour a realist position. It is an application of positions I defend on the debate of explanatory realism more generally. The key thing to notice is that surely there are types of explanations we come up with that involve describing what happened in the world. What I have been referring to as causal descriptions would be a great example when it comes to causation, as we could naturally attempt to explain why a given effect occurred by putting together a "causal story" spelt out merely in terms of those descriptive statements. For this type of explanations, which we may call "descriptive explanations" for the lack of a better term, the realist has a simple and plausible account of what differentiates those explanations that "get things right" from those that do not: the ones that describe what happened in a way that corresponds to what actually happened "in the world", between the things evoked in the description, get things right. They give successful, accurate explanations. This is the kind of explanation we usually aim for (if we have sincere, cognitive aims)! The crucial point is that the antirealist has a much harder time when trying to come up with a criterion for separating the explanations that "get things right" from those that do not. We can surely come up with stories that make perfect sense, would render the *explanandum* intelligible, but that are false. If all that was involved in the success (in the broad sense of success) of explanations were the cognitive virtues they exemplify, as antirealists would normally defend, then it seems there is no way the antirealist can distinguish an explanatorily virtuous story that is false from those that are true when it comes to their being successful. Appealing to accurately describing what transpired or "being true" as a cognitive virtue is not an available

move, as it would mean the quality, the success of a descriptive explanation would hinge on whether its description corresponds or not to what goes on in the world, which would be a collapse into realism. But if not that, it is not clear (at least I cannot envision) what else antirealists can appeal to that is available to them in order to account for this key difference. In a sentence, then, my conclusion is this: the realist can, while it seems the antirealist cannot, offer an account of what differentiates descriptive explanations (in this case, *causal* descriptive explanations) that get things right from those that do not. Unless some surprising move is available to get antirealism out of this deadlock, I believe we have good reasons to favour causal realism.

When it comes to further developments to my project, next steps that I clearly envision at the moment to further my research would involve exploring other theories of mental causation that might show that my Quausal Exclusion Argument is not sound; I have especially in mind addressing trope theories of causation[84], theories of causation in terms of transmission of a physical quantity (such as the Salmon-Dowe theory; e.g., Dowe, 1992), and powers-based accounts of causation, which have become more popular in 21st century. Offering a proper theory of property causation to supplement my interpretation of mental quausation would also be a fruitful task if possible.

[84] Though the broad picture of property causation I presented seems quite similar to trope theories of causation, I believe Robb's verdict about his position solving the exclusion problem does not work. Here, I believe Noordhof (1998) and Gibb (2004) are correct in their objections and that they are decisive.

References:

Árnadóttir, S.T. & Crane, T. (2013). "There is no exclusion problem". In Gibb, Lowe & Ingthorsson (eds) 2013. pp. 248–265.

Bennet, M. R. & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Wiley-Blackwell.

Bernstein, S. & Wilson, J. M., 2016, "Free will and mental quausation." *Journal of the American Philosophical Association* 2 (2): pp. 310-331.

Block, N. (1990). "Inverted Earth", *Philosophical Perspectives*, 4: pp. 53–79.

Campbell, N. (2003). "Causes and causal explanations: Davidson and his critics." *Philosophia* 31 (1-2):149-157.

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

_____ (2010). *The Character of Consciousness*. Oxford University Press USA.

Crane, T. (1995). "The Mental Causation Debate", *Proceedings of the Aristotelian Society*, Supplementary Vol. 69: pp. 211–36.

Cummins, R. (2010). *The World in the Head*. Oxford University Press.

Davidson, D. (1963). "Actions, Reasons, and Causes", *Journal of Philosophy*, 60: 685–700. Reprinted in Davidson 1980, pp. 3–19.

—— (1970). "Mental Events", in L. Foster and J. W. Swanson (eds.), *Experience and Theory*, Amherst, MA: University of Massachusetts Press, pp. 79–101. Reprinted in Davidson 1980, pp. 207–25.

—— (1980). *Essays on Actions and Events*, Oxford: Clarendon Press.

____ (1993). "Thinking Causes", in J. Heil and A. Mele (eds) *Mental Causation*, Oxford: Clarendon Press.

Dowe, P. (1992), "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory", *Philosophy of Science* 59: pp. 195-216.

____ (2000). *Physical Causation*. Cambridge University Press.

____ (2001). "A Counterfactual Theory of Prevention and 'Causation' by Omission," *Australasian Journal of Philosophy*, 79: pp. 216–26.

Dretske, F., (1989). "Reasons and Causes", *Philosophical Perspectives*, 3: pp. 1–15.

Fine, K., (1994). "Essence and Modality: The Second Philosophical Perspectives Lecture", *Philosophical Perspectives*, 8: pp. 1–16.

Fodor, J. (1974). "Special Sciences (or: The Disunity of Science as a Working Hypothesis)", *Synthese*, 28(2): pp. 97–115.

____ (1989). "Making mind matter more", *Philosophical Topics 17* (11): pp. 59-79.

Garber, D. (1983). "Understanding Interaction: What Descartes Should Have Told Elisabeth," *Southern Journal of Philosophy* (Supplement), 21: pp. 15–37.

Gibb, S. C. (2004). "The Problem of Mental Causation and the Nature of Properties," *Australasian Journal of Philosophy*, 82(3): 464–476.

____. (2013). "Introduction", in Gibb, Lowe & Ingthorsson (eds.) 2013, pp. 1–17.

____ (2015). "The Causal Closure Principle", *Philosophical Quarterly* 65 (261): pp. 626-647.

Gibb, S. C., Lowe, E. J. and Ingthorsson, V. (eds.) (2013). *Mental Causation and Ontology*, Oxford: Oxford University Press

Hitchcock, C. (2012). "Theories of Causation and the Causal Exclusion Argument", *Journal of Consciousness Studies* 19 (5-6): pp. 40-56.

Honderich, T. (1982). "The Argument for Anomalous Monism", *Analysis*, 42: pp. 59–64.

Horgan, T. (1989). "Mental quausation". *Philosophical Perspectives* 3:47-74.

—— (1998). "Kim on Mental Causation and Causal Exclusion", *Philosophical Perspectives* 11, pp. 165-84.

—— (2001). "Causal Compatibilism and the Exclusion Problem". *Theoria* 16, pp. 95-116.

—— (2011). "The Phenomenology of Agency and Freedom: Lessons from Introspection and Lessons from Its Limits." *Humana Mente* 15 (Jan. 27, 2011), pp. 77-97. Issue: *Agency: From Embodied Cognition to Free Will*

—— (2015). "Injecting the Phenomenology of Agency into the Free Will Debate". In D. Shoemaker (ed.), *Oxford Studies in Agency and Responsibility*, 3. Oxford (2015), pp. 34-61.

—— , forthcoming, "Agentive Self-Awareness and the Nature of the Conscious Self." In J. Bugnon, M. Nida-Rümelin, and D. O'Conaill (eds.), *The Phenomenology of Self-Awareness and the Nature of Conscious Subjects*. Routledge, forthcoming.

Hurka, T. (2015). "Moore's Moral Philosophy", *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2015/entries/moore-moral/>.

Jackson, F. (1982). "Epiphenomenal Qualia", *Philosophical Quarterly*, 32: pp. 127–36. Kim, J. (1989). "Mechanism, purpose, and explanatory exclusion". *Philosophical Perspectives* 3:77-108.

Kim, J. (1990). "Supervenience as a philosophical concept." *Metaphilosophy* 21 (1-2): pp. 1-27. Reprinted in Kim, 1993.

_____ (1992). "The nonreductivist's troubles with mental causation". In John Heil & Alfred R. Mele (eds.), *Mental Causation*. Oxford University Press.

_____ (1993). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge University Press

_____ (1993b), "Can Supervenience and 'Non-Strict Laws' Save Anomalous Monism?", in Heil and Mele, 1993, pp. 19–26.

_____ (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*, MIT Press.

_____ (2005). *Physicalism, or something near enough*, Princeton University Press.

_____ (2006). "Emergence: Core ideas and issues". *Synthese* 151 (3): pp. 547-559.

_____ (2008). "Reduction and reductive explanation : is one possible without the other?" In Jakob Hohwy & Jesper Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation.* Oxford University Press.

_____ (2010). *Essays in the Metaphysics of Mind*. Oxford University Press.

_____ (2010a). "Taking the agent's point of view seriously in action explanation", in Kim, 2010, pp. 125-147.

_____ (2010b). "Two Concepts of Realization, Mental Causation, and Physicalism", in Kim, 2010, pp. 263–281.

Leibniz, G.W. (1991). "Monadology", in Rescher, N., ed. *G.W. Leibniz's Monadology: An Edition for Students*. Pittsburgh: University of Pittsburgh Press.

LePore, E., and Loewer, B. (1987). "Mind Matters", *Journal of Philosophy*, 84: pp. 630–42.

Lowe, E. J. (2000). "Causal Closure Principles and Emergentism". *Philosophy*, 75: pp. 571–86.

Macdonald, C. and Macdonald, G. (2006). 'The Metaphysics of Mental Causation'. *Journal of Philosophy*, 103: pp. 539–76.

Mackie, J. L. (1965). Causes and Conditions. *American Philosophical Quarterly* 2 (4): pp. 245 - 264.

Malcolm, N. (1968). "The conceivability of mechanism". *Philosophical Review* 77 (January): pp. 45-72.

Menzies, P. and List, C. (2009). "Nonreductive Physicalism and the Limits of the Exclusion Principle". *Journal of Philosophy*, 1006: pp. 475–502.

_____ (2010). "The Causal Autonomy of the Special Sciences". In G. and C. Macdonald (eds.), *Emergence in Mind*. Oxford: Oxford University Press: pp. 108–28.

_____ (2017). "My brain made me do it: The exclusion argument against free will, and what's wrong with it." In H. Beebee, C. Hitchcock & H. Price (eds.), *Making a Difference: Essays on the Philosophy of Causation*. Oxford: Oxford University Press.

Menzies, P. (2009). "Platitudes and counterexamples". In Helen Beebee, Peter Menzies & Christopher Hitchcock (eds.), *The Oxford Handbook of Causation*. Oxford University Press. pp. 341-367.

Merricks, T. (2001). *Objects and persons*, New York: Oxford University Press.

Montero, B. (2003). "Varieties of Causal Closure", in Walter and Heckmann 2003, pp. 173–87.

—— (2006). "What Does the Conservation of Energy Have to Do with Physicalism?", *Dialectica*, 60: pp. 383–96.

Noordhof, P. (1998). "Do Tropes Resolve the Problem of Mental Causation?" *Philosophical Quarterly*, 48(191): 221–226.

Papineau, D. (2000). "The Rise of Physicalism", in M. W. F. Stone and J. Wolff (eds.), *The Proper Ambition of Science*, New York: Routledge, pp. 174–208

Paul, L. A. (2000). "Aspect Causation". *Journal of Philosophy* 97 (4):235.

Pessoa Jr., O. and Melo, L. (2015). "O dualismo interacionista não precisa violar leis de conservação da física", in M. Aiub, M. C. Broens, & M. E. Q. Gonzalez (orgs.), *Filosofia da mente, ciência cognitiva e o pós-humano: para onde vamos?*, São Paulo: FiloCzar, pp. 119-121.

Polger, T. and Shapiro, L. (2016). *The Multiple Realization Book*, New York: Oxford University Press.

Putnam, H. (1967). "Psychological Predicates", in W.H. Capitan and D.D. Merrill (eds.), *Art, Mind, and Religion*, Pittsburgh: University of Pittsburgh Press, pp. 37–48.

Robb, D. (1997). "The Properties of Mental Causation," *The Philosophical Quarterly*, 47(187): 178–194.

—— (2001). "Reply to Noordhof on Mental Causation," *The Philosophical Quarterly*, 51(202): pp. 90–94.

_____ (2013). "The Identity Theory as a Solution to the Exclusion Problem". In

Gibb, Lowe & Ingthorsson (eds.) 2013, pp. 215–232.

Robb, D. and Heil, J. (2021). "Mental Causation", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/mental-causation/>.

Russell, B. (1956) "Mind and matter", in *Portraits from memory and other essays*, Simon & Schuster, New York, pp.145-65.
_____ (1959). *My Philosophical Development*, London: Routledge, 1995 edition.

Schaffer, J. (2016). "The Metaphysics of Causation", *The Stanford Encyclopedia of Philosophy* (Fall 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/>

Shoemaker, S. (1982). "The Inverted Spectrum", *Journal of Philosophy*, 79: pp. 357–81.

_____ (2003). 'Realization and Mental Causation'. In *Identity, Cause, and Mind*, Expanded Edition. Oxford: Clarendon Press: 427–51.

_____ (2007). *Physical Realization*. Oxford University Press, UK.

Smith, M. (1994). *The Moral Problem*. Blackwell.

Sosa, E. (1984). "Mind-Body Interaction and Supervenient Causation." *Midwest Studies in Philosophy*, 9, pp. 271-282.

Sosa, E. & Tooley, M. (eds.) (1993). *Causation*. Oxford University Press.

Strawson, G. (2006). "Realistic monism: why physicalism entails panpsychism". *Journal of Consciousness Studies* 13 (10-11):3-31. Reprinted in Skrbina, David (ed.) (2009). Mind That Abides. Panpsychism in the New Millennium. John Benjamins.

van Gulick, R. (2001). "Reduction, emergence and other recent options on the mind/body problem: A philosophic overview". *Journal of Consciousness Studies* 8 (9-10): pp. 1-34.

Walter, S. and H. Heckmann (eds.) (2003). *Physicalism and Mental Causation: The Metaphysics of Mind and Action*, Exeter: Imprint Academic.

Witmer, D. (2000). "Locating the overdetermination problem". *British Journal for the Philosophy of Science* 51 (2): pp. 273-286.

Yablo, S. (1992). "Mental causation". *Philosophical Review* 101 (2):245-280.

Yalowitz, S. (2021). "Anomalous Monism", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/anomalous-monism/>.

Zhong, L. (2011). Can Counterfactuals Solve the Exclusion Problem? *Philosophy and Phenomenological Research* 83 (1):129-147.

Appendix A:
Causation, Communication, and the transmission of information

In any kind of philosophical inquiry, recurring to some platitudes concerning the concepts involved seems inescapable when we try to establish a measure of adequacy of a given account. For example, if someone offers a theory of knowledge that has as a result that we can know things that are false (such as knowing that 2+2=5 or that it is raining when in fact it is not), we would readily find such an account improper. After all, it is a platitude about knowledge that we can only know something that is the case, i.e., that knowledge is *factive*.

What I mean by "platitude" is the simple idea of a patent truth, especially truths involving a concept that seem "pre-theoretical" and that we carry along when we theorise about that concept and what it involves. They are truths the apprehension of which might be required to master a concept and that not only guide and restrain our investigations, but also help us frame the problems in the first place. If one is not so content with the term "platitudes", I intend what I have in mind to be just as appropriately captured by "pre-theoretical results", if that does any help. Platitudes can be conceptual truths of the classical kind (such as the fact about the factivity of knowledge) or even results deemed obvious (e.g., "torturing innocent people for fun is wrong" is a moral truth if anything is, and any theory in normative ethics must include this result in order to be minimally plausible) [85]. The case of causation, of course, is no exception and there are many platitudes about it. To cite a few: "causes happen before their effects", "causation is asymmetrical", "had the cause not been present, its effects would not ensue (at least in the absence "causal back-ups")", "the occurrence of the cause necessitates the occurrence of the effect", and "causes generate their effects". Trying to come up with a theory that accommodates *all* the platitudes is a quest for an ideal philosophical Holy Grail. What usually happens, however, is that some platitudes are favoured in detriment of others, with an attempt to explain why some of the apparent platitudes were left out. In fact, focus on different causal platitudes might be a plausible

[85] My ideas about the role of platitudes in philosophy were influenced by Michael Smith's book "The Moral Problem" (Smith, 1994), where they satisfy this role of pre-theoretical results we would seek to preserve and organise with our philosophical theories.

explanation for the feeling one sometimes get that different parties in the debates about causation disagree so wildly that they probably have different concepts in mind (for this impression, see Hitchcock, 2012, Menzies, 2009 and Sosa & Tooley, 1993). In this work, I will follow the practice of favouring a platitude in detriment of others. The platitude in question is the last in my list of examples, and, since I take it to be so important, I will highlight it and refer to it as the "Generative maxim":

Generative maxim: causes generate their effects.

Now, before I proceed any further, I must recognise that some will probably object to the generative maxim being a platitude about causation, or at least for it being as much of a platitude as the other examples I cited.[86] Those who subscribe to regularist views of causation would probably be good candidates, for example. That is fair enough; there has been enough disagreement about this very topic for this claim to be reasonable. In reply, there are a few remarks I would like to make: 1. Though I will end up favouring a rather "oomphy" interpretation of the generative maxim (and, therefore, of causation), I intended the generative maxim to express something broader, less charged and, thus, for it to be in some sense compatible with other, less "connectivist", views. For example, the notion of "making a difference" defended by Menzies & List and explored in Part III can be interpreted in a way that attempts to accommodate the generative maxim in its own terms. We could try to say that causes generate their effects by "making them happen", and then defend that this notion is to be treated in mere counterfactual terms like Menzies & List do. In this case, even though there is no real "causal connection" being assumed between cause and effect, the alleged platitude can be accommodated in some way. 2. I believe that those who would vehemently deny that the generative maxim is at all part of our pre-theoretical *desiderata* about causation would probably simply have a different conception of what causation is from those do accept it (see my remark at the end of the preceding paragraph). I take this to be evidence in favour of this probable multiplicity of causal concepts, and it is interesting to think what we should make of it. 3. If one does not want to concede to the generative maxim being a platitude, then what follows can be faced as an exercise in exploration of some results we could come up with if we explored this path. I hope my exposition can be seen as a valuable endeavour either way, especially given the problematic exclusionary results I arrive at (again!) at the end, which the opponents of the generative maxim might welcome gladly.

[86] I thank Wilson Bezerra for this important observation.

One alternative way of phrasing the maxim is by saying that "causes make their effects happen". I take these to be equivalent because the idea that the cause *makes* the effect happen seems to me to convey the same overall idea of some kind of "exertion" worked by the cause in producing the effect that the notion of "generation" conveys. Since, as I have discussed, there already exist theories of causation centred on the notion of "making a difference" (which sounds quite like "making something happen") that are importantly different from the view I am espousing here (see Menzies, 2004; Menzies & List, 2009; 2010; 2017), I prefer to write in terms of "generation" or "production" to avoid confusion[87]. What the maxim means is quite simple, as is to be expected of a platitude. It means that a billiard ball hitting another in a given way is what produced the latter's motion, and the origin of its movement can be traced to the one that hit it. It means that the sharp tip of a pin in motion literally generated a hole in the board. It means, and this is more relevant, that there is some kind of influence exerted by the cause on its effect. This might not seem to be saying much, but the kind of influence alluded to here is enough to distance my view from regularist or counterfactual theories of causation, since they describe causal transactions in terms of events happening or not together in a given order in given scenarios. They do not require any kind of *real* interaction between causes and effects, no kind of "causal glue" binding them together; in effect, they cannot maintain that causes *generate* their effects.

In fact, I take the Generative maxim to be so central that it can actually be taken to be equivalent with causation. That is, any statement of the form "A causes B" could be rephrased as "A generates B" and have the same truth-conditions. I take the maxim to take precedence over other platitudes because, as long as it is maintained, counterexamples to the other platitudes would not constitute counterexamples to there being causation. To illustrate my point, consider the idea that causes must occur before their effects. This is widely regarded as a central platitude concerning causation. However, some possible counterexamples can be raised against it. The phenomenon of quantum entanglement is a famous one that concerns the possibility of simultaneous causal interaction. If it really is the case that affecting an entangled particle would result in an instantaneous response in the other particle it is paired with, then this looks like a counterexample to the maxim. After all, it seems true to say that the alteration in the first particle produced the change in the second; the other particle underwent a change in virtue of the other being affected. This seems

[87] Jaegwon Kim favours a conception of causation along the lines of the "generation-centric" view I am defending here in his Kim, 2005.

enough to say that this is a case of causation, and so it could be a case of *instantaneous* causation (if actually occurring as described).

Conversely, if an account manages to preserve all the other platitudes but fails to preserve the Generative maxim, it will not be a plausible account of *causation*. This seems to be precisely the point raised by many counterexamples to counterfactual theories of causation: they consist of cases that satisfy the criteria stipulated by those theories but that involve no kind of generative interaction (or interaction whatsoever) between the occurrences in question[88].

One way to capture this broad idea of generation is through the notions of *transmission* or *"communication"*. Consider the following examples of causal statements:

I. 'The table was smashed by the hammer'

II. 'Lightning boosted the concentration of ozone in the air, which made the townspeople feel the smell of rain'

III. 'The announcement of the success of the candidate in the recent polls caused a fall in the stock market'

IV. 'Robert jolted down the answer to the riddle because he remembered that "altius" means "higher" in Latin'

V. 'Nathan paid attention to his pain and then told the doctor it feels like "muscle pain", not "organ pain"'

All of these cases seem to involve some kind of transmission. A still general but slightly more specific interpretation is to take these cases to involve transmission of information. The problem is that it is not clear what notions of 'information' and 'transmission' are in play here. The usual account of 'transmission of information' we find in the literature about causation is one centred on the notion of physical information (or physical quantity). This is the kind of view espoused by philosophers such as Phil Dowe and Wesley Salmon. According to these views, causal transactions involve the transmission of some kind of physical quantity (such as momentum or energy). This is perhaps a plausible

[88] A way this can happen is if C and B are both effects of A, but do not interact between themselves. One example is the relationship between lightning (the luminous flash), thunder and the electric discharge that causes both. The relationship between lightning and thunder satisfies the criteria of most counterfactual or regularist accounts of causation: if lightning occurs, thunder occurs afterwards and, if lightning does not occur, then neither does thunder. Yet, as we know, lightning does not cause thunder: they are just effects of a same cause.

account of what we can call "straightforward physical causation", which is the kind involved in I and (at least the first half of) II; after all, the transmission of momentum and energy from the hammer to the table seems to be all the transmission we need to bring up in order to explain what happened in the first case. The first half of II might be a bit more complicated, but there should be no problem in explaining the transactions involved in the production of ozone by an electrostatic discharge in terms of 'transference of physical quantities'. The second half of II involves physical-to-mental causation, which might have its problems when irreducible mental states are involved (and sensations are preferred candidates when it comes to irreducibility), but the physiological process of feeling the smell of rain can fit in this scheme.

It might seem strange to posit that information, a seemingly *abstract* thing, might be what is involved in the very concrete relation of causation. In response to this worry, I ask the reader to consider two things: 1. We can understand information as a physical item, one that is realised in physical data. While an air of mystery seems to remain, I just want to draw attention to the view that there might be such a thing as "physical information". 2. There is at least one good reason to prefer an account in terms of information instead of a clear physical quantity, like energy or momentum. Consider a case of entangled particles, the best real candidate we have for "spooky action at a distance". Affecting the state of one entangled particle will cause a change in the state of the other. However, there surely is no physical quantity like energy or momentum being transmitted from one particle to other. It seems we can, however, make sense of saying that information of some kind is being sent from one to the other (even though this might not involve a medium in the usual sense of transmission). This does not help much; the case of entanglement still seems very strange and perhaps even this proposal will not do (perhaps "transmission" of any kind is not the kind of thing that can "leap" and be carried out "at a distance"). But at least it is not obviously a non-starter like other proposals in terms of identifiable physical items such as energy or momentum.

As we have seen, the account of transmission of information in terms of physical quantities manages to deal well enough with cases of straightforward physical causation. However, it does not fare well at all when it comes to more complex cases, such as III. Trying to analyse the causal link in III in terms of physical quantities will not do. Surely, this case poses no threat to the causal closure principle. After all, we can see that it ultimately involves many transactions that, individually, fit in the "physicalist scheme" previously stated. For instance, "the announcement", as a happening, is surely a physical event (or process). And, surely, the communication of what is being announced is made through other physical

transactions (e.g, the audio of the announcement being registered by a camera, which records it in a disk, which is read by a computer, through which it is posted on the internet and so on; the same goes for the actions of the many stockholders). The point, however, is that bringing up this massive collection of physical transactions will not explain what is said in III. This is a limitation of the physical quantity account: it works well in a restricted level of description, but we can (and often do) make causal claims in much higher levels of abstraction. This seems like a good reason to believe that this cannot be a general account of causation. So what is going on, then, in cases like III? What kind of transmission does it involve? And what kind of information is in play? I do not yet know.

Examples IV and V are cases of mental causation. What the notion of quausation helps us see, I argued, is that the peculiarly mental characteristics of the causes are indispensable to explain the occurrence of the effects in these cases. Where can transmission of information enter here? In V, it seems that the information that the feeling of pain was of a particular kind (muscle pain) must have been passed on, or "communicated" (in quotes since it does not involve a linguistic communication) to the later stages on the causal chain. The idea is that in order for Nathan to say what he said to the doctor the information that the feeling of pain was of a specific kind must have been communicated. This might be one way of interpreting the quausal requirement stated in the fourth clause of Horgan's definition of the quausal relation (that is, that these mental aspects are explanatorily relevant to explain why the effect occurred). So, in this interpretation, the effect under consideration occurred in virtue of the fact that some information that is indispensable for bringing about the effect and that is peculiar to the mental cause was transmitted. Since pain is a sensation, then if sensations are presumably irreducible mental states, this information cannot be physical. So what kind of information could it be? And how could it be passed on to physical things? It is tempting to say that it is "mental information", "qualitative information", or "phenomenal information," but until we explain what *these* are supposed to mean we have made no progress. It surely does not seem to be some kind of semantic information either, though that might be precisely what is involved in cases such as IV. So, given that it cannot be a kind of physical information, it is hard to see how this suggestion can possibly lead to anything other than some kind of interactionist dualism.

If what was said in the last paragraph is correct, then irreducible mental causation leads to some kind of interactionist dualist picture of the world. But there is a serious problem with basically all kinds of interactionist dualism: they seem to require a violation of the causal closure of the physical domain, for, irrespective of their specific formulations, they all posit

that there are non-physical causes of physical effects. If we really have good reason to maintain that the CC principle is true, then things do not look good for interactionist dualist theses (and any thesis that entails some version of it, for that matter). This provides us with another avenue for causal exclusion: if irreducible mental causation requires interactionist dualism and interactionist dualism is incompatible with the CC principle, then there can be no irreducible mental causation. We can try to capture this idea with the following argument, which is another version of the exclusion argument in its own right:

P1) If causation requires communication/transmission from cause to effect, then, if there are irreducible mental causes of physical effects, there is communication/transmission from irreducible mental entities to physical entities.

P2) It is impossible for irreducible mental entities to communicate/transmit something to physical entities.

Therefore,

C. If causation requires communication/transmission from cause to effect, then it is impossible for irreducible mental entities to cause physical effects.

A first thing to note is that this argument actually seeks to establish a stronger conclusion than the one previously stated: it does not only state that there can be no irreducible mental causation; it states that irreducible mental causation is *impossible*.

There are two main reasons for believing that P2 is true. The first is that any kind of influence exerted by an irreducible non-physical entity on physical entities would presumably incur in a violation of the conservation of energy. The second reason is a more general worry about any kind of interactionist dualism: that the mode of interaction of non-physical things is incompatible with the mode of interaction of physical things. They are, in this sense, incommunicable. This is basically the point raised by Princess Elisabeth of Bohemia in her objection to Descartes' dualism (see Garber, 1983): physical objects can only affect and be affected in certain characteristic ways that involve physical attributes (according to the dominant mechanistic view of the time, this mode of interaction ultimately consisted of "pushing and pulling"). Non-physical things, lacking any physical attributes, would thus be unable to interact with physical objects. This point appeared stronger when the picture of the physical world was of a realm of "colliding billiard balls", but I think the general point stands. After all, the idea that physical changes can only occur in some ways and that these

ways involve physical attributes remains plausible (and empirically adequate). For how *could* a non-physical entity generate a change in a physical system? The possibilities of transference of energy or momentum are unavailable; non-physical entities would not possess those quantities to begin with. The cruder notions of pushing, pulling, repelling and the like also seem forbidden, since those entities would not have mass or even surface to exert any kind of contact. It seems, then, that there is no possible way for non-physical things to cause physical effects, since the effects require a mode of interaction that is unavailable to the kind of entity that figures in the causes. We are, once again, lead to the threat of exclusion.